EVALUATING GLOBAL FIT INDEX CUT-OFFS FOR MULTIDIMENSIONAL IRT MODELS

by

BENJAMIN LISTYG

(Under the Direction of Nathan T. Carter)

ABSTRACT

Multidimensional item response theory (MIRT) model selection has only recently come about following the development of the M_r-family of statistics (Maydeu-Olivares & Joe, 2005; 2006). Global model fit indices are now available for a wide range of MIRT models and are theoretically equivalent to typical χ^2-based fit indices (e.g. RMSEA, SRMSR, CFI, and TLI). These fit indices are evaluated relative to popular cut-offs, such as those derived by Hu and Bentler (1998, 1999). The purpose of the present study was to establish whether these popular cut-offs achieve acceptable levels of Type 1 error for model selection or have adequate Power for detecting misfit in MIRT models. Results from two simulation studies suggest the performance of these cut-offs varies depending on the MIRT model of interest and study design characteristics and the power to detect model misspecification is contingent on the type of misspecification present.

INDEX WORDS: Item Response Theory, M_2 Statistics, Dominance Models, Unfolding Models, Model-Data Fit

EVALUATING GLOBAL FIT INDEX CUT-OFFS FOR MULTIDIMENSIONAL IRT MODELS

by

BENJAMIN LISTYG

B.S., University of Georgia, 2016

A Thesis Proposal Submitted to the Graduate Faculty of the University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2022

EVALUATING GLOBAL FIT INDEX CUT-OFFS FOR MULTIDIMENSIONAL IRT MODELS

by

BENJAMIN LISTYG

Major Professor: Nathan T. Carter

Committee: Dorothy Carter

Allan Cohen

Electronic Version Approved:

Ron Walcott
Dean of the Graduate School
The University of Georgia
December 2022

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

The measurement of latent variables underlies nearly all empirical research in the psychological and organizational sciences. There are numerous strategies and modelling techniques researchers can use when fitting latent variable models to survey data. One family of models in particular, confirmatory Item Response Theory (IRT) models, provide benefits that cannot be obtained by alternative models, such as structural equation models (SEM) and confirmatory factor analysis (CFA). First, IRT models are known to be a more faithful representation of item responses as categorical (Tay et al., 2015). Second, IRT latent trait estimates have more desirable properties than SEM factor scores. Lastly, IRT scoring results in more accurate conclusions in tests of interactions (Embertson, 1996; Morse et al., 2012) and curvilinearity (Carter et al., 2017). While IRT is also associated with several downsides, including the need for greater sample size and slow estimation, advances in optimization algorithms and numerical analysis techniques (e.g., the BFGS algorithm; Fletcher, 2013) have lessened these concerns.

One problem historically associated with IRT that has led to its slow adoption by applied researchers is a lack of global model-data fit indices. Indeed, the seminal work by Embretson and Riese (2000) claims that "the science of statistical model comparison in IRT is deficit, especially when compared to structural equation models" (p. 124). For CFA and SEM models, researchers have been able to use indices based on the χ^2 statistic to evaluate discrepancies between the sample and model-implied correlation matrix. These same χ^2 tests can, in theory, be used to

evaluate model-data fit for IRT models by examining discrepancies between the observed and expected contingency tables of response patterns. However, the size of such contingency tables is dependent on the number of items and number of response options for a given measurement scale and will yield K^n cells in total (assuming an equal number of response options across all items) where K is the number of response options, n is the total number of items in a survey, and each cell represents a given response pattern a member of the sample could have generated. In practice, these tables suffer from the issue of sparseness, such that only a small portion of response patterns are observed relative to all possible patterns that exist for a given survey, thereby generating many empty table cells. This issue of sparseness is what makes the χ^2 statistic unusable for evaluating IRT model-data fit as the Type 1 error rates increases as a function of the sparseness in such a table (Read & Cressie, 1988).

Within the past decade, advances in statistical theory (Maydeu-Olivares & Joe, 2005; Cai & Hansen, 2013) have led to the creation of "limited information" statistics (e.g. M_2 and M_2^*) that resolve the issues traditional χ^2 tests have for assessing IRT model fit. Limited information statistics remedy this sparseness issue by using information contained in the margins or moments of these contingency tables, creating a family of test statistics that retain the same distributional properties as the χ^2 statistic. Such developments now allow researchers to estimate global-model fit indices similar to those based on χ^2 for SEM and CFA (e.g. CFI, TLI, and RMSEA). Although these indices and their respective cut-off values (Hu & Bentler, 1998; 1999) are well-known and widely used across all areas of psychology for evaluating SEMs and CFAs, it is not clear how these fit indices operate in the context of IRT model selection. The lack of guidance on model-data fit indices for IRT has resulted in researchers making educated guesses regarding what

values of these fit indices lead to adequate model selection (Chalmers, 2017) or relying on traditional cut-offs for model evaluation and selection (Nye et al., 2020).

Here, I conducted two simulation studies that evaluate these fit indices and their cut-off values on (a) the Type 1 error of selecting a multidimensional IRT model capturing the "true" data generating process of a survey scale and (b) their power to detect various forms of multidimensional IRT model misspecification. I begin by providing an overview of the χ^2-based fit statistics frequently used by researchers for evaluating latent variable models. Next, I introduce the limited information, M_r, family of statistics and move on to introducing multidimensional extensions of several popular IRT models used in organizational research. After providing an overview of the simulation procedures used to examine these fit statistics and presenting my results, I discuss the implications of these findings for researchers evaluating multidimensional IRT models.

CHAPTER 2

GLOBAL MODEL DATA FIT IN SEM VERSUS IRT

In psychological research, evaluating competing theoretical models of interest is often accomplished by comparing the discrepancy between the sample covariance matrix, S_{xx} , with the covariance matrix implied or expected by the model under consideration, $\hat{\Sigma}(\Theta)$. This discrepancy can be calculated as follows using a test statistic, χ^2:

$$\chi^2 = n[trace(S_{xx}\hat{\Sigma}(\Theta)^{-1}) - \ln|S_{xx}\hat{\Sigma}(\Theta)^{-1}| - k]$$

where *n* is the sample size and *k* is the number of items in a scale. This test statistic follows a χ^2 distribution with $df = p - q$ where p is the number of unique parameters in the covariance matrix, computed as $(\frac{k(k+1)}{2})$, and q is the number of freely estimated parameters in $\hat{\Sigma}(\Theta)$.

Evaluating χ^2 by itself leads to poor model selection decisions, such that for trivial degrees of model misspecification, the power to detect misfit approaches 1.0 as the sample size increases. Because of this issue, numerous fit indices have been developed based on this χ^2 statistic to evaluate how much this discrepancy is indicative of a model sufficiently fitting the data. The most popular of these fit statistics are the Comparative Fit Index (CFI), the Tucker Lewis Index (TLI), and the Root Mean Squared Error of Approximation (RMSEA). In addition to these χ^2-based statistics, another commonly reported assessment of model-data fit is the Standardized Root Mean Square Residuals (SRMSR) which, as the name implies, computes the standardized difference between the observed correlation matrix and the model-implied correlation matrix and is not based on the $\chi 2$ statistic. These fit statistics can be computed as follows:

$$\text{CFI} = 1 - \frac{\chi 2 - df}{\chi 2_{Nuil} - df_{null}} \quad (1)$$

$$\text{TLI} = \frac{\chi 2 - df}{\chi 2_{Nuil} - df_{null}} \quad (2)$$

$$\text{RMSEA} = \sqrt{\frac{\left(\frac{\chi 2}{df - 1}\right)}{N - 1}} \quad (3)$$

$$\text{SRMSR} = \sqrt{\sum (\hat{\Sigma}(\Theta) - (S_{xx}))^2} \quad (4)$$

In their seminal papers, Hu and Bentler (1998, 1999) demonstrated via simulation what values CFI, TLI, RMSEA, and SRMSR should taken on to ensure researchers are selecting the "true" latent variable model of interest. For TFI and CLI, the obtained fit statistic should be greater than 0.95, while the RMSEA should be less than 0.05 and SRMSR should be less than 0.08.

CHAPTER 3

M_2 STATISTICS

Maydeu-Olivares and Joe (2005; 2006) derived the "limited information" family of statistics, referred to as M_r statistics, to provide researchers with the same capabilities as χ^2 when examining model-data fit for categorical variables. These M_r statistics can be computed as follows:

$$M_r = N(\mathbf{p}_\mathrm{r} - \boldsymbol{\pi}_\mathrm{r})'\widehat{\boldsymbol{C}}_\mathrm{r}(\mathbf{p}_\mathrm{r} - \boldsymbol{\pi}_\mathrm{r}) \quad (5)$$

where N is the sample size and r is the moment of the contingency table of response patterns. Maydeu-Olivares and Joe (2005) demonstrate that model selection is optimal when r goes up to 2, meaning only the univariate and bivariate moments of the contingency table are used in model evaluation. At $r = 1$, these univariate moments are simply the marginal probability or proportion of responses for the category of given item (see p. 307, Maydeu-Olivares & Joe, 2014). Thus, p_1 represents the *observed* marginal probabilities for the category of a given item obtained from the sample, π_1 represents the marginal, *expected* probability for the category of a given item, and $\mathrm{p}_1 - \pi_1$ represents the difference between these two values (i.e. the marginal residuals at the first moment). At $r = 2$, these bivariate moments represent the joint probability of item responses (e.g. the probability of selecting response option 1 for item 1 *and* selecting response option 1 on item 2), such that p_2 is the joint probability of item responses obtained from the sample, π_2 represents the expected probability of joint item responses, and $\mathrm{p}_2 - \pi_2$ represents the residuals at the second moment.

The $\hat{C}_{\mathrm{r}}$ portion of this equation, referred to as the weight matrix, can be further decomposed into the following matrices:

$$\boldsymbol{\Delta}_{\boldsymbol{r}}^{(\boldsymbol{c})}\left(\boldsymbol{\Delta}_{\boldsymbol{r}}^{(\boldsymbol{c})\prime}\boldsymbol{\Xi}_{\boldsymbol{r}}\boldsymbol{\Delta}_{\boldsymbol{r}}^{(\boldsymbol{c})}\right)^{-1}\boldsymbol{\Delta}_{\boldsymbol{r}}^{(\boldsymbol{c})\prime} \tag{6}$$

where $\boldsymbol{\Xi}_{\boldsymbol{r}}$ is simply referred to as the asymptotic covariance matrix whose elements are the product of the marginal probabilities of item responses subtracted from the joint probabilities of item responses (e.g. $\pi_{Response\ Option\ 1\ for\ Item\ 1\ \&\ Response\ Option\ 1\ for\ Item\ 2} - \pi_{Response\ Option\ 1\ for\ Item\ 1}\ x\ \pi_{Response\ Option\ 1\ for\ Item\ 2}$). The computational challenge of computing the M_2 statistic comes from generating the elements of the $\boldsymbol{\Delta}_{\boldsymbol{r}}^{(\boldsymbol{c})}$ matrix. Each element in this matrix is a partial derivative of the population-implied joint and marginal probabilities with respect to the item parameters being estimated. In simpler terms, the elements of $\boldsymbol{\Delta}_{\boldsymbol{r}}^{(\boldsymbol{c})}$ represent the values of the item parameters that maximize the probability of obtaining the responses contained in the contingency table.

While initially designed for dichotomous items, Cai and Hansen (2013) demonstrated that M_2 can be viewed as a special case of a more general family of statistics, M_2^*. M_2^* further reduces the size of these high dimensional contingency tables by summing the residuals across the response categories for a given item and can be applied to both dichotomous and polytomous survey scales with varying response options.

M_2-Based Global Model Fit Indices

As noted previously, global model fit statistics for CFA and SEM models are based on the χ^2 statistic. These same fit statistics can be adapted for M_2 and M_2^* statistics and are represented as follows:

$$\text{CFI} = 1 - \frac{M_2^* - df}{M_{2_{Nuil}}^* - df_{null}} \tag{7}$$

$$\text{TLI} = \frac{M_2^* - df}{M_{2_{Nuil}}^* - df_{null}} \tag{8}$$

$$\text{RMSEA} = \sqrt{\frac{\left(\frac{M_2^*}{df - 1}\right)}{N - 1}} \tag{9}$$

while SRMSR remains unchanged.

To date, RMSEA and SRMSR have been the only fit indices to be empirically examined in the context of IRT model selection. Simulation results from Maydeu-Olivares and Joe (2014) looking at M_r-based RMSEA and SRMSR suggest that a candidate IRT model under consideration would be considered an adequate fit to the data if the RMSEA is less than 0.089 and SRMSR is less than 0.05, a close fit if RMSEA is less than 0.05 and SRMSR is less 0.027, and an excellent fit if RMSEA is less than 0.05 / (K – 1) and SRMSR is less than 0.027 / (K – 1) where K is the number of item response categories. These cut-offs derived by Maydeu-Olivares and Joe (2014) were subsequently used by Nye et al. (2020) for evaluating IRT model-data fit in the context of *response process* misspecification. In their study, Nye et al. fit a series of unidimensional (e.g. Generalized Graded Unfolding Model; GGUM and Graded Response Model; GRM) and multidimensional (e.g. Multidimensional GRM; MGRM and Multidimensional 2-Parameter Logistic Model; M2PL) IRT models to data generated from these same models, each representing a specific response style (i.e. the GGUM was fit to data from the 2PL, 3PL, GGUM, and M2PL, see p. 472 for an example). The authors then examined if a variety of fit statistics were able to detect this type of model misspecification including: Yen's QI, posterior predictive p-values, χ^2 Singles, χ^2 Doubles, χ^2 Triples, and SRMSR.

CHAPTER 4

THE PRESENT STUDY

The development of the M_r and M_2^* statistics now allows researchers to confirm model-data fit for IRT models. However, with the exception Nye et al.'s examination of the M2PL and MGRM, prior work in this area has predominantly examined unidimensional IRT models leaving no guidance for researchers to turn to when making decisions regarding multidimensional IRT model selection. Such lack of guidance is problematic given the increasing adoption of IRT modelling in psychology (Lang & Tay, 2021) and most constructs studied in psychology being multidimensional, such as personality (Tellegan & Waller, 2008), vocational interests (Wetzel & Hell, 2014; Tay et al., 2009), and psychopathology (Lillienfeld, 2018). The present study aims to provide guidance in this domain this by assessing the performance of the aforementioned global model fit statistics and their respective cut-offs for MIRT model selection and detecting multidimensional IRT model misspecification.

CHAPTER 5

METHOD

To assess the performance of the aforementioned fit index cutoff values for multidimensional IRT models, I conduct two simulations in which I first generate data from a known, "true" IRT model with preset item parameters. Next, these data are fit to both a correct specification of the data generating process as well as an incorrect specification in addition to varying sample size, test length, response options, and the number of dimensions assessed. Most relevant to the current investigation are the types of model misspecifications: the percentage of items in the "wrong" model that incorrectly have discrimination parameters (i.e., item factor loadings) equal to zero (i.e., number of α_i=0) and the number of between-factor correlations set to zero in the "incorrect model" (i.e., number of $\rho_{\theta_1\theta_2} = 0$). I evaluate these factors in data with different types of multidimensional response models: the multidimensional Graded Response Model (Samejima, 1969), the multidimensional Generalized Partial Credit Model (Muraki, 1992) and the multidimensional Generalized Graded Unfolding Model (Roberts, Donoghue, & Laughlin, 2000). These three types of response process models were chosen given their prevalence in organizational and educational research settings. The final simulation contains 864 conditions in total (see Table 1).

Multidimensional IRT Models

The unidimensional Graded Response Model (GRM) is an extension of the Two-Parameter Logistic (2PL) Model for polytomous data (Samejima, 1969). For a single latent trait, the GRM can be expressed as:

$$P(u_{ij} = k|\theta_j, a_i, b_{ik}) = P(u_{ij} \geq k|\theta_j, a_i, b_{ik}) - P(u_{ij} \geq k+1|\theta_j, a_i, b_{i(k+1)}) \quad (10)$$

where $P(u_{ij} \geq k|\theta_j, a_i, b_{ik}) = \frac{1}{1+e^{[-(a_i(\theta'_j+b_{ik}))]}}$ and $P(u_{ij} \geq k+1|\theta_j, a_i, b_{ik}) = \frac{1}{1+e^{[-(a_i(\theta'_j+b_{ik+1}))]}}$. Here, $P(u_{ij} = k|\theta_j, a_i, b_{ik})$ denotes the conditional probability of the observed rating *u* made an individual *j* selecting response option *k* on item *i*, θ_j denotes individual *j*'s stance on the latent trait, a_i is the discrimination parameter for item *i,* and b_{ik} is the threshold parameter for item *i* in response category *k*. These threshold parameters represent the level of θ at which a person is equally likely to choose the response option *k* rather than *k*-1, the response option below it. Thus, the number of threshold parameters is one less than the number of response options for a given item. Extending the GRM to multiple dimensions, the MGRM is typically written as:

$$P(u_{ij} = k|\boldsymbol{\theta}_j, \boldsymbol{a}_i, d_{ik}) = \frac{1}{1+e^{[-(\boldsymbol{a}_i\boldsymbol{\theta}'_j+d_{ik}))]}} - \frac{1}{1+e^{[-(\boldsymbol{a}_i\boldsymbol{\theta}'_j+d_{ik+1})]}} \quad (11)$$

where $\boldsymbol{a}_i$ is a vector of item discrimination parameters for item *i* indexing each dimension measured, $\boldsymbol{\theta}_j$ is a vector of trait levels for person *j* over each dimension measured. Finally, d_{ik} represents the intercept parameter for item *i* on response category *k*. Wang et al. (2018) note that when a multidimensional test displays a simple structure, such that each item *i* corresponds to a unique latent factor, d_{ik} can be recomputed as the product -$a_i b_{ik}$.

The unidimensional generalized partial credit model (Reckase, 2009) is written as:

$$P(u_{ij} = k|\theta_j, a_i, b_{ik}) = \frac{e^{[ka_i\theta_j - \Sigma_{u=0}^{k} a_i b_{iu}]}}{\Sigma_{h=0}^{C} e^{[ha_i\theta_j - \Sigma_{u=0}^{k} a_i b_{iu}]}} \quad (12)$$

Where $P(u_{ij} = k|\theta_j, a_i, b_{ik})$ is the conditional probability of the observed rating *u* made an individual *j* selecting response option *k* on item *i,* θ_j is person *j*'s stance on the latent trait, a_i

is the item discrimination parameter for item *i*, b_i is the category threshold parameter for item *i*, *C* is the total number of rating categories minus 1 for item *i*. The multidimensional extension of the GPCM allows for a vector of item discrimination parameters, $\boldsymbol{a}_i$, that indexes each items discrimination over the number of latent traits being estimated as well as a vector of latent trait scores, $\boldsymbol{\theta}_j$, and can be written as:

$$P(u_{ij} = k|\boldsymbol{\theta_j}, \boldsymbol{a_i}, d_{ik}) = \frac{e^{[k\boldsymbol{a}_i\boldsymbol{\theta}_j + d_{iu}]}}{\Sigma_{h=0}^{C} e^{[h\boldsymbol{a}_i\boldsymbol{\theta}_j + d_{iu}]}} \quad (13)$$

Similar to the MGRM, the MGPCM can be re-parameterized in slope-intercept form with each item intercept computed as $-a_i\Sigma_{u=1}^{k} b_k$ when the test displays a simple structure (Matlock et al., 2018).

The generalized graded unfolding model (Roberts et al., 2000) is a class of ideal point models based on the GPCM expressed as:

$$P(u_{ij} = k|\theta_j, \alpha_i, \delta_i, \tau_{ik}) = \frac{e^{\alpha_i[k(\theta_j - \delta_i) - \Sigma_{u=0}^{k}\tau_{iu}]} + e^{\alpha_i[M-k(\theta_j - \delta_i) - \Sigma_{u=0}^{k}\tau_{iu}]}}{\Sigma_{h=0}^{C} e^{\alpha_i[h(\theta_j - \delta_i) - \Sigma_{u=0}^{h}\tau_{iu}]} + e^{\alpha_i[M-h(\theta_j - \delta_i) - \Sigma_{u=0}^{h}\tau_{iu}]}} \quad (14)$$

where $P(u_{ij} = k|\theta_j, \alpha_i, \delta_i, \tau_{ik})$ is the conditional probability of the observed rating *u* made an individual *j* selecting response option *k* on item *i,* θ_j is person *j*'s stance on the latent trait, α_i is the item discrimination parameter for item *i*, δ_i is the item location parameter, τ_{ik} represents the location of the k^{th} subjective response option, *C* represents the number of item categories minus 1, and $M = 2C + 1$ denoting the total number of τ_{ik} parameters to be estimated. Wang and Wu (2016) extend the GGUM to multiple dimensions as follows:

$$P(u_{ij} = k|\boldsymbol{\theta}_j, \alpha_i, \delta_i, \tau_{ik}) = \frac{e^{\alpha_i[k(\boldsymbol{d}_i'\boldsymbol{\theta}_j - \delta_i) - \Sigma_{u=0}^{k}\tau_{iu}]} + e^{\alpha_i[M-k(\boldsymbol{d}_i'\boldsymbol{\theta}_j - \delta_i) - \Sigma_{u=0}^{k}\tau_{iu}]}}{\Sigma_{h=0}^{C} e^{\alpha_i[h(\boldsymbol{d}_i'\boldsymbol{\theta}_j - \delta_i) - \Sigma_{u=0}^{h}\tau_{iu}]} + e^{\alpha_i[M-h(\boldsymbol{d}_i'\boldsymbol{\theta}_j - \delta_i) - \Sigma_{u=0}^{h}\tau_{iu}]}} \quad (15)$$

where $\boldsymbol{\theta}_j$ now denotes a vector of latent traits indexed over the number of dimensions being estimated and $\boldsymbol{d}_i'$ is a design vector specifying which dimensions are measured by item *i.* In the

case of a single dimension, d_i reduces to 1 for every item. As with the MGRM and MGPCM, computational implementations for estimating the MGGUM parameterize this model in slope-intercept form such that $P(u_{ij} = k|\boldsymbol{\theta}_j, a_i, b_i, t_{ik})$. When estimating a simple-form test with the slope-intercept parameterization of the MGGUM, $\alpha_i = a_i$, $\delta_i = b_i$, and $\tau_{ik} = -t_{ik}$ (Liu & Chalmers, 2018).

Item Parameter Generation

Item parameters for the MGRM[1] were generated following the procedures used in Roberts et al. (2000), Stark et al. (2006), and Kieftenbeld and Natesan (2012). For the dichotomous condition, the discrimination parameters will be sampled from a random uniform distribution [0.5, 2.0]. The location parameters will be sampled from a random uniform distribution [–2.0, 2.0]. Discrimination parameters for the four and six response category conditions will be generated following the same procedure as dichotomous conditions. In addition, the threshold parameters for the four response option condition, b_1, b_2, b_3, and six response option condition, b_1, b_2, b_3, b_4, and b_5, will be obtained from random uniform distributions $b_1 \sim U$[–2.0, -0.67], $b_2 \sim U$[–0.67, 0.67], and $b_3 \sim U$[0.67, 2.0], and $b_1 \sim U$[–2, –1.2], $b_2 \sim U$[–1.2, -0.4], $b_3 \sim U$[-0.4, 0.4], $b_4 \sim U$ [0.4, 1.2], $b_5 \sim U$[1.2, 2.0], respectively.

Item parameters were for the MGPCM[2] were generated following the procedure described in Kang et al. (2006), Kang et al. (2009), and Sung and Kang (2008). The discrimination parameters (α_i) will be randomly sampled from a lognormal (0, 0.5^2) distribution. For the six category conditions, five item category parameters, $\delta_{ci} = \beta_i - \tau_{ci}$, where c = 1, 2, 3,

[1] Given Chalmers (2012) uses a slope-intercept parameterization of the MGRM, slope parameters will remain unchanged (i.e. $a_i = a_i$) while data generation for intercept parameters will be recomputed as follows: $d_{ik} = -a_i b_{ik}$ where *k* represents the k^{th} response option to the i^{th} item.

[2] For the MGPCM, data generation for item parameters will be computed as follows: $a_i = a_i$ and $d_k = -a_i \sum_k b_{ik}$ where *k* represents the k^{th} response option to the i^{th} item.

4, or 5, will be drawn from five normal distributions with a common standard deviation of .5 and means of –2, -1, 0.5, 1, and 2, respectively. The mean of these five step parameters will then be used as the item difficulty parameter (β_i), and the difference between β_i and δ_{ci} is taken as the item threshold parameter, τ_{ci}. For the four category condition, three item category parameters, $\delta_{ci} = \beta_i - \tau_{ci}$, where c = 1, 2, or 3, will be drawn from three normal distributions with a common standard deviation of .5 and means of –2, 0.5, and 2, respectively. The mean of these three step parameters will then be used as the item difficulty parameter (β_i), and the difference between β_i and δ_{ci} is taken as τ_{ci}. Finally, for the dichotomous condition, the single threshold parameter, τ_1, will be drawn from a $N(0,1)$ distribution. As is common for GPCM models, τ_0 will be set to 0 for model identification purposes.

Item parameters for the MGGUM[3] were generated following the procedure listed in Roberts et al. (2002). Item locations, δ_{ij}, will be simulated to range from -2 to 2 with equal spacing, reflecting Thurstone's (1928) notion that items should be distributed across the latent attribute continuum. Discrimination parameters, α, will be sampled from a uniform distribution bounded between 0.5 and 2. The highest option threshold, τ_{ijk}, for each item will be drawn from a uniform (-1.4, 0.4) distribution. Successive τ_{ijk} parameters will be generated with the following recursive equation: $\tau_{ijk-1} = \tau_{ijk} - .25 + e_{ijk-1}, for\ k = 2, 3, \ldots C$ where C is the higher number of response options and $e_{ijk-1} \sim N(0,0.04)$. For all MIRT models under examination, person ability parameters, θ_j, will be distributed multivariate normal with means of 0 and standard deviations of 1. Correlations between dimensions, $\rho_{\theta_1\theta_2}$, $\rho_{\theta_1\theta_3}$, and $\rho_{\theta_2\theta_3}$ will be set to 0.5.

[3] For the MGGUM, item parameters will be converted as follows for data generation: $a_i = \alpha_i$, $b_i = \delta_i$, $t_{ik} = -\tau_{ik}$ where *k* represents the k^{th} response option to the i^{th} item.

CHAPTER 6

STUDY 1 METHOD

Analytic Design

For study 1, I evaluated the ability of the cut-offs to determine a given candidate model is the model representing the true data generating process. In the context of this study, a Type 1 error occurs when the *true* model was incorrectly determined to not fit the data. The simulation procedure for Study 1 is as follows:

1) Generate item parameters following the previously mentioned approaches
2) Simulate response data using the item parameters from Step 1
3) Fit the "true" data generating model to the simulated data
4) Compute the number of number of times the M_2^*-based fit statistics indicate the true model did not fit the data

This three-step procedure will be repeated 100 times per simulation condition, resulting in 14,400 total simulation runs.

CHAPTER 7

STUDY 1 RESULTS

To ensure all MIRT models were correctly estimated, I computed the mean bias, root mean squared error (RMSE), and correlation between the sample and population discrimination, location, and category threshold parameters for each simulation condition. Mean bias was computed as the average difference between the sample item parameter value and its corresponding population item parameter. RMSE was computed as the square root of the average squared difference between sample and population item parameter values. The correlation between sample and item parameters was computed as the typical standardized covariance between parameters. As shown in Figure 1 through Figure 9, the bias and RMSE between the population and sample item parameters approach 0 and the correlation between the population and sample item parameters approaches 1.0 as the sample size increases for all three MIRT models. These results indicate that the sample item parameters for all three models converged towards the population item parameters as sample size increased. These results replicate prior simulation studies examining sample size requirements for the MGPCM (Kose & Demirtasli, 2012), MGGUM (Wang & Wu, 2016), and MGRM (Jiang et al., 2016).

Type 1 Error

Type 1 Error results are displayed in Figure 16 through Figure 18. Overall, the efficacy of global model fit index cut-offs for MGPCM, MGRM, and MGGUM model selection is mixed. While some conditions had Type 1 Error steadily decrease as sample size increased (e.g. the Multidimensional 2PL, see Figure 16), other statistics had a T1E of either 100% (e.g. the three-

dimensional GPCM with dichotomous items) or 0% (e.g. the CFI, TLI, and RMSEA of the two-dimensional GGUM with four response options) across all simulation conditions. These results can best be explained by Figures 10 through 15 that demonstrate how some simulation conditions have fit statistics from the *correct* model fit to the data that were either all higher or lower than the cut-off value for that condition. Results for each fit statistics are presented in the following and are organized by fit statistic, MIRT model type, and simulation factor and are collapsed across all other simulation factors.

Across all dichotomous response option conditions for the 0.95 CFI cut-off applied to the MGRM, 37.5% (6/16) achieved nominal levels of. Across all four response option conditions for the 0.95 cut-off applied to the MGRM, 100% (16/16) were able to reach nominal levels of T1E. Across all six response option conditions of the 0.95 cut-off applied to the MGRM, 100% (16/16) were able to reach nominal levels of T1E. Similarly, across all two response option conditions for the 0.95 cut-off applied to the MGPCM, 25% (4/16) were able to reach nominal levels of T1E. Across all four response option conditions for the 0.95 cut-off applied to the MGPCM, 75% (12/16) were able to reach nominal levels of T1E. Across all six response option conditions of the 0.95 cut-off applied to the MGPCM, 100% (16/16) were able to reach nominal levels of T1E. Lastly, across all two response option conditions for the 0.95 cut-off applied to the MGGUM, 43.75% (7/16) were able to reach nominal levels of T1E. Across all four response option conditions for the 0.95 cut-off applied to the MGGUM, 100% (16/16) were able to reach nominal levels of T1E. Across all six response option conditions of the 0.95 cut-off applied to the MGGUM, 100% (16/16) were able to reach nominal levels of T1E.

Across all 18 item conditions for the 0.95 CFI cut-off applied to the MGRM, 75% (18/24) were able to reach nominal levels of T1E, while 83.33% (20/24) of the 36 item condition

were able to reach nominal levels of T1E. Similarly, 50% (12/24) of all 18 item conditions for the 0.95 CFI cut-off applied to the MGPCM and 83.33% (20/24) of the 36 item conditions were able to reach nominal levels of T1E. Lastly, Across all 18 item conditions for the 0.95 CFI cut-off applied to the MGGUM, 79.17% (19/24) were able to reach nominal levels of T1E, while 83.33% (20/24) of the 36 item condition were able to reach nominal levels of T1E.

Across all two-dimension conditions for the 0.95 CFI cut-off applied to the MGRM, 91.67% (22/24) were able to reach nominal levels of T1E, while 66.67% (16/24) of the three-dimension conditions were able to reach nominal levels of T1E. Similarly, 83.33% (20/24) of all two-dimension conditions for the 0.95 CFI cut-off applied to the MGPCM and 50% (12/24) of the three-dimension condition were able to reach nominal levels of T1E. Lastly, Across all two-dimension conditions for the 0.95 CFI cut-off applied to the MGGUM, 95.83% (23/24) were able to reach nominal levels of T1E, while 66.67% (16/24) of the three-dimension condition were able to reach nominal levels of T1E.

Across all dichotomous response option conditions for the 0.95 TLI cut-off applied to the MGRM, 31.25% (5/16) were able to reach nominal levels of. Across all four response option conditions for the 0.95 cut-off, 93.75% (15/16) were able to reach nominal levels of T1E. Across all six response option conditions of the 0.95 cut-off 100% (16/16) were able to reach nominal levels of T1E. Similarly, across all two response option conditions for the 0.95 cut-off applied to the MGPCM, 25% (4/16) were able to reach nominal levels of T1E. Across all four response option conditions for the 0.95 cut-off, 68.75% (11/16) were able to reach nominal levels of T1E. Across all six response option conditions of the 0.95 cut-off, 100% (16/16) were able to reach nominal levels of T1E. Lastly, across all two response option conditions for the 0.95 cut-off applied to the MGGUM, 31.25% (5/16) were able to reach nominal levels of T1E. Across all

four response option conditions for the 0.95 cut-off, 81.25% (13/16) were able to reach nominal levels of T1E. Across all six response option conditions of the 0.95 cut-off, 93.75% (15/16) were able to reach nominal levels of T1E.

Across all 18 item conditions for the 0.95 TLI cut-off applied to the MGRM, 75% (16/24) were able to reach nominal levels of T1E, while 83.33% (20/24) of the 36 item condition were able to reach nominal levels of T1E. Similarly, 45.83% (11/24) of all 18 item conditions for the 0.95 TLI cut-off applied to the MGPCM, while 83.33% (20/24) of the 36 item condition were able to reach nominal levels of T1E. Lastly, Across all 18 item conditions for the 0.95 TLI cut-off applied to the MGGUM, 54.17% (13/24) were able to reach nominal levels of T1E, while 83.33% (20/24) of the 36 item condition were able to reach nominal levels of T1E.

Across all two-dimension conditions for the 0.95 TLI cut-off applied to the MGRM 87.5% (21/24) were able to reach nominal levels of T1E, while 62.5% (15/24) of the three-dimension conditions were able to reach nominal levels of T1E. Similarly, 79.17% (19/24) of all two-dimension conditions for the 0.95 TLI cut-off applied to the MGPCM and 50% (12/24) of the three-dimension condition were able to reach nominal levels of T1E. Lastly, Across all two-dimension conditions for the 0.95 TLI cut-off applied to the MGGUM, 87.5% (21/24) were able to reach nominal levels of T1E, while 50% (12/24) of the three-dimension condition were able to reach nominal levels of T1E.

Across all dichotomous response option conditions for the 0.05 RMSEA cut-off applied to the MGRM, 62.5% (10/16) were able to reach nominal levels of T1E. Across all four response option conditions for the 0.05 cut-off, 93.75% (15/16) were able to reach nominal levels of T1E. Across all six response option conditions of the 0.05 cut-off, 100% (16/16) were able to reach nominal levels of T1E. Similarly, across all two response option conditions for the 0.05 cut-off

applied to the MGPCM, 93.75% (15/16) were able to reach nominal levels of T1E. Across all four response option conditions for the 0.05 cut-off, 87.5% (14/16) were able to reach nominal levels of T1E. Across all six response option conditions of the 0.05 cut-off, 100% (16/16) were able to reach nominal levels of T1E. Lastly, across all two response option conditions for the 0.05 cut-off applied to the MGGUM, 93.75% (15/16) were able to reach nominal levels of T1E. Across all four response option conditions for the 0.05 cut-off, 68.75% (11/16) were able to reach nominal levels of T1E. Across all six response option conditions of the 0.05 cut-off, 62.5% (10/16) were able to reach nominal levels of T1E.

Aross all 18 item conditions for the 0.05 RMSEA cut-off applied to the MGRM, 75% (18/24) were able to reach nominal levels of T1E, while 95.83% (23/24) of the 36 item condition were able to reach nominal levels of T1E. Similarly, 87.5% (21/24) of all 18 item conditions for the 0.05 RMSEA cut-off applied to the MGPCM and 100% (24/24) of the 36 item condition were able to reach nominal levels of T1E. Lastly, Across all 18 item conditions for the 0.05 RMSEA cut-off applied to the MGGUM, 62.5% (15/24) were able to reach nominal levels of T1E, while 87.5% (21/24) of the 36 item condition were able to reach nominal levels of T1E.

Across all two dimension conditions for the 0.05 RMSEA cut-off applied to the MGRM, 95.83% (23/24) were able to reach nominal levels of T1E, while 75% (18/24) of the three-dimension conditions were able to reach nominal levels of T1E. Similarly, 100% (24/24) of all two-dimension conditions for the 0.05 RMSEA cut-off applied to the MGPCM and 87.5% (21/24) of the three-dimension condition were able to reach nominal levels of T1E. Lastly, Across all two-dimension conditions for the 0.05 RMSEA cut-off applied to the MGGUM, 100% (24/24) were able to reach nominal levels of T1E, while 50% (12/24) of the three-dimension condition were able to reach nominal levels of T1E.

Across all dichotomous response option conditions for the 0.08 SRMSR cut-off applied to the MGRM, 0% (0/16) were able to reach nominal levels of T1E. Across all four response option conditions for the 0.08 cut-off, 0% (0/16) were able to reach nominal levels of T1E. Across all six response option conditions of the 0.08 cut-off, 0% (0/16) were able to reach nominal levels of T1E. Similarly, across all two response option conditions for the 0.08 cut-off applied to the MGPCM, 25% (4/16) were able to reach nominal levels of T1E. Across all four response option conditions for the 0.08 cut-off, 0% (0/16) were able to reach nominal levels of T1E. Across all six response option conditions of the 0.08 cut-off, 0% (0/16) were able to reach nominal levels of T1E. Lastly, across all two response option conditions for the 0.08 cut-off applied to the MGGUM, 81.25% (13/16) were able to reach nominal levels of T1E. Across all four response option conditions for the 0.08 cut-off, 0% (0/16) were able to reach nominal levels of T1E. Across all six response option conditions of the 0.08 cut-off, 0% (0/16) were able to reach nominal levels of T1E.

Across all 18 item conditions for the 0.08 SRMSR cut-off applied to the MGRM, 0% (0/24) were able to reach nominal levels of T1E, while 0% (0/24) of the 36-item condition were able to reach nominal levels of T1E. Similarly, 4.17% (1/24) of all 18 item conditions for the 0.08 SRMSR cut-off applied to the MGPCM and 12.5% (3/24) of the 36 item conditions were able to reach nominal levels of T1E. Lastly, Across all 18 item conditions for the 0.08 SRMSR cut-off applied to the MGGUM, 25% (6/24) were able to reach nominal levels of T1E, while 29.17% (7/24) of the 36 item condition were able to reach nominal levels of T1E.

Across all two-dimension conditions for the 0.08 SRMSR cut-off applied to the MGRM, 0% (0/24) were able to reach nominal levels of T1E, while 0% (0/24) of the 3-dimension conditions were able to reach nominal levels of T1E. Similarly, 25% (4/24) of all 2-dimension

conditions for the 0.08 SRMSR cut-off applied to the MGPCM and 0% (0/24) of the three-dimension condition were able to reach nominal levels of T1E. Lastly, Across all two-dimension conditions for the 0.08 SRMSR cut-off applied to the MGGUM, 33.33% (8/24) were able to reach nominal levels of T1E, while 20.83% (5/24) of the three-dimension condition were able to reach nominal levels of T1E.

CHAPTER 8

STUDY 2 METHOD

Analytic Design

For study 2, I evaluate how well these fit statistics are able to detect various forms of model misspecification. In the context of this study, a Type 2 error is the erroneous decision that the *incorrect* candidate model was deemed to be the *true* data generating process. Thus, Power for study 2 is the frequency of correctly determining that the incorrect model was *not* the true model that generated the data. Like study 1, data were generated from the three aforementioned MIRT models using the previously stated item parameter generation procedures. However, for study 2, I fit a number of models that deviate from the true data generating process in several ways. These incorrect models contain either (a) a varying percentage of misspecified item loadings (e.g. 10% or 20% of items with $\alpha_i = 0$), (b) a varying percentage of misspecified inter-factor correlations (e.g. $\rho_{\theta 1 \theta 2} = 0$), or (c) a combination of misspecified item loadings and misspecified inter-factor correlations. The simulation procedure for study 2 is as follows:

1) Generate item parameters following the previously mentioned approaches
2) Simulate response data using the item parameters from Step 1
3) Fit a misspecified model to the simulated data
4) Compute the number of number of times the M_2^*-based fit statistics from the incorrect model was greater than or less than the $5^{th}/95^{th}$ percentile[4] fit statistic of the same simulation condition from Study 1.

[4] This is done to ensure the Type 1 Error is held constant at 5%

CHAPTER 9

STUDY 2 RESULTS

Power results are displayed in Figures 19 through 34. Overall, results indicate that detecting model misspecification for confirmatory MIRT models is conditional on the degree of misspecification between the true data generating process and the candidate model fit to the data at hand. These findings suggest that it is easier to detect model misspecification when the deviation from the "true" model is large. For example, detecting model misspecification when 20% of the items and 2 out of 3 of the inter-factor correlations are mis-specified occurred in 100% of simulation conditions with samples as small as 250 individuals. In contrast to detecting large degrees of model misspecification, detecting small amounts of misspecification in a confirmatory MIRT model is substantially more difficult with realistic sample sizes. For example, detecting *only* a mis-specified inter-factor correlation failed to achieve nominal levels of power in *all* simulation conditions. Given the fully crossed nature of each simulation condition for the Power study and the large number of conditions, results are presented in ascending order of misspecification for each fit statistic and simulation condition pairing. For a given misspecification condition, results are collapsed across all other conditions.

No simulation condition was able to achieve nominal levels of power when an inter-factor correlation was misspecified. Similarly, no simulation condition was able to achieve nominal levels of power when two inter-factor correlations were misspecified.

Across all dichotomous response option conditions for the 10% of Items with $\alpha_i = 0$ misspecification condition, 72.92% (35/48) were able to reach nominal levels of power. Across

all four response option conditions, 97.92% (47/48) were able to reach nominal levels of power. Across all six response option conditions of the, 100% (48/48) were able to reach nominal levels of power.

Across all 18 item conditions for the 10% of Items with $\alpha_i = 0$ misspecification condition, 81.94% (59/72) were able to reach nominal levels of power while 98.61% (71/72) of the 36 item conditions reached nominal levels of power.

Across all two-dimension conditions for the 10% of Items with $\alpha_i = 0$ misspecification condition, 94.44% (68/72) were able to reach nominal levels of power while 86.11% (62/72) of the three-dimension conditions reached nominal levels of power.

Across all dichotomous response option conditions for the 10% of Items with $\alpha_i = 0$ misspecification condition, 70.83% (34/48) were able to reach nominal levels of power. Across all four response option conditions, 97.92% (47/48) were able to reach nominal levels of power. Across all six response option conditions of the, 100% (48/48) were able to reach nominal levels of T1E.

Across all 18 item conditions for the 10% of Items with $\alpha_i = 0$ misspecification condition, 81.94% (59/72) were able to reach nominal levels of power while 98.61% (71/72) of the 36 item conditions reached nominal levels of power.

Across all two-dimension conditions for the 10% of Items with $\alpha_i = 0$ misspecification condition, 95.83% (69/72) were able to reach nominal levels of power while 83.33% (60/72) of the three-dimension conditions reached nominal levels of power.

Across all dichotomous response option conditions for the 10% of Items with $\alpha_i = 0$ misspecification condition, 37.5% (18/48) were able to reach nominal levels of power. Across all four response option conditions, 41.67% (20/48) were able to reach nominal levels of power.

Across all six response option conditions of the, 100% (24/48) were able to reach nominal levels of power.

Across all 18 item conditions for the 10% of Items with $\alpha_i = 0$ misspecification condition, 34.72% (25/72) were able to reach nominal levels of power while 51.39% (37/72) of the 36 item conditions reached nominal levels of power.

Across all two dimension conditions for the 10% of Items with $\alpha_i = 0$ misspecification condition, 51.39% (37/72) were able to reach nominal levels of power while 34.72% (25/72) of the three dimension conditions reached nominal levels of power.

Across all dichotomous response option conditions for the 20% of Items with $\alpha_i = 0$ misspecification condition, 97.92% (47/48) were able to reach nominal levels of power. Across all four response option conditions, 97.92% (47/48) were able to reach nominal levels of power. Across all six response option conditions of the, 100% (45/48) were able to reach nominal levels of power.

Across all 18 item conditions for the 20% of Items with $\alpha_i = 0$ misspecification condition, 97.22% (70/72) were able to reach nominal levels of power while 95.83% (69/72) of the 36 item conditions reached nominal levels of power.

Across all two-dimension conditions for the 20% of Items with $\alpha_i = 0$ misspecification condition, 98.61% (71/72) were able to reach nominal levels of power while 94.44% (68/72) of the three-dimension conditions reached nominal levels of power.

Across all dichotomous response option conditions for the 20% of Items with $\alpha_i = 0$ misspecification condition, 100% (48/48) were able to reach nominal levels of power. Across all four response option conditions, 100% (48/48) were able to reach nominal levels of power.

Across all six response option conditions of the, 95.83% (46/48) were able to reach nominal levels of power.

Across all 18 item conditions for the 20% of Items with $\alpha_i = 0$ misspecification condition, 100% (72/72) were able to reach nominal levels of power while 100% (70/72) of the 36 item conditions reached nominal levels of power.

Across all two-dimension conditions for the 20% of Items with $\alpha_i = 0$ misspecification condition, 100% (72/72) were able to reach nominal levels of power while 97.22% (70/72) of the three-dimension conditions reached nominal levels of power.

Across all dichotomous response option conditions for the 20% of Items with $\alpha_i = 0$ misspecification condition, 70.83% (34/48) were able to reach nominal levels of power. Across all four response option conditions, 87.5% (42/48) were able to reach nominal levels of power. Across all six response option conditions of the, 100% (43/48) were able to reach nominal levels of power.

Across all 18 item conditions for the 20% of Items with $\alpha_i = 0$ misspecification condition, 76.39% (55/72) were able to reach nominal levels of power while 88.89% (64/72) of the 36 item conditions reached nominal levels of power.

Across all two dimension conditions for the 10% of Items with $\alpha_i = 0$ misspecification condition, 97.22% (70/72) were able to reach nominal levels of power while 68.06% (49/72) of the three dimension conditions reached nominal levels of power.

Across all dichotomous response option conditions for the 1 factor correlation misspecified ($\rho_{\theta_1\theta_2} = 0$) and 10% of Items with $\alpha_i = 0$ misspecification condition, 72.92% (35/48) were able to reach nominal levels of power. Across all four response option conditions,

97.92% (47/48) were able to reach nominal levels of power. Across all six response option conditions of the, 100% (48/48) were able to reach nominal levels of power.

Across all 18 item conditions for the 1 factor correlation misspecified ($\rho_{\theta_1\theta_2} = 0$) and 10% of Items with $\alpha_i = 0$ misspecification condition, 81.94% (59/72) were able to reach nominal levels of power while 98.61% (71/72) of the 36 item conditions reached nominal levels of power.

Across all two-dimension conditions for the 1 factor correlation misspecified ($\rho_{\theta_1\theta_2} = 0$) and 10% of Items with $\alpha_i = 0$ misspecification condition, 93.06% (67/72) were able to reach nominal levels of power while 87.50% (63/72) of the three-dimension conditions reached nominal levels of power.

Across all dichotomous response option conditions for the 1 factor correlation misspecified ($\rho_{\theta_1\theta_2} = 0$) and 10% of Items with $\alpha_i = 0$ misspecification condition, 72.92% (35/48) were able to reach nominal levels of power. Across all four response option conditions, 97.92% (47/48) were able to reach nominal levels of power. Across all six response option conditions of the, 100% (48/48) were able to reach nominal levels of power.

Across all 18 item conditions for the 1 factor correlation misspecified ($\rho_{\theta_1\theta_2} = 0$) and 10% of Items with $\alpha_i = 0$ misspecification condition, 84.72% (61/72) were able to reach nominal levels of power while 97.22% (70/72) of the 36 item conditions reached nominal levels of power.

Across all two-dimension conditions for the 1 factor correlation misspecified ($\rho_{\theta_1\theta_2} = 0$) and 10% of Items with $\alpha_i = 0$ misspecification condition, 95.83% (69/72) were able to reach nominal levels of power while 86.11% (62/72) of the three-dimension conditions reached nominal levels of power.

Across all dichotomous response option conditions for the 1 factor correlation misspecified ($\rho_{\theta_1\theta_2} = 0$) and 10% of Items with $\alpha_i = 0$ misspecification condition, 35.42%

(17/48) were able to reach nominal levels of power. Across all four response option conditions, 41.67% (20/48) were able to reach nominal levels of power. Across all six response option conditions of the, 100% (22/48) were able to reach nominal levels of power.

Across all 18 item conditions for the 1 factor correlation misspecified ($\rho_{\theta_1\theta_2} = 0$) and 10% of Items with $\alpha_i = 0$ misspecification condition, 33.33% (24/72) were able to reach nominal levels of power while 48.61% (35/72) of the 36 item conditions reached nominal levels of power.

Across all two dimension conditions for the 1 factor correlation misspecified ($\rho_{\theta_1\theta_2} = 0$) and 10% of Items with $\alpha_i = 0$ misspecification condition, 48.61% (35/72) were able to reach nominal levels of power while 33.33% (24/72) of the three dimension conditions reached nominal levels of power.

Across all dichotomous response option conditions for the 1 factor correlation misspecified ($\rho_{\theta_1\theta_2} = 0$) and 20% of Items with $\alpha_i = 0$ misspecification condition, 100% (48/48) were able to reach nominal levels of power. Across all four response option conditions, 100% (48/48) were able to reach nominal levels of power. Across all six response option conditions of the, 100% (48/48) were able to reach nominal levels of power.

Across all 18 item conditions for the 1 factor correlation misspecified ($\rho_{\theta_1\theta_2} = 0$) and 20% of Items with $\alpha_i = 0$ misspecification condition, 100% (72/72) were able to reach nominal levels of power while 100% (72/72) of the 36 item conditions reached nominal levels of power.

Across all two-dimension conditions for the 1 factor correlation misspecified ($\rho_{\theta_1\theta_2} = 0$) and 20% of Items with $\alpha_i = 0$ misspecification condition, 100% (72/72) were able to reach nominal levels of power while 100% (72/72) of the three-dimension conditions reached nominal levels of power.

Across all dichotomous response option conditions for the 1 factor correlation misspecified ($\rho_{\theta_1\theta_2} = 0$) and 20% of Items with $\alpha_i = 0$ misspecification condition, 100% (48/48) were able to reach nominal levels of power. Across all four response option conditions, 100% (48/48) were able to reach nominal levels of power. Across all six response option conditions of the, 100% (48/48) were able to reach nominal levels of power.

Across all 18 item conditions for the 1 factor correlation misspecified ($\rho_{\theta_1\theta_2} = 0$) and 20% of Items with $\alpha_i = 0$ misspecification condition, 100% (72/72) were able to reach nominal levels of power while 100% (72/72) of the 36 item conditions reached nominal levels of power.

Across all two-dimension conditions for the 1 factor correlation misspecified ($\rho_{\theta_1\theta_2} = 0$) and 20% of Items with $\alpha_i = 0$ misspecification condition, 100% (72/72) were able to reach nominal levels of power while 100% (72/72) of the three-dimension conditions reached nominal levels of power.

Across all dichotomous response option conditions for the 1 factor correlation misspecified ($\rho_{\theta_1\theta_2} = 0$) and 20% of Items with $\alpha_i = 0$ misspecification condition, 64.58% (31/48) were able to reach nominal levels of power. Across all four response option conditions, 81.25% (39/48) were able to reach nominal levels of power. Across all six response option conditions of the, 83.33% (40/48) were able to reach nominal levels of power.

Across all 18 item conditions for the 1 factor correlation misspecified ($\rho_{\theta_1\theta_2} = 0$) and 20% of Items with $\alpha_i = 0$ misspecification condition, 69.44% (50/72) were able to reach nominal levels of power while 83.33% (60/72) of the 36 item conditions reached nominal levels of power.

Across all two dimension conditions for the 1 factor correlation misspecified ($\rho_{\theta_1\theta_2} = 0$) and 20% of Items with $\alpha_i = 0$ misspecification condition, 98.61% (71/72) were able to reach

nominal levels of power while 54.17% (39/72) of the three dimension conditions reached nominal levels of power.

Across all dichotomous response option conditions for the 2 factor correlation misspecified ($\rho_{\theta_1\theta_2} = 0$) and 10% of Items with $\alpha_i = 0$ misspecification condition, 66.67% (16/24) were able to reach nominal levels of power. Across all four response option conditions, 95.83% (23/24) were able to reach nominal levels of power. Across all six response option conditions of the, 100% (24/24) were able to reach nominal levels of power.

Across all 18 item conditions for the 2 factor correlation misspecified ($\rho_{\theta_1\theta_2} = 0$) and 10% of Items with $\alpha_i = 0$ misspecification condition, 77.78% (28/36) were able to reach nominal levels of power while 97.22% (35/36) of the 36 item conditions reached nominal levels of power.

Across all three-dimension conditions for the 2 factor correlation misspecified ($\rho_{\theta_1\theta_2} = 0$) and 10% of Items with $\alpha_i = 0$ misspecification condition, 87.5% (63/72) were able to reach nominal levels of power.

Across all dichotomous response option conditions for the 2 factor correlation misspecified ($\rho_{\theta_1\theta_2} = 0$) and 10% of Items with $\alpha_i = 0$ misspecification condition, 58.33% (14/24) were able to reach nominal levels of power. Across all four response option conditions, 95.83% (23/24) were able to reach nominal levels of power. Across all six response option conditions of the, 100% (24/24) were able to reach nominal levels of power.

Across all 18 item conditions for the 2 factor correlation misspecified ($\rho_{\theta_1\theta_2} = 0$) and 10% of Items with $\alpha_i = 0$ misspecification condition, 75% (27/36) were able to reach nominal levels of power while 94.44% (34/36) of the 36 item conditions reached nominal levels of power.

Across all three-dimension conditions for the two-factor correlations misspecified ($\rho_{\theta_1\theta_2} = 0$) and 10% of Items with $\alpha_i = 0$ misspecification condition, 84.72% (61/72) were able to reach nominal levels of power.

Across all dichotomous response option conditions for the 2 factor correlations misspecified ($\rho_{\theta_1\theta_2} = 0$) and 10% of Items with $\alpha_i = 0$ misspecification condition, 33.33% (8/24) were able to reach nominal levels of power. Across all four response option conditions, 33.33% (8/24) were able to reach nominal levels of power. Across all six response option conditions of the, 100% (8/24) were able to reach nominal levels of power.

Across all 18 item conditions for the 2 factor correlation misspecified ($\rho_{\theta_1\theta_2} = 0$) and 10% of Items with $\alpha_i = 0$ misspecification condition, 33.33% (12/36) were able to reach nominal levels of power while 33.33% (12/36) of the 36 item conditions reached nominal levels of power.

Across all three dimension conditions for the 2 factor correlation misspecified ($\rho_{\theta_1\theta_2} = 0$) and 10% of Items with $\alpha_i = 0$ misspecification condition, 48.61% (24/72) were able to reach nominal levels of power while 48.61% (35/72) of the three dimension conditions reached nominal levels of power.

Across all dichotomous response option conditions for the 2 factor correlation misspecified ($\rho_{\theta_1\theta_2} = 0$) and 20% of Items with $\alpha_i = 0$ misspecification condition, 100% (24/24) were able to reach nominal levels of power. Across all four response option conditions, 100% (24/24) were able to reach nominal levels of power. Across all six response option conditions of the, 100% (24/24) were able to reach nominal levels of power.

Across all 18 item conditions for the 2 factor correlation misspecified ($\rho_{\theta_1\theta_2} = 0$) and 20% of Items with $\alpha_i = 0$ misspecification condition, 100% (36/36) were able to reach nominal levels of power while 100% (36/36) of the 36 item conditions reached nominal levels of power.

Across all three dimension conditions for the 2 factor correlation misspecified ($\rho_{\theta_1\theta_2} = 0$) and 20% of Items with $\alpha_i = 0$ misspecification condition, 100% (72/72) were able to reach nominal levels of power.

Across all dichotomous response option conditions for the 2 factor correlation misspecified ($\rho_{\theta_1\theta_2} = 0$) and 20% of Items with $\alpha_i = 0$ misspecification condition, 100% (24/24) were able to reach nominal levels of power. Across all four response option conditions, 100% (24/24) were able to reach nominal levels of power. Across all six response option conditions of the, 100% (24/24) were able to reach nominal levels of power.

Across all 18 item conditions for the 2 factor correlation misspecified ($\rho_{\theta_1\theta_2} = 0$) and 20% of Items with $\alpha_i = 0$ misspecification condition, 100% (36/36) were able to reach nominal levels of power while 100% (36/36) of the 36 item conditions reached nominal levels of power.

Across all three dimension conditions for the 2 factor correlation misspecified ($\rho_{\theta_1\theta_2} = 0$) and 20% of Items with $\alpha_i = 0$ misspecification condition, 100% (72/72) were able to reach nominal levels of power.

Across all dichotomous response option conditions for the 2 factor correlation misspecified ($\rho_{\theta_1\theta_2} = 0$) and 20% of Items with $\alpha_i = 0$ misspecification condition, 33% (8/24) were able to reach nominal levels of power. Across all four response option conditions, 62.5% (15/24) were able to reach nominal levels of power. Across all six response option conditions of the, 100% (16/24) were able to reach nominal levels of power.

Across all 18 item conditions for the 2 factor correlation misspecified ($\rho_{\theta_1\theta_2} = 0$) and 20% of Items with $\alpha_i = 0$ misspecification condition, 41.67% (15/36) were able to reach nominal levels of power while 66.67% (24/36) of the 36 item conditions reached nominal levels of power.

Across all three dimension conditions for the 2 factor correlation misspecified ($\rho_{\theta_1\theta_2} = 0$) and 20% of Items with $\alpha_i = 0$ misspecification condition, 54.17% (39/72) were able to reach nominal levels of power.

CHAPTER 10

DISCUSSION

The use of multidimensional item response theory in psychology has increased dramatically over the past decade thanks, in part, to recent advances in statistical theory and statistical computing. With the development of the M_2 family of statistics by Maydeu-Olivares and Joe (2005; 2006) and its computational implementation in the R package "mirt" (Chalmers, 2012), researchers are now able to compute a variety of M_2-based global model fit statistics (e.g. CFI, TLI, RMSEA, SRMSR) across a large gamut of MIRT models. Despite this, there is little clarity for how to best make adequate model selection decisions when using these fit statistics. In practice, researchers using MIRT models in conjunction with M2-based statistics have turned to the commonly used Hu and Bentler cut-offs for evaluating these M_2 fit statistics despite there being little evidence of their performance for adequate model selection (see Kilgus et al., 2018, Hyatt et al., 2022, and Harris et al., 2020 for examples in school, clinical, and organizational psychology, respectively).

The purpose of the present simulation study was to provide guidance for researchers using M2 statistics to make MIRT model selection decisions. I accomplished this by examining the behavior of commonly used global model fit statistic cut-offs for determining the correct DGP as well as detecting varying forms of model misspecification for MIRT model selection. Results indicate that existing global model fit statistic cut-offs have mixed success for determining whether a candidate model is the correct or "true" DGP. Of particular note was the poor performance of SRMSR for detecting model misspecification in comparison to CFI, TLI,

and RMSEA. In the event researchers have conflicting findings from fit statistic values they derive in their own data, these results would suggest discarding the SRMSR in favor of the other statistics if misfit is of concern. Overall, the results from this study indicate that researchers should be cautious when using the common Hu and Bentler model-data fit index cut-offs to evaluate multidimensional IRT models given the potential study design factors that could influence the ability to make correct model selection decisions and inferences.

Despite the large number of conditions, response models, and misspecifications examined in this study, there remain several open questions for MIRT model selection should explore. Future research on CFA and SEM model selection has begun shifting from a "one-size-fits-all" approach to viewing cut-offs as being dependent on the research process and question. Such *dynamic* fit index cut-offs (Wolf & McNeish, 2021) could provide a means with which to reduce researcher degrees of freedom (Gelman & Loken, 2013) and increase transparency with respect to reporting what MIRT models a researcher wishes to examine. DFIs serve as a form of simulation-based inference, whereby researcher simulate potentially plausible alternative models in addition to the theoretical model of interest and simulate cut-offs beforehand to aid with inference and model selection. Results from the present work suggest this would be a worthwhile pursuit given the mixed performance of global cut-offs. Lastly, one area of work that has recently taken shape is the use of *predictive* fit index measures (Stenhaug & Domingue, 2022) for IRT models that evaluate a model's ability to predict both person and item parameters, ignoring its ability to recover the structural characteristics of a latent variable. It would be beneficial to examine their performance for multidimensional models and compare their performance directly with M2 statistics.

REFERENCES

Cai, L., & Hansen, M. (2013). Limited□information goodness□of□fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology, 66*(2), 245-276. https://doi.org/10.1111/j.2044-8317.2012.02050.x

Carter, N. T., Dalal, D. K., Guan, L., LoPilato, A. C., & Withrow, S. A. (2017). Item response theory scoring and the detection of curvilinear relationships. *Psychological Methods, 22*(1), 191-203. https://psycnet.apa.org/doi/10.1037/met0000101

Chalmers, R.P. (2017). model fit using M2 and anova [Online forum comment]. Message□ posted to https://groups.google.com/g/mirt-package/c/J695Rb77viA

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment, 8*(4), 341–349. https://doi.org/10.1037/1040-3590.8.4.341

Fletcher, R. (2013). Practical methods of optimization. John Wiley & Sons.

Hu, L. T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological methods, 3*(4), 424-453. https://psycnet.apa.org/doi/10.1037/1082-989X.3.4.424

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal, 6*(1), 1-55. https://doi.org/10.1080/10705519909540118

Kang, T., Cohen, A. S., & Sung, H. J. (2005, March). IRT model selection methods for polytomous items. In annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Kang, T., Cohen, A. S., & Sung, H. J. (2009). Model selection indices for polytomous items. *Applied Psychological Measurement, 33*(7), 499-518. https://doi.org/10.1177/0146621608327800

Kieftenbeld, V., & Natesan, P. (2012). Recovery of graded response model parameters: A comparison of marginal maximum likelihood and Markov chain Monte Carlo estimation. *Applied Psychological Measurement, 36*(5), 399-419. https://doi.org/10.1177/0146621612446170

Lang, J. W., & Tay, L. (2020). The science and practice of item response theory in organizations. *Annual Review of Organizational Psychology and Organizational Behavior*, *8*, 311-338. https://doi.org/10.1146/annurev-orgpsych-012420-061705

Lilienfeld, S. O. (2018). The multidimensional nature of psychopathy: Five recommendations for research. *Journal of Psychopathology and Behavioral Assessment, 40*(1), 79-85. https://doi.org/10.1007/s10862-018-9657-7

Liu, C. W., & Chalmers, R. P. (2018). Fitting item response unfolding models to Likert-scale data using mirt in R. *PloS one, 13*(5), e0196292. https://doi.org/10.1371/journal.pone.0196292

Matlock, K. L., Turner, R. C., & Gitchel, W. D. (2018). A study of reverse-worded matched item pairs using the generalized partial credit and nominal response models. *Educational and psychological measurement, 78*(1), 103-127. https://doi.org/10.1177/0013164416670211

Maydeu-Olivares, A., & Joe, H. (2005). Limited-and full-information estimation and goodness-of-fit testing in 2 n contingency tables: A unified framework. *Journal of the American Statistical Association, 100*(471), 1009-1020. https://doi.org/10.1198/016214504000002069

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika, 71*(4), 713-732. https://doi.org/10.1007/s11336-005-1295-9

Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research, 49*(4), 305-328. https://doi.org/10.1080/00273171.2014.911075

Morse, B. J., Johanson, G. A., & Griffeth, R. W. (2012). Using the graded response model to control spurious interactions in moderated multiple regression. *Applied Psychological Measurement, 36*(2), 122-146. https://doi.org/10.1177/0146621612438725

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *ETS Research Report Series, 1992*(1), i-30. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x

Muraki, E., & Carlson, J. E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement, 19*(1), 73-90. https://doi.org/10.1177/014662169501900109

Nye, C. D., Joo, S. H., Zhang, B., & Stark, S. (2020). Advancing and evaluating IRT model data fit indices in organizational research. *Organizational Research Methods, 23*(3), 457-486. https://doi.org/10.1177/1094428119833158

Read, T. R., & Cressie, N. A. (1988). Historical Perspective: Pearson's X 2 and the Loglikelihood Ratio Statistic G 2. In Goodness-of-Fit Statistics for Discrete Multivariate Data (pp. 133-153). Springer, New York, NY.

Reckase, M. D. (2009). Multidimensional item response theory models. In Multidimensional item response theory (pp. 79-112). Springer, New York, NY.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement, 24*(1), 3-32. https://doi.org/10.1177/01466216000241001

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*(4, Pt. 2), 100.

Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. A. (2006). Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology, 91*(1), 25-39. https://psycnet.apa.org/doi/10.1037/0021-9010.91.1.25

Sung, H. J., & Kang, T. (2006, April). Choosing a polytomous IRT model using Bayesian model selection methods. In National Council on Measurement in Education Annual Meeting (pp. 1-36).

Tay, L., Drasgow, F., Rounds, J., & Williams, B. A. (2009). Fitting measurement models to vocational interest data: Are dominance models ideal? *Journal of Applied Psychology, 94*(5), 1287-1304. https://psycnet.apa.org/doi/10.1037/a0015899

Tay, L., Meade, A. W., & Cao, M. (2015). An overview and practical guide to IRT measurement equivalence analysis. *Organizational Research Methods, 18*(1), 3-46. https://doi.org/10.1177/1094428114553062

Tellegen, A., & Waller, N. G. (2008). Exploring personality through test construction: Development of the Multidimensional Personality Questionnaire. The SAGE handbook of personality theory and assessment, 2, 261-292.

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology, 33*(4), 529-554. https://doi.org/10.1086/214483

Wang, W. C., & Wu, S. L. (2016). Confirmatory multidimensional IRT unfolding models for graded-response items. *Applied Psychological Measurement, 40*(1), 56-72. https://doi.org/10.1177/0146621615602855

Wang, C., Su, S., & Weiss, D. J. (2018). Robustness of parameter estimation to assumptions of normality in the multidimensional graded response model. *Multivariate Behavioral Research, 53*(3), 403-418. https://doi.org/10.1080/00273171.2018.1455572

Wetzel, E., & Hell, B. (2014). Multidimensional item response theory models in vocational interest measurement: An illustration using the AIST-R. *Journal of Psychoeducational Assessment, 32*(4), 342-355. https://doi.org/10.1177/0734282913508244

Table 1. *Summary of the Monte Carlo Simulation Design*

Factor	Number of Levels	Levels of Factor
Sample Size	4	150, 250, 500, 750, 1000
Number of Items	2	18, 36
Number of Factors	2	2, 3
Number of Response Options	3	2, 4, 6
IRT Model Studied	3	MGRM, MGGUM, MGPCM
% of Items with $\alpha_i = 0$	2	10%, 20%
Number of $\rho_{\theta_1\theta_2} = 0$	3	0, 1, 2
RMSEA Cutoffs	4	0.06 (Hu & Bentler, 1999); 0.089, 0.05, 0.05 / (K-1) (Maydeu-Olivares & Joe, 2014)
SRMSR Cutoffs	4	0.08 (Hu & Bentler, 1999); 0.05, 0.027, 0.027 / (K-1) (Maydeu-Olivares & Joe, 2014)
CFI Cutoffs	1	0.95 (Hu & Bentler, 1999)
TLI Cutoffs	1	0.95 (Hu & Bentler, 1999)

Note. MGRM=Multidimensional Graded Response Model; GGUM= Multidimensional Generalized Graded Unfolding Model; MGPCM=Multidimensional Generalized Partial Credit Model; the GRM is the polytomous case of the two-parameter logistic model; α_i is the discrimination parameter; $\rho_{\theta_1\theta_2}$ is the inter-factor correlation.

FIGURES

Figure 1. *RMSE Between Multidimensional Graded Response Model Sample and Population Item Parameters*

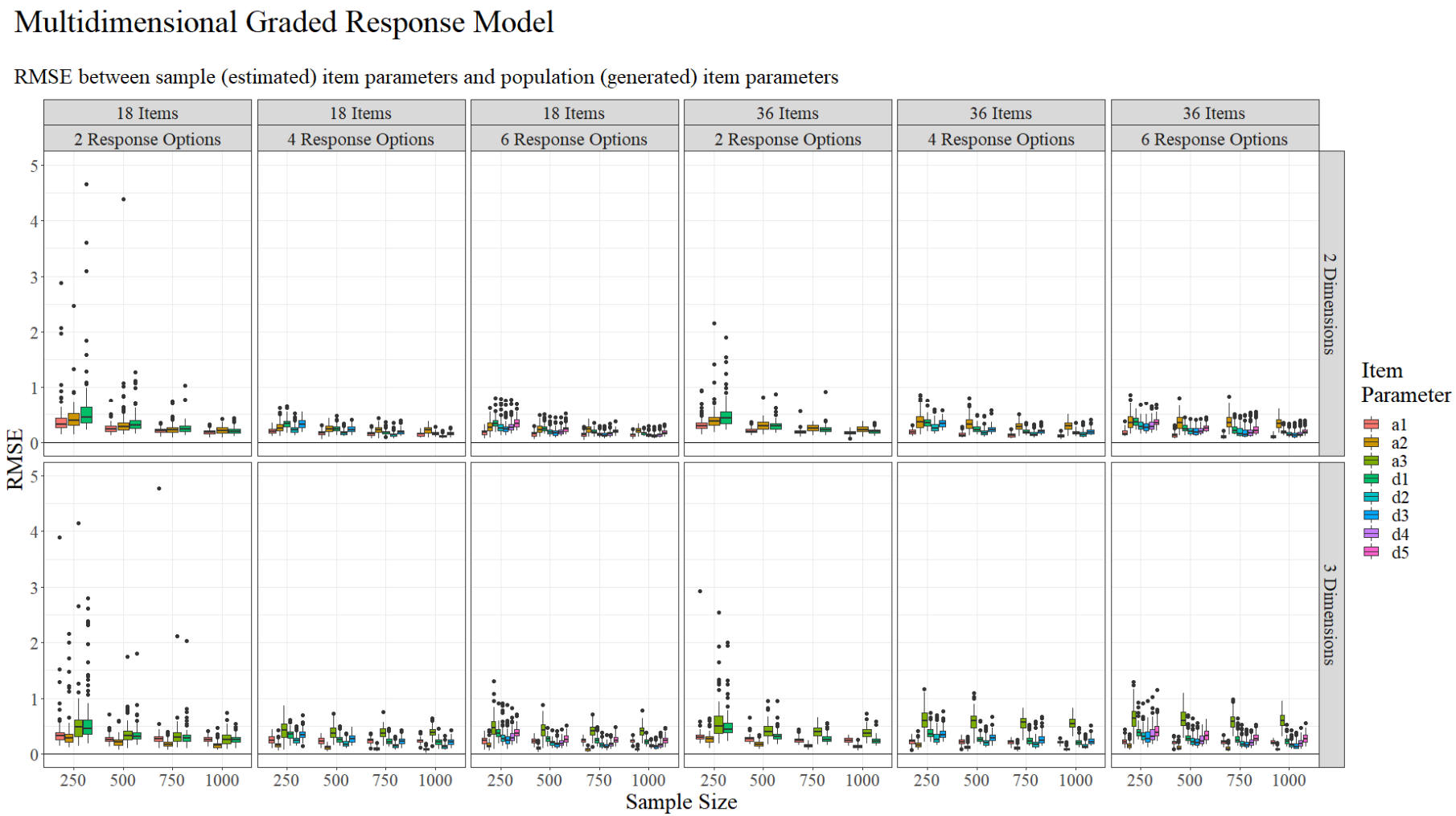

Figure 2. *Mean Bias Between Multidimensional Graded Response Model Sample and Population Item Parameters*

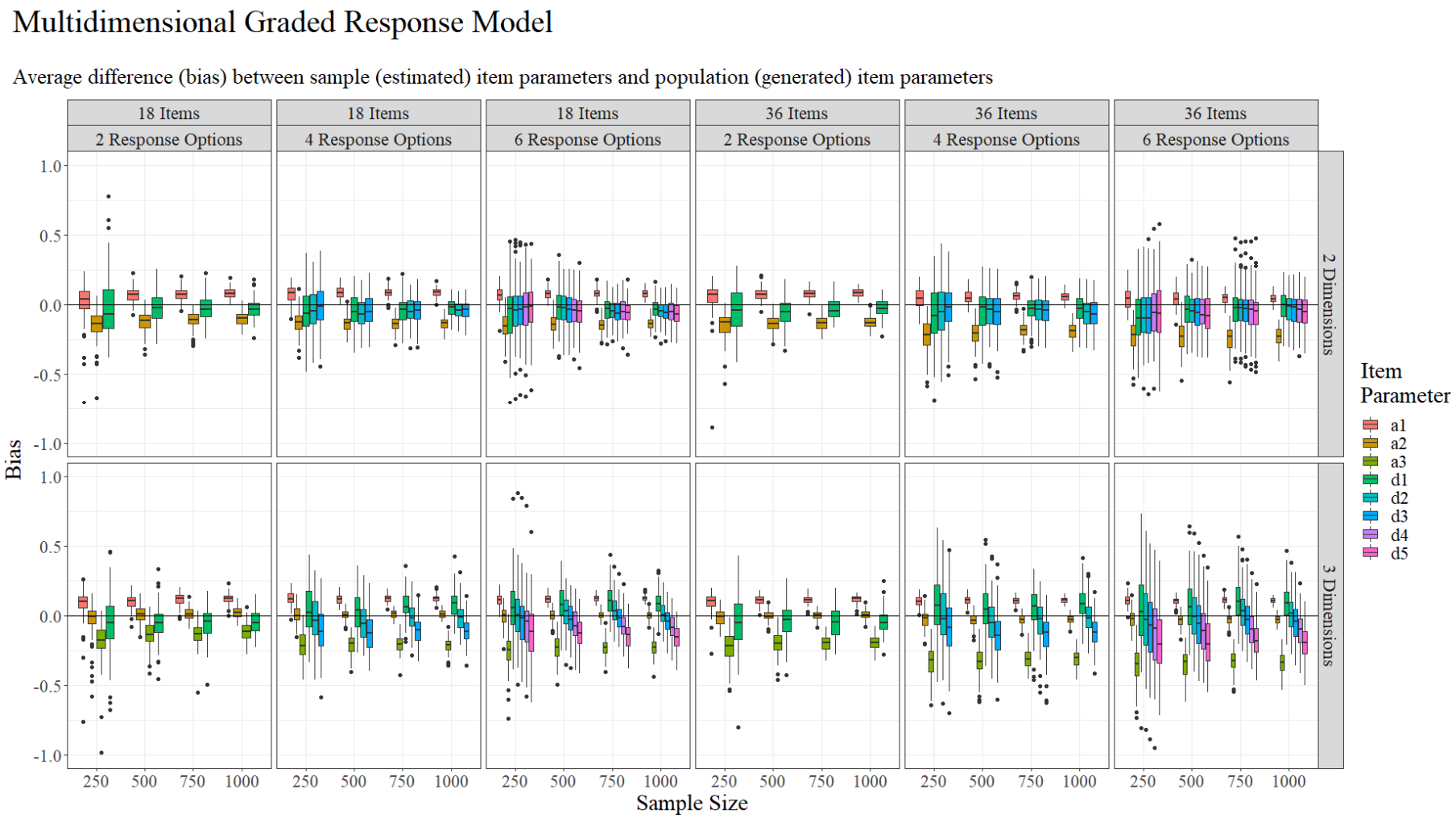

Figure 3. *Correlation Between Multidimensional Graded Response Model Sample and Population Item Parameters*

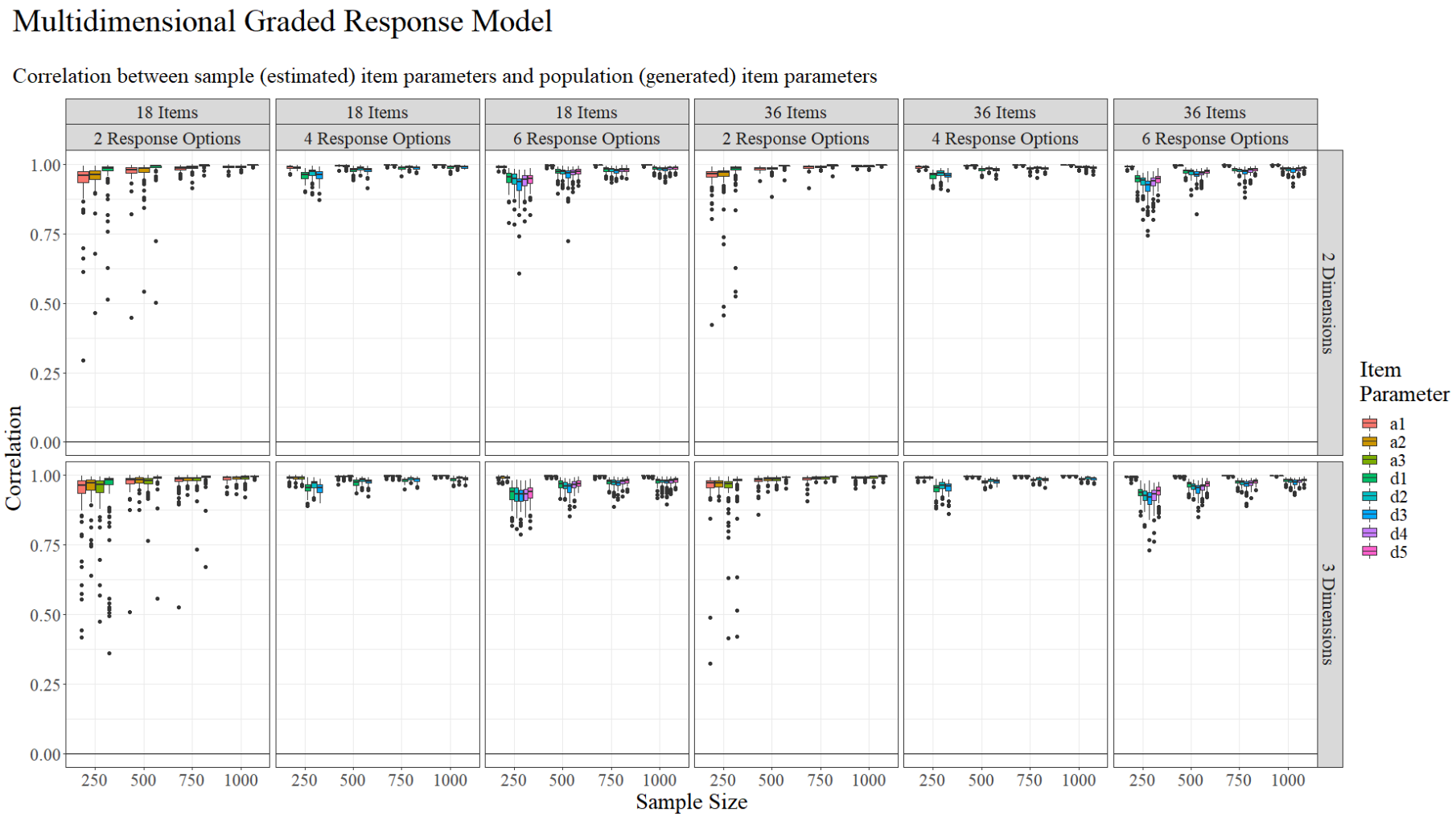

Figure 4. *RMSE Between Multidimensional Generalized Partial Credit Model Sample and Population Item Parameters*

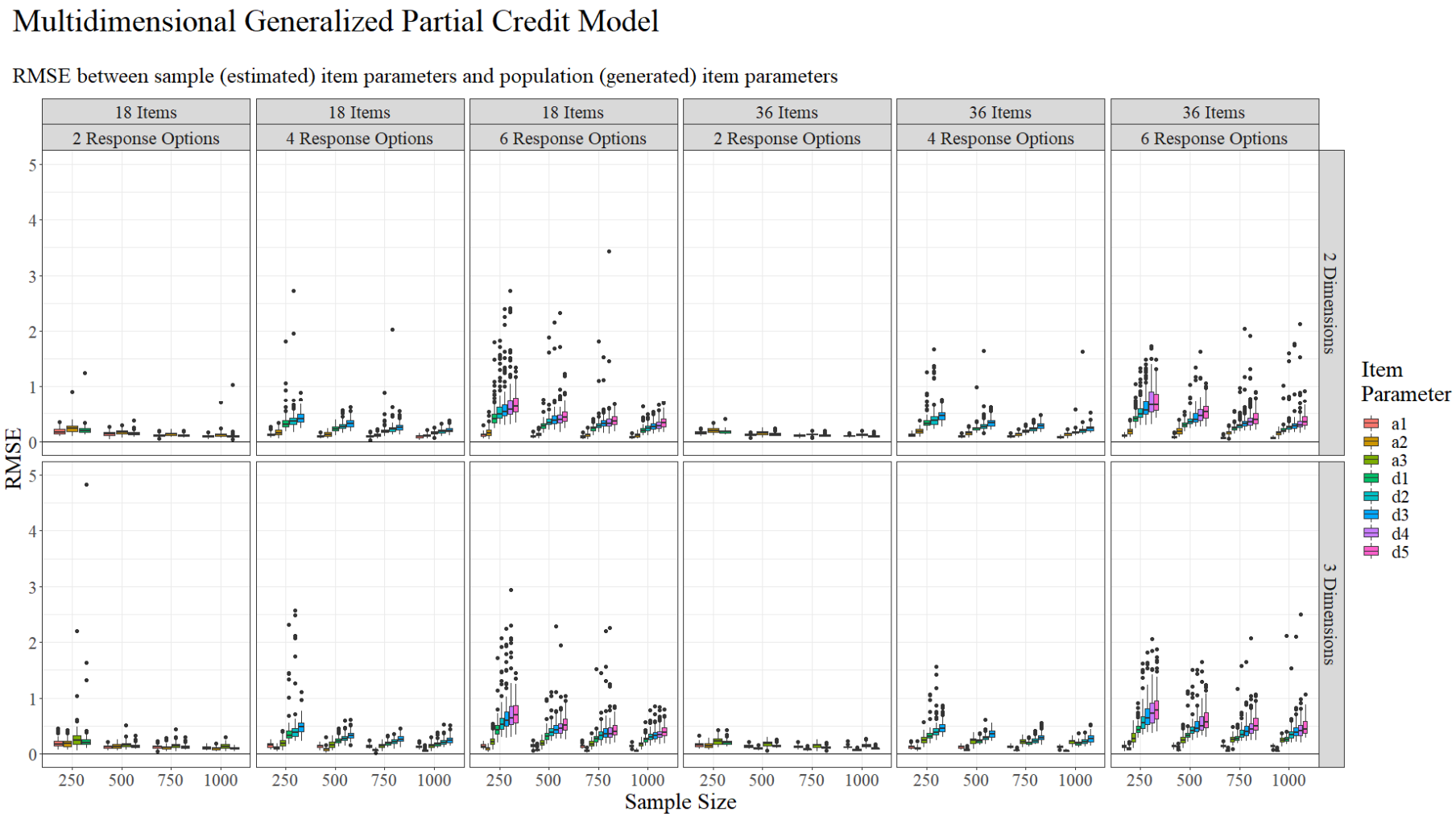

Figure 5. *Mean Bias Between Multidimensional Generalized Partial Credit Model Sample and Population Item Parameters*

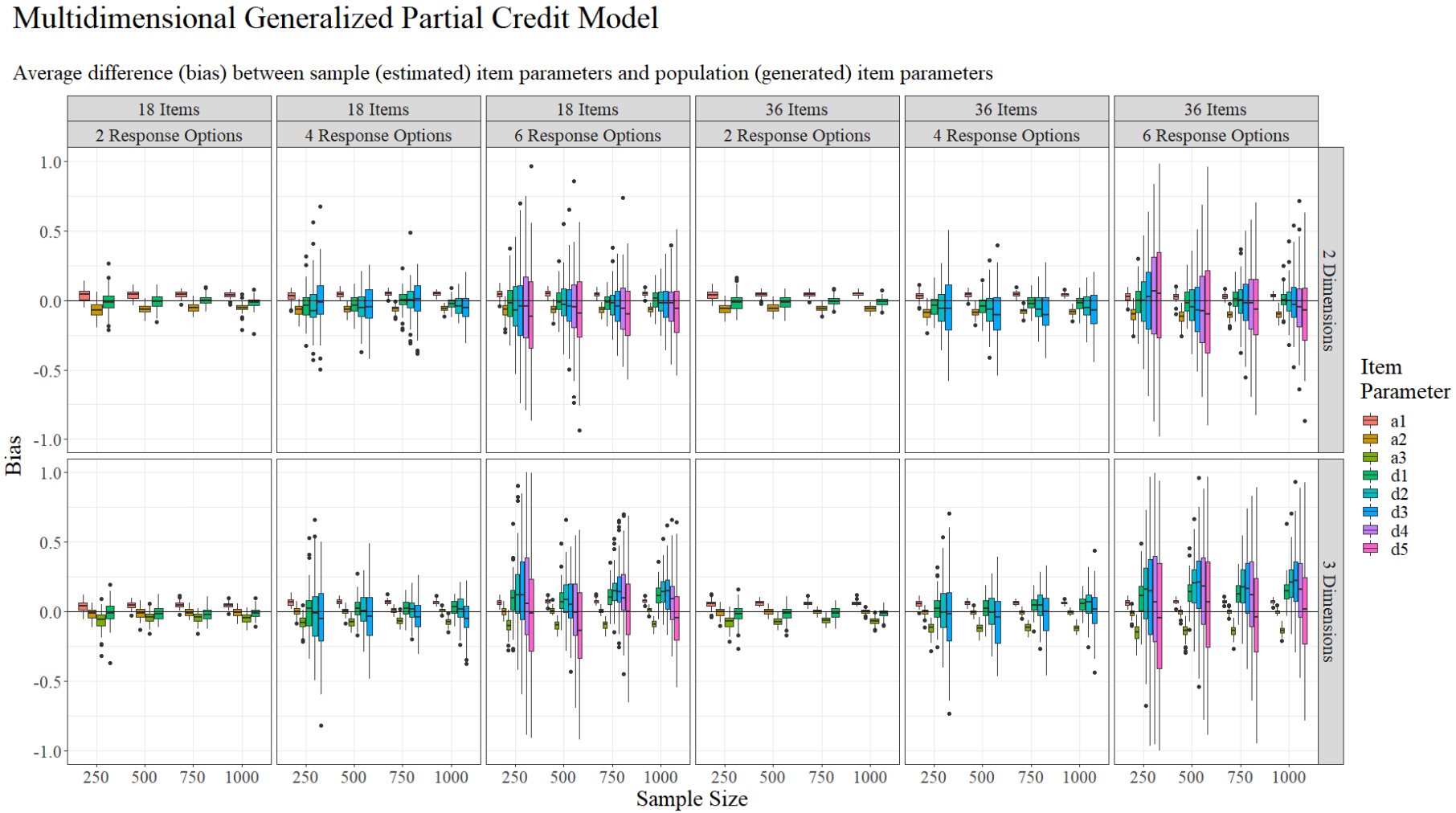

Figure 6. *Correlation Between Multidimensional Generalized Partial Credit Model Sample and Population Item Parameters*

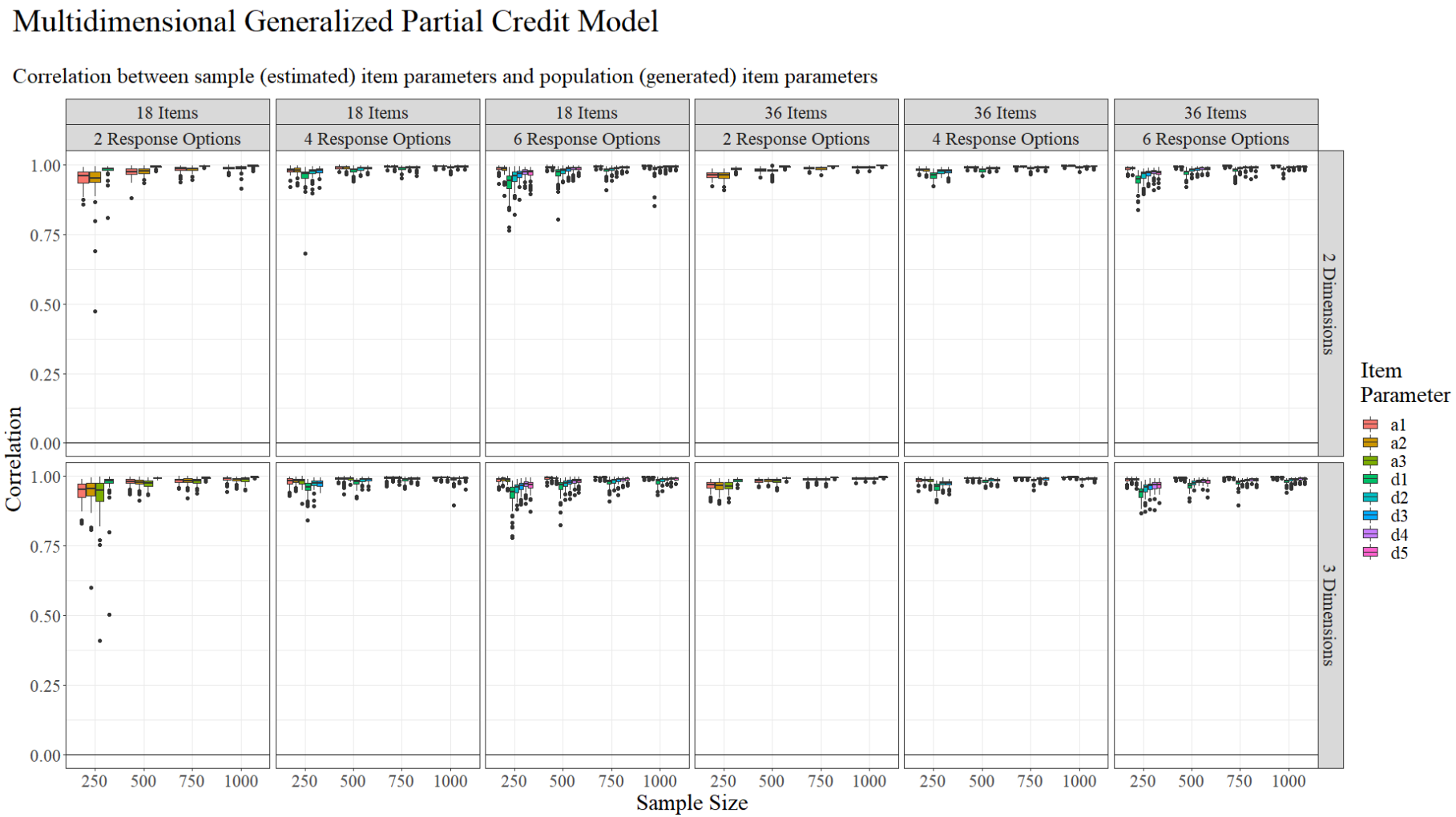

Figure 7. *RMSE Between Multidimensional Generalized Graded Unfolding Model Sample and Population Item Parameters*

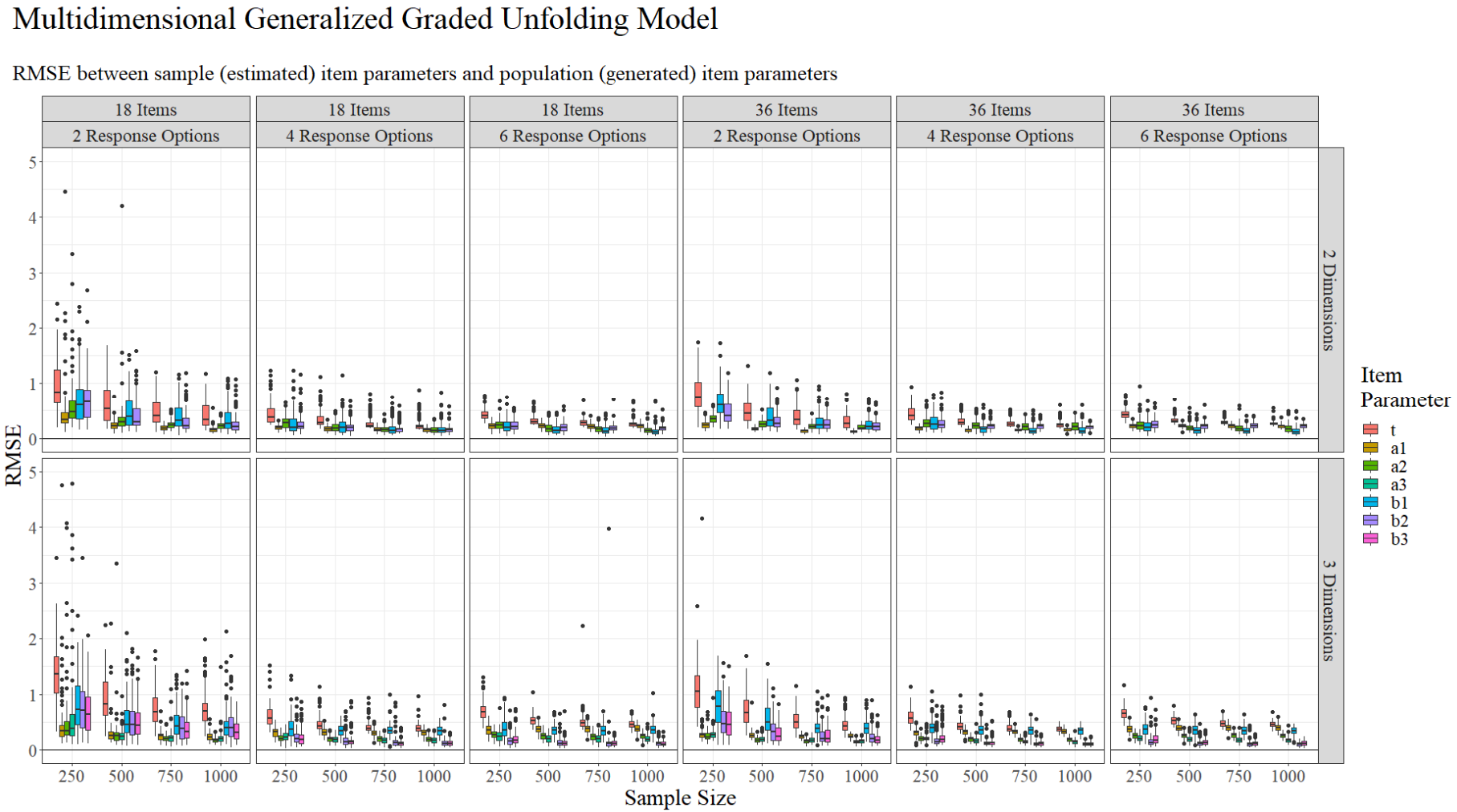

Figure 8. *Mean Bias Between Multidimensional Generalized Graded Unfolding Model Sample and Population Item Parameters*

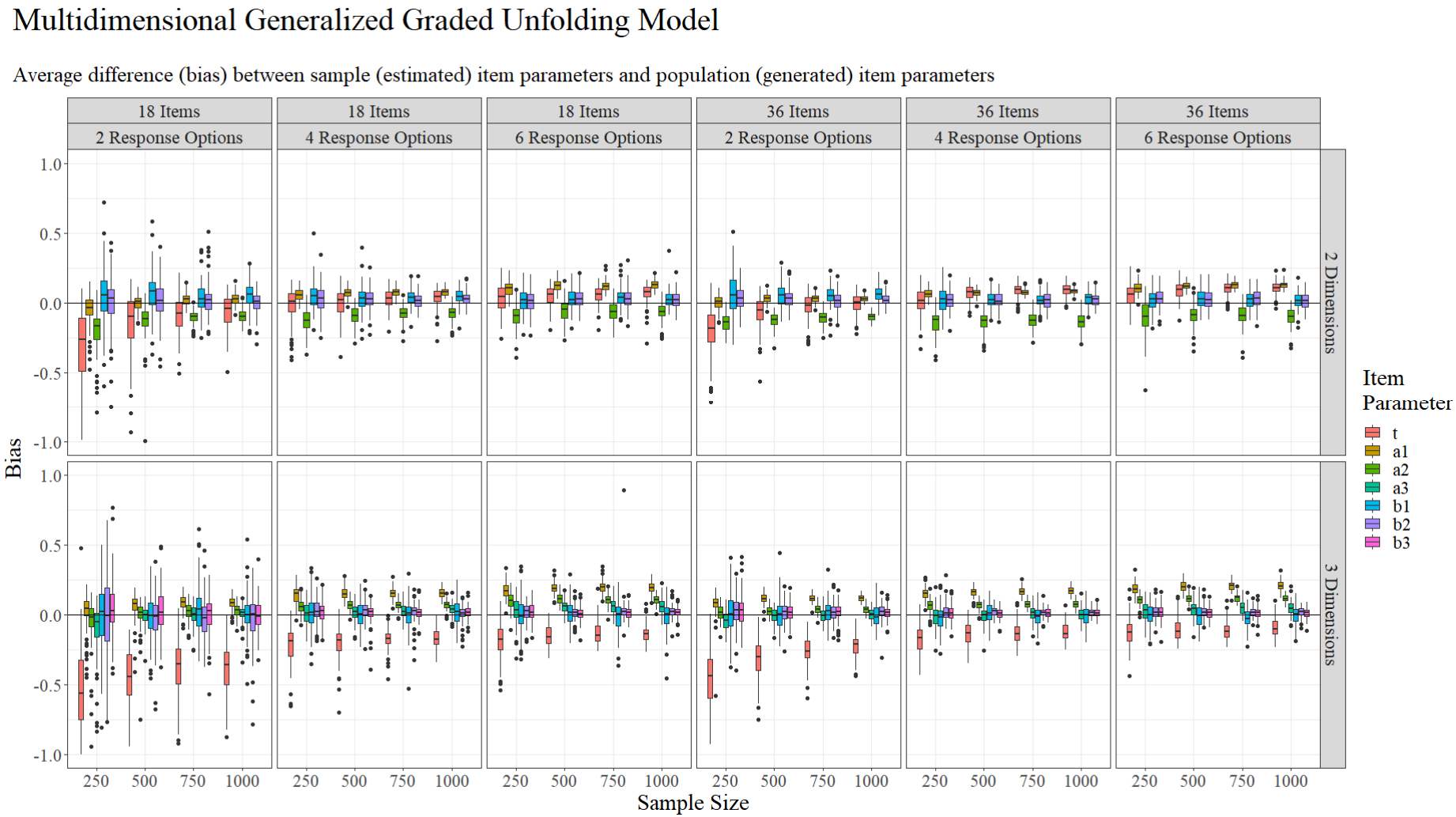

Figure 9. *Correlation Between Multidimensional Generalized Graded Unfolding Model Sample and Population Item Parameters*

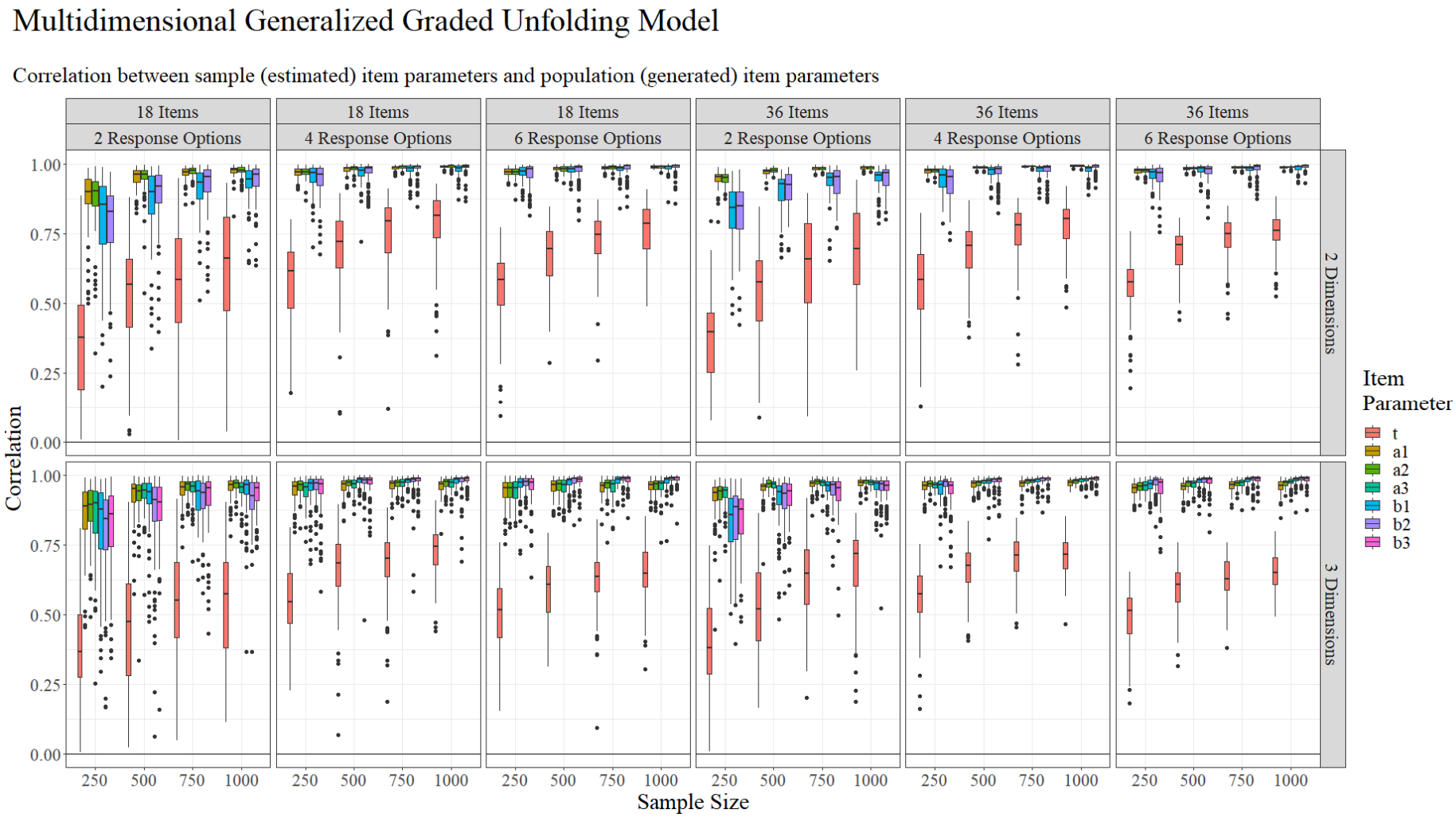

Figure 10. *Boxplot of Model Fit Statistics for the Multidimensional Graded Response Model*

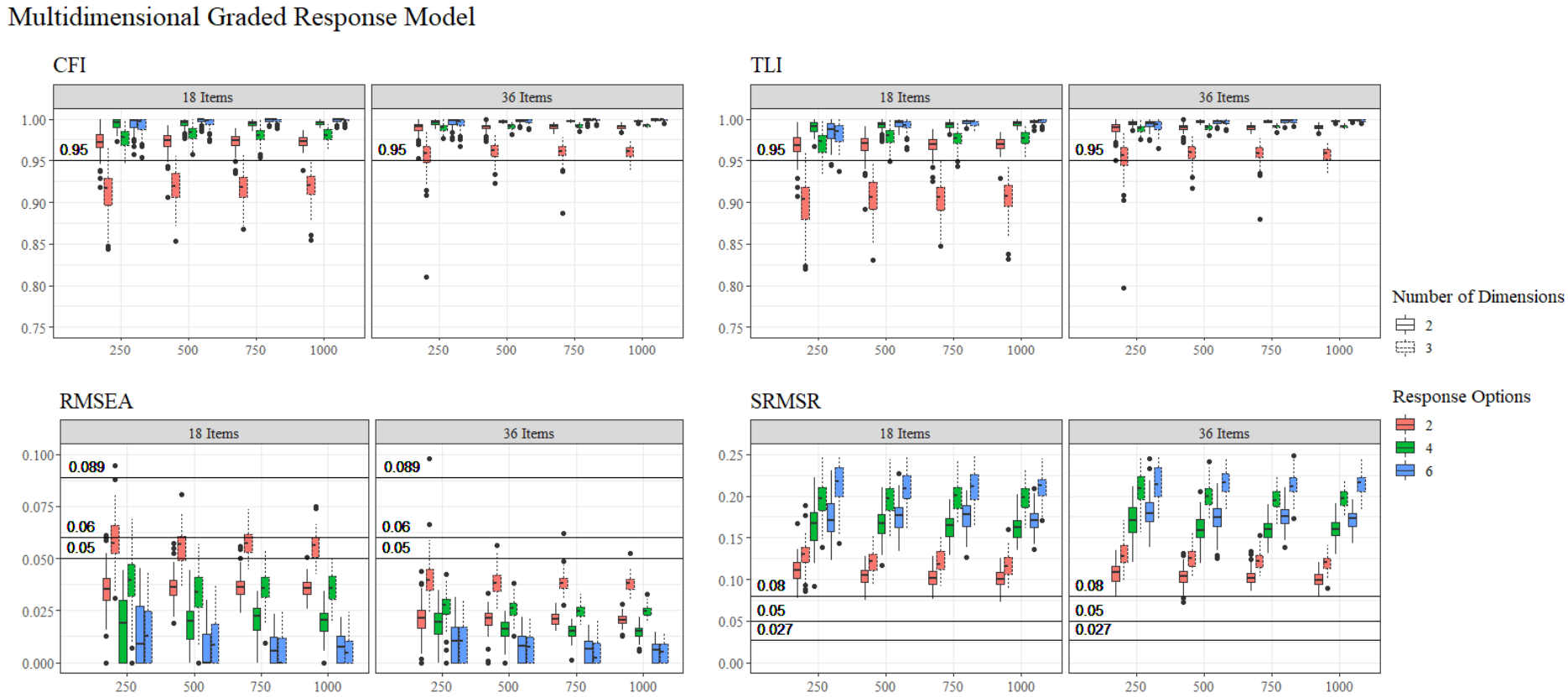

Figure 11. *Absolute Fit Statistics and Their Respective Cut-Offs for the Multidimensional Graded Response Model*

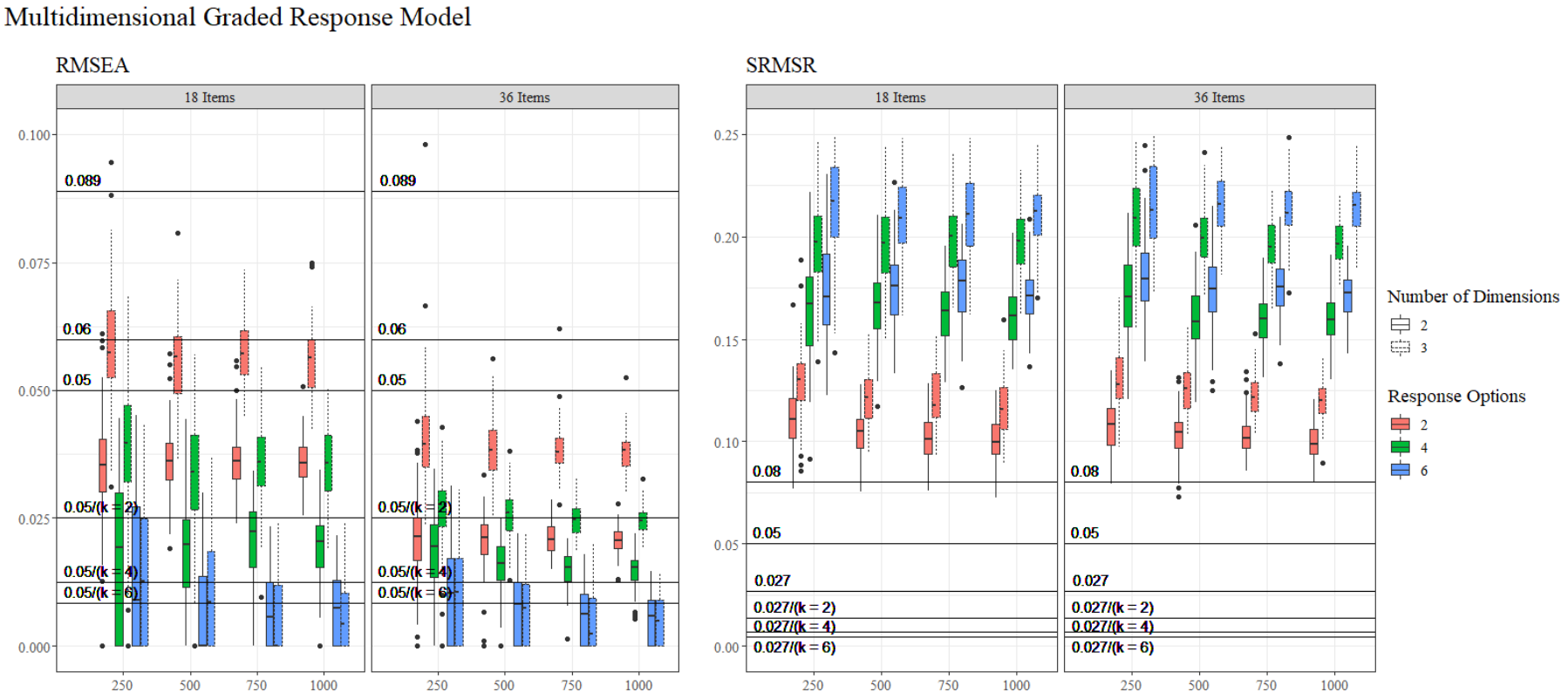

Figure 12. *Boxplot of Model Fit Statistics for the Multidimensional Generalized Partial Credit Model*

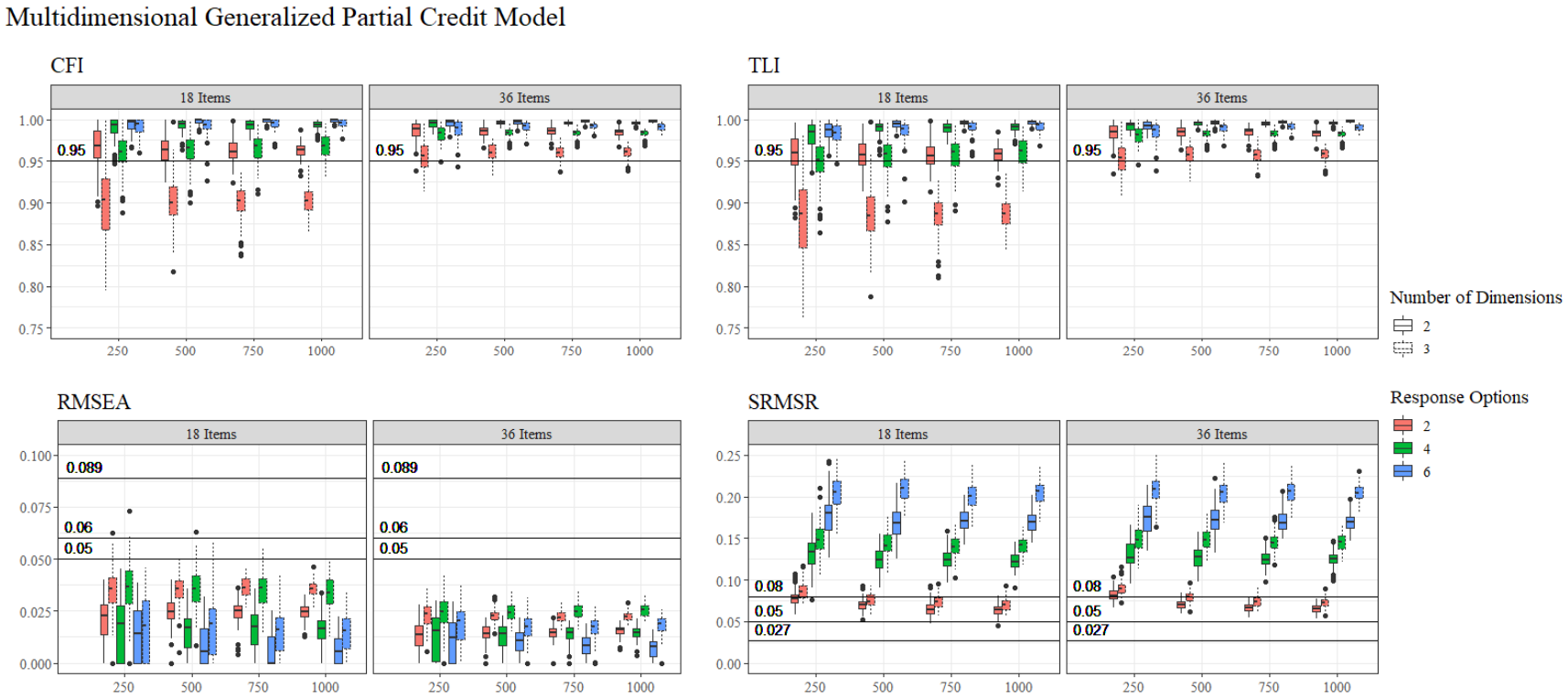

Figure 13. *Absolute Fit Statistics and Their Respective Cut-Offs for the Multidimensional Generalized Partial Credit Model*

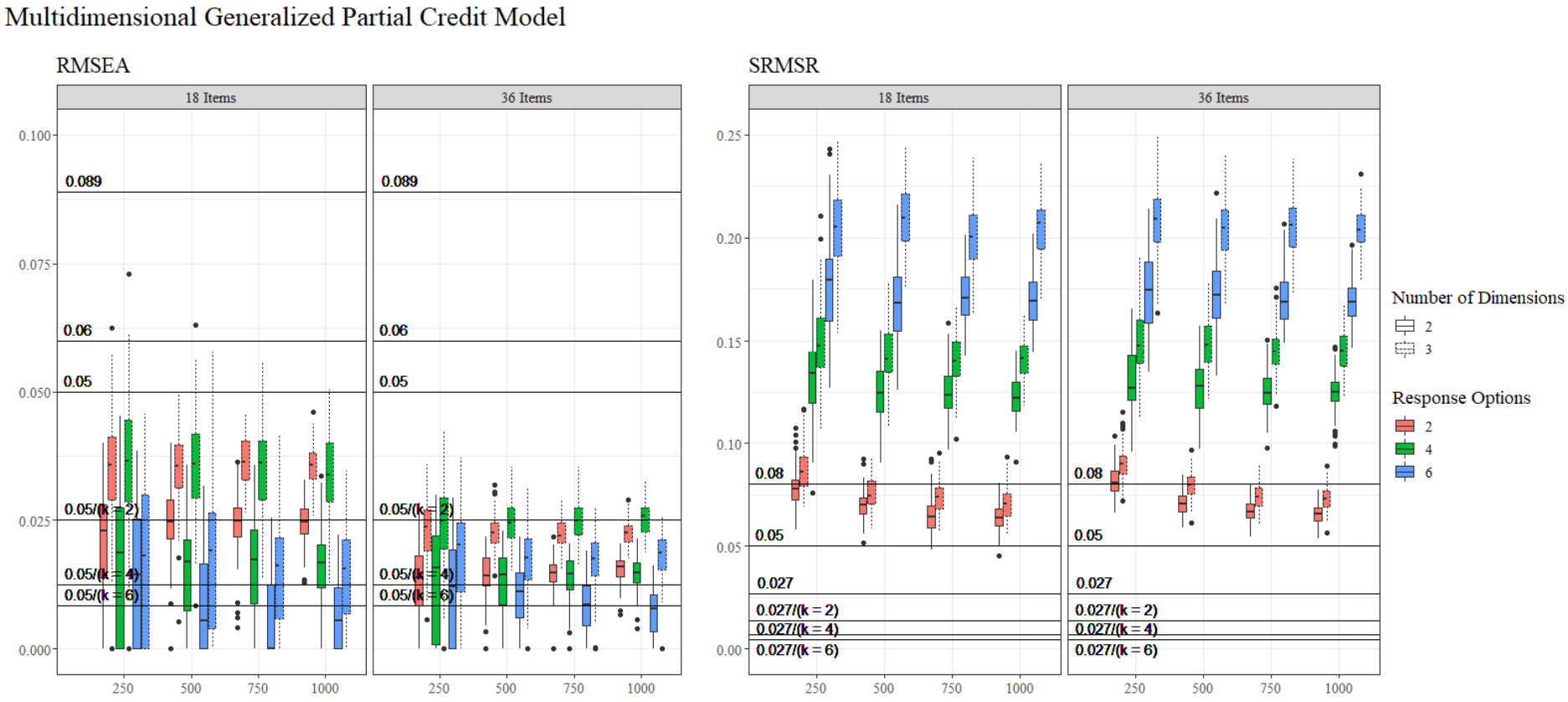

Figure 14. *Boxplot of Model Fit Statistics for the Multidimensional Generalized Graded Unfolding Model*

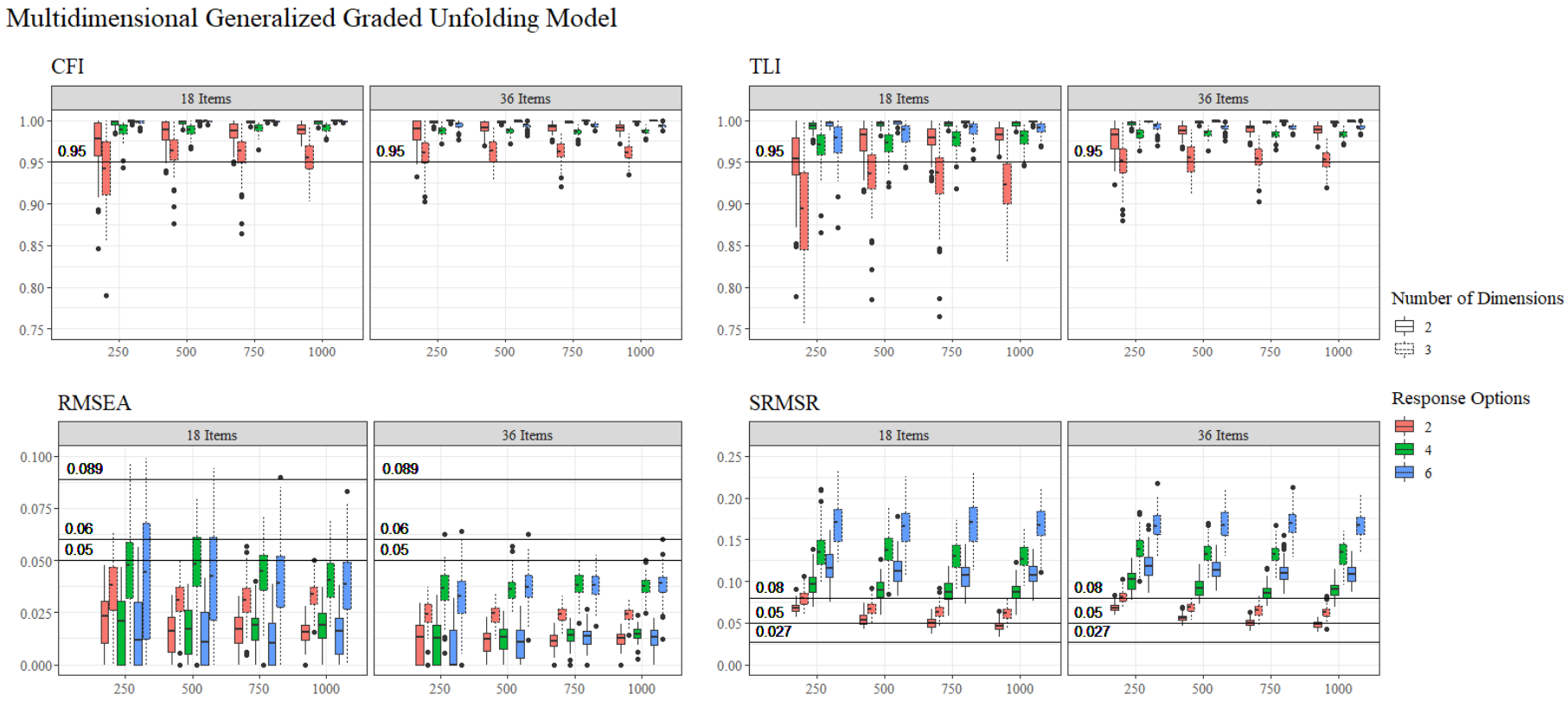

Figure 15. *Absolute Fit Statistics and Their Respective Cut-Offs for the Multidimensional Generalized Graded Unfolding Model*

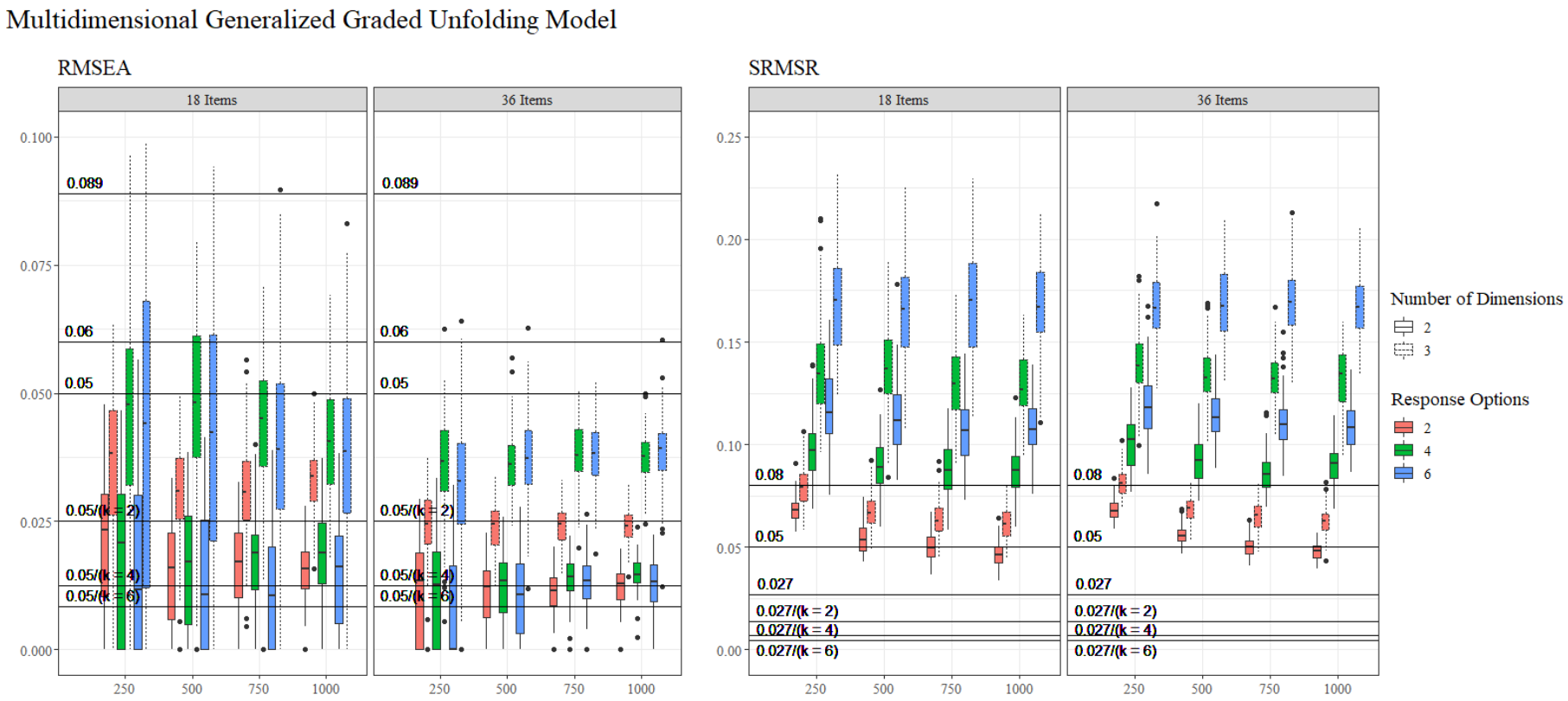

Figure 16. *Type 1 Error Rates for the Multidimensional Graded Response Model*

Figure 17. *Type 1 Error Rates for the Multidimensional Generalized Partial Credit Model*

Figure 18. *Type 1 Error Rates for the Multidimensional Generalized Graded Unfolding Model*

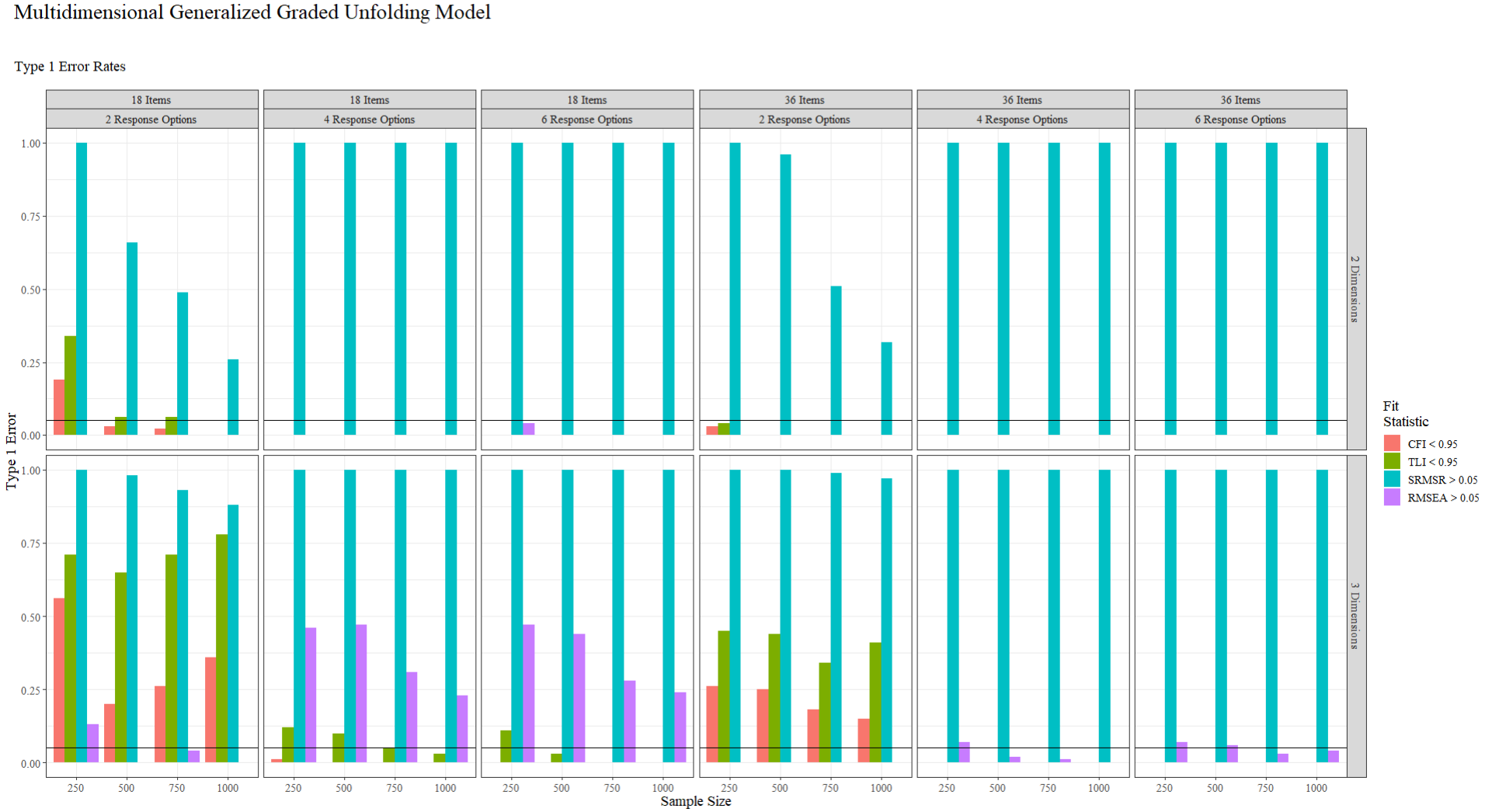

Figure 19. *Power Results for 1 Correlation Misspecified Condition (18 items)*

Figure 20. *Power Results for 1 Correlation Misspecified Condition (36 items)*

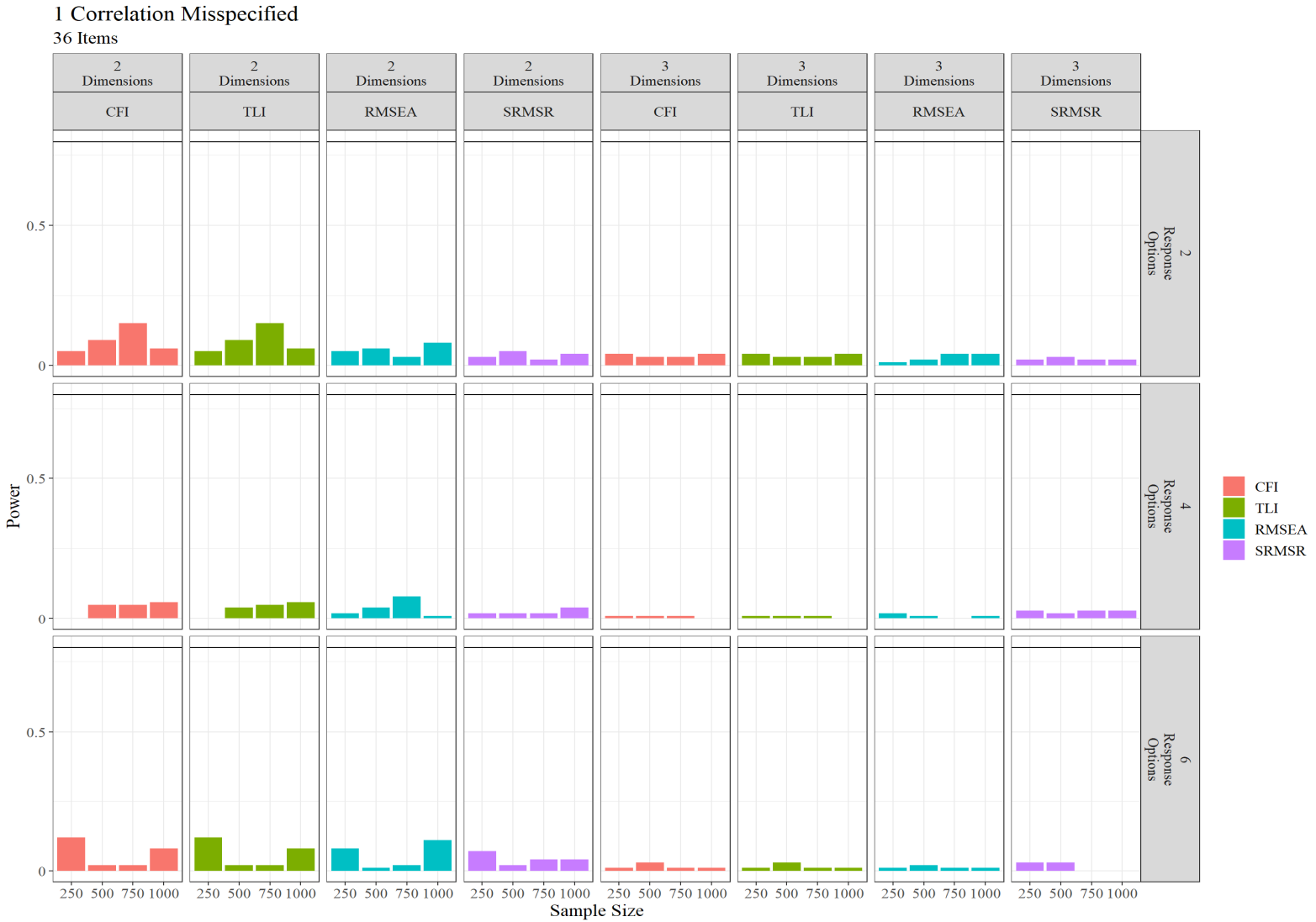

Figure 21. *Power Results for 2 Correlations Misspecified (18 items)*

Figure 22. *Power Results for 2 Correlations Misspecified (36 items)*

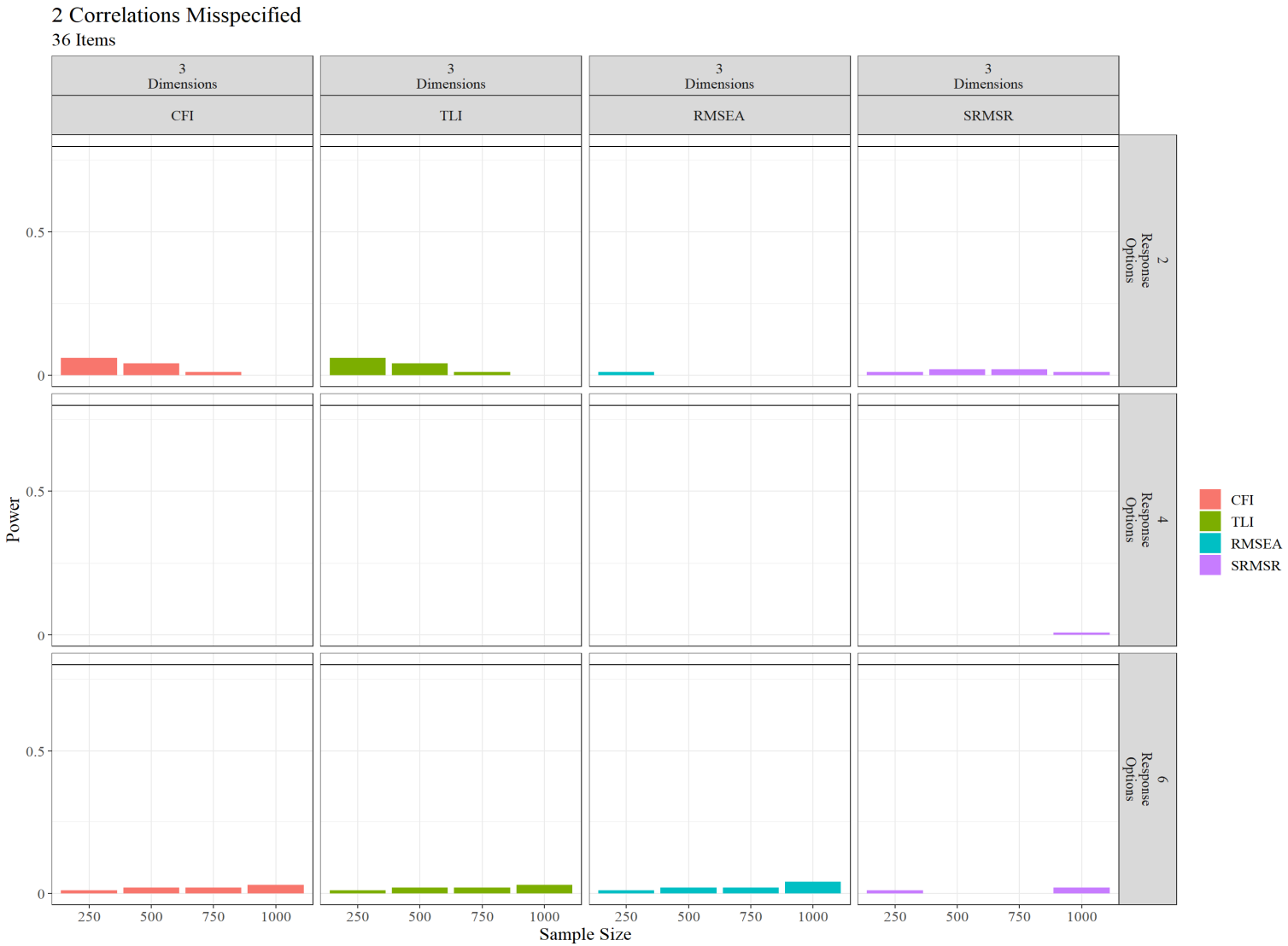

Figure 23. *Power Results for 10% of Items Mis-Loaded (18 items)*

Figure 24. *Power Results for 10% of items mis-loaded (36 items)*

Figure 25. *Power Results for 20% of Items Mis-Loaded (18 items)*

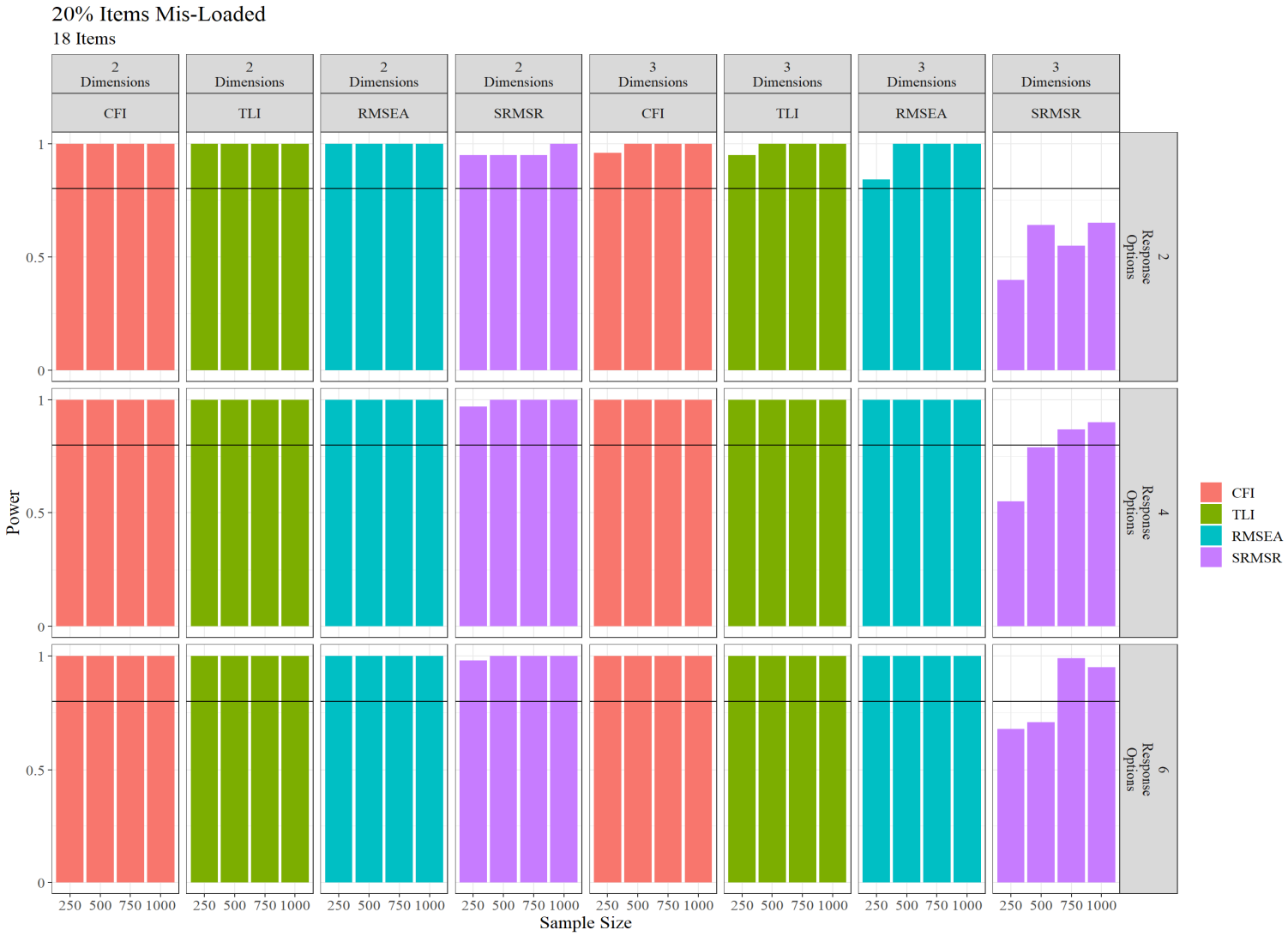

Figure 26. *Power Results for 20% of Items Mis-Loaded (36 items)*

Figure 27. *Power Results for 1 Correlation Misspecified and 10% of Items Mis-Loaded (18 items)*

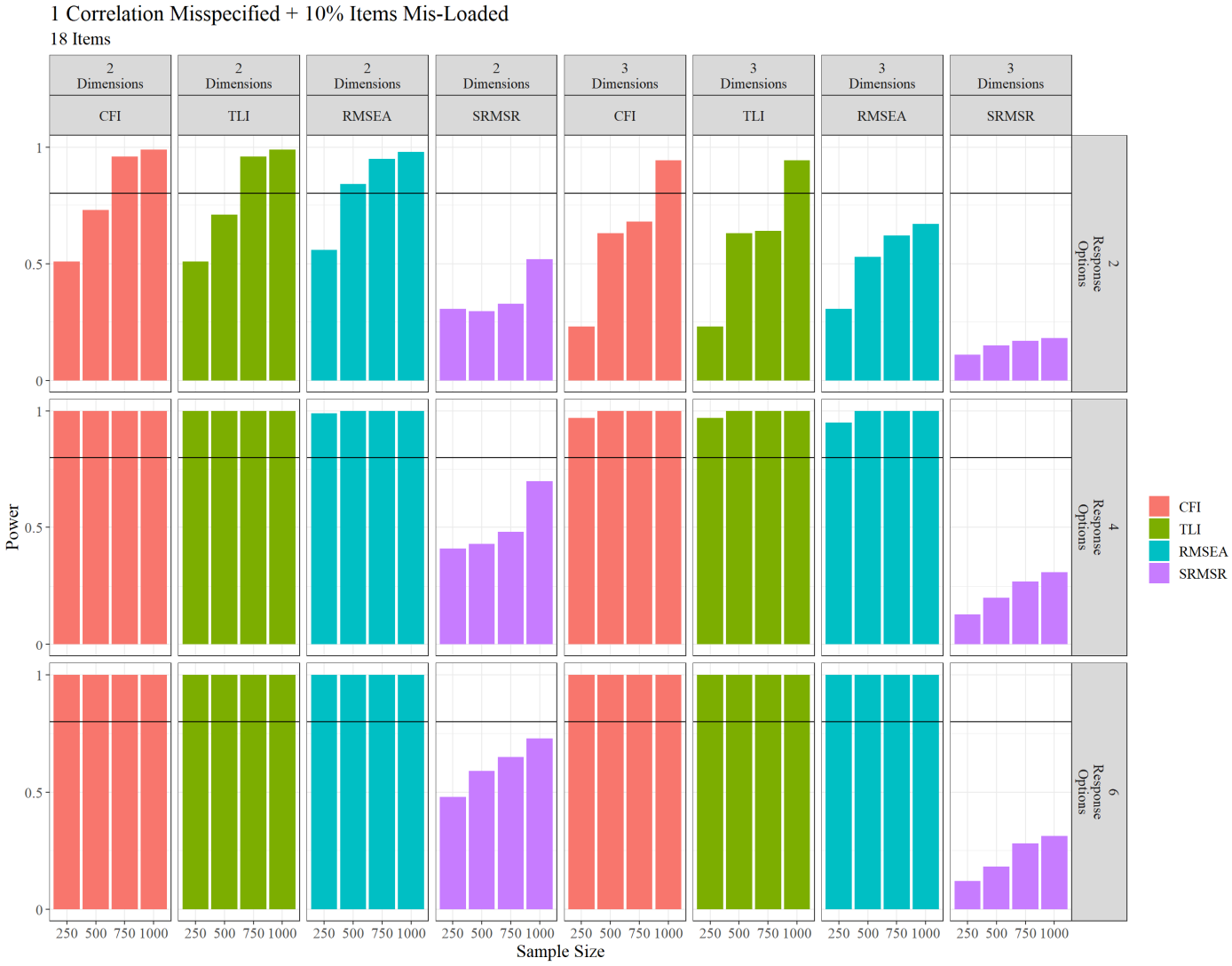

Figure 28. *Power Results for 1 Correlation Misspecified and 10% of Items Mis-Loaded (36 items)*

Figure 29. *Power Results for 1 Correlation Misspecified and 20% of Items Mis-Loaded (18 items)*

Figure 30. *Power Results for 1 Correlation Misspecified and 20% of Items Mis-Loaded (36 items)*

Figure 31. *Power Results for 2 Correlations Misspecified and 10% of Items Mis-Loaded (18 items)*

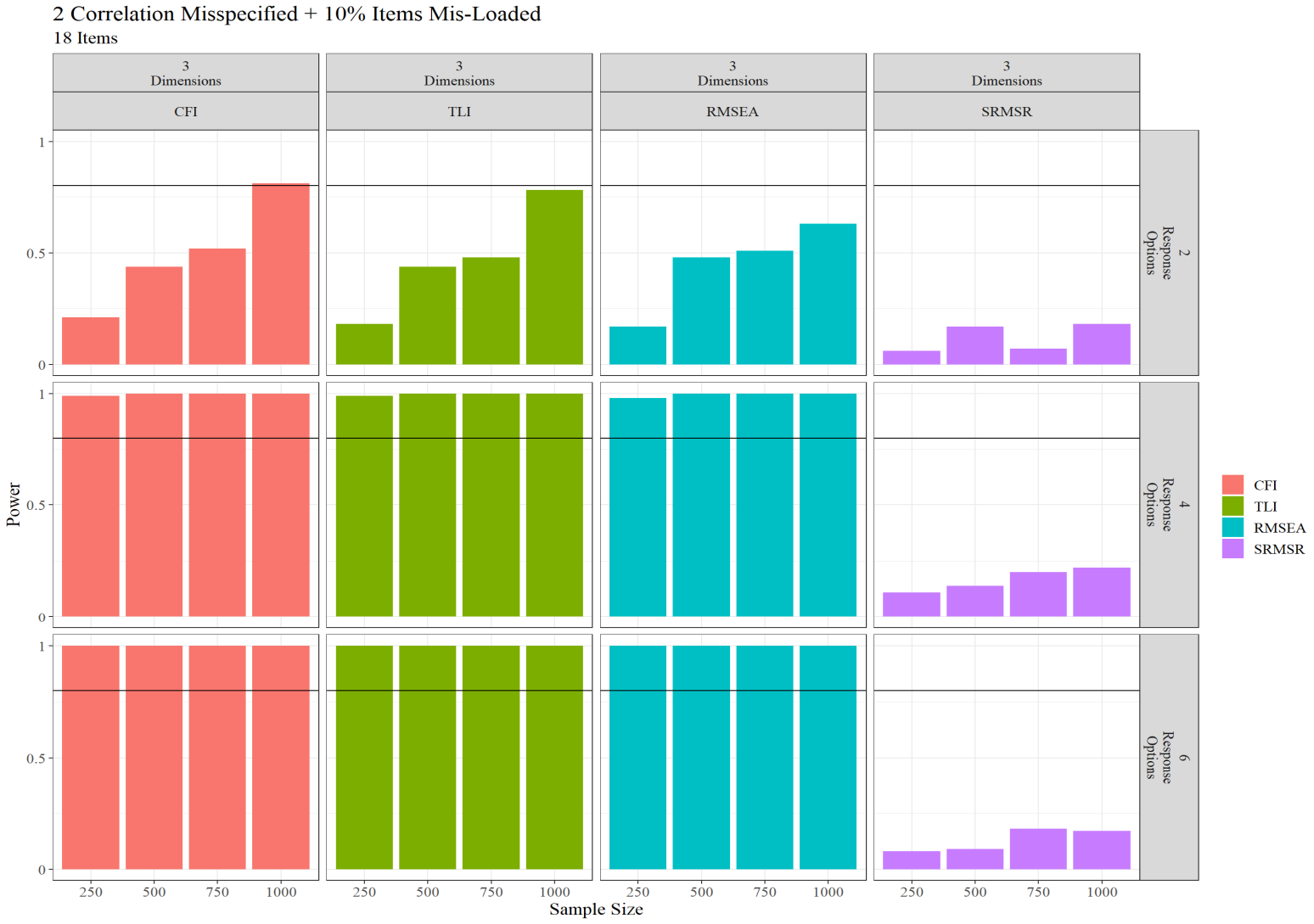

Figure 32. *Power Results for 2 Correlations Misspecified and 10% of Items Mis-Loaded (36 items)*

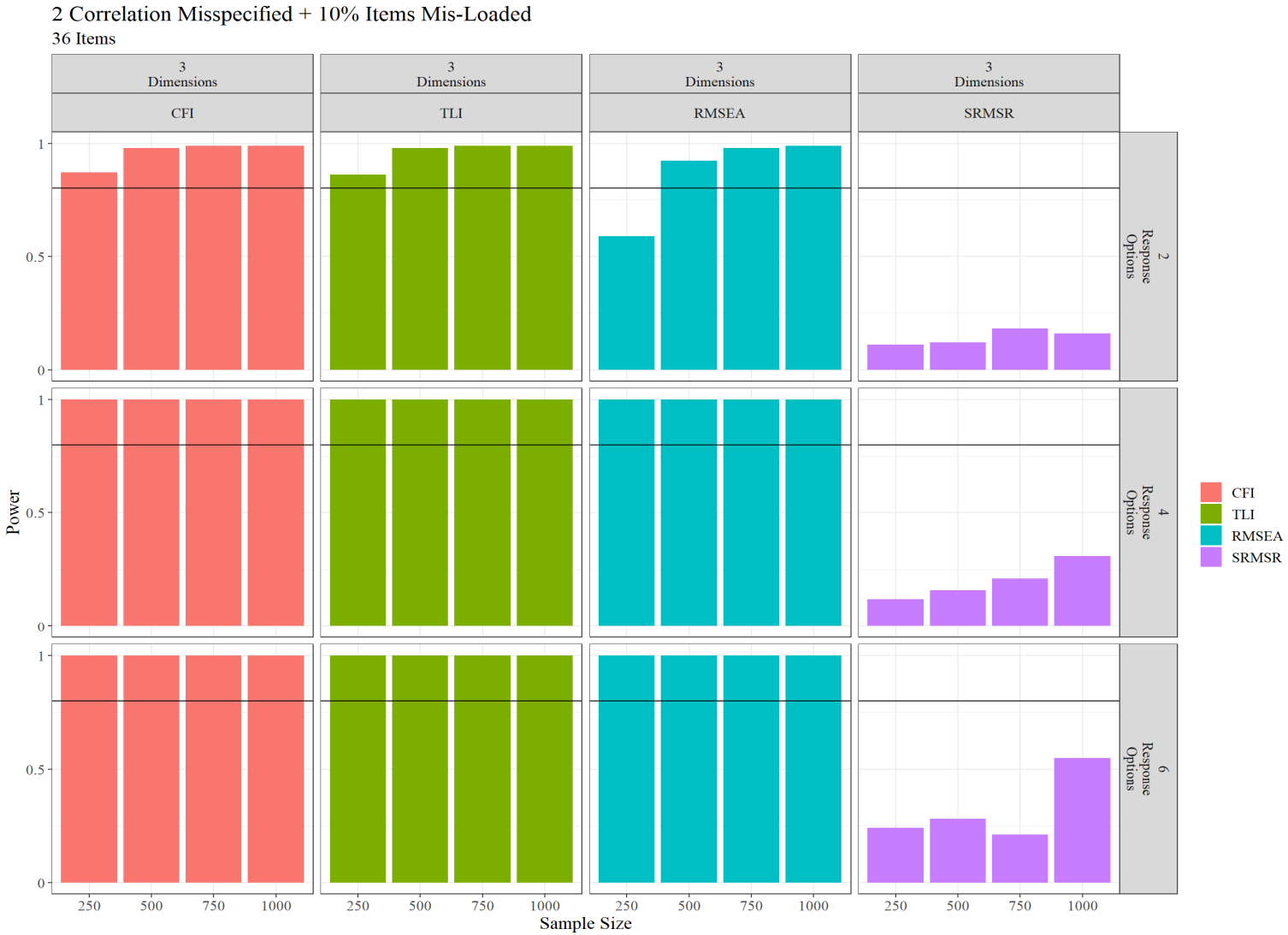

Figure 33. *Power Results for 2 Correlations Misspecified and 20% of Items Mis-Loaded (18 items)*

Figure 34. *Power Results for 2 Correlations Misspecified and 20% of Items Mis-Loaded (36 items)*