Assessing Dimensionality of Latent Structures Underlying Dichotomous Item Response Data with Imperfect Models

A DISSERTATION
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Cengiz Zopluoglu

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Ernest C. Davenport, Adviser

July 2013

Acknowledgements

I want to express the deepest appreciation to my advisor, Dr. Ernest C. Davenport, Jr., for the support and supervision during my graduate education. I am grateful to Dr. Davenport because he always shared his valuable time with me to talk, discuss, and clarify my ideas. I would not have finished writing this thesis without his patience.

I would like to thank Dr. Mark Davison, Dr. Michael Rodriguez, and Dr. Niels Waller, members of my dissertation committee, for their valuable feedback throughout the various iterations of this study.

This work was carried out in part using computing resources at the University of Minnesota Supercomputing Institute. I am grateful for resources provided by the University of Minnesota Supercomputing Institute in order to complete the study.

I would also like to thank the Ministry of National Education of Turkey for their financial support granted through doctoral fellowship, particularly in my first two years of graduate education.

I am indebted to Dr. Soner Durmus and Dr. Zulbiye Toluk Ucar who supported and encouraged me to pursue an academic career in educational sciences.

I would further like to thank my parents for supporting me to pursue my studies. Finally and most importantly, thanks go to my wife and daughter who had a great patience with me and my intense and abnormal working schedule during my graduate education.

Dedication

To the greatest scholars of the lost civilization who lived between 8th and 13th century.

Abstract

The purpose of this study was to investigate the effect of model misspecification due to minor latent factors on a variety of dimensionality assessment methods proposed in the literature by using both real and simulated data. Several dimensionality assessment procedures based on eigenvalue examination (i.e., parallel analysis), conditional covariances (i.e., DETECT), and model selection approach (e.g., NOHARM and Mplus based chi-square statistics, RMSEA, GFI, AIC) were considered in the study.

Two studies were conducted. In Study 1, the average, standard deviation, and range of the number of dimensions suggested by different approaches were investigated using sample datasets drawn from a very large real item response dataset treated as the population. In Study 2, a comprehensive simulation study was run, and the performances of the analytical methods were evaluated using the number of major dimensions in the true generating model as a reference.

The current study provides some interesting and provoking results regarding the performances of some well-known and most commonly used practices under certain conditions. The results of the current study suggest that most of the methods proposed in the literature and available for practitioners are not necessarily useful tools in dimensionality assessment, particularly if the goal of dimensionality assessment is to identify the latent traits with major influences, when the underlying factor structure is complex and minor factors are present. The current study provides some insight for the performance of different dimensionality assessment approaches with misspecified models when the underlying latent structure was factorially complex.

Table of Contents

List of Tables

List of Figures

CHAPTER 1: Introduction

Dichotomous data is a typical measurement outcome in educational and psychological assessments (e.g., TRUE/FALSE items, multiple-choice items). Different statistical models that link observed dichotomous outcomes to latent theoretical constructs have been developed and widely used. While these models are extensively used in modeling dichotomous response data, a challenging early step is to determine the necessary number of latent traits in the model. Multiple latent traits can occur in educational and psychological testing due to either intended or unintended sources. The intended sources of multiple latent traits may be the planned content structure (e.g., a test can include various sub-components such as algebra, geometry, and probability) or different item formats within the test. The unintended sources of multiple latent traits may be construct-irrelevant abilities (e.g., reading ability in a math problem), speed in the test's administration, student motivation due to the testing day and conditions, or dependencies among a set of items related to the same reading passage (Tate, 2009). Although the number of underlying latent traits can be hypothesized a priori in a confirmatory approach, the researchers' judgments may not always fit well to the item response data due to unintended sources of variability.

An exploratory analysis may be helpful to identify unintended sources of variability in item response data (if any) as well as the amount of variability due to those unintended sources, and dimensionality assessment is recommended as "*part of a standard set of analyses conducted after each test administration* (Ackerman, 2005, p. 24)." Several standards in Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999) are also established to encourage such analyses. Some of these standards are as follows:

> "*If the rationale for a test use or interpretation depends on premises about the relationships among parts of the test, evidence concerning the internal structure of the test should be provided (Standard 1.11, p. 20)*."

> "*When interpretation of subscores, score differences, or profiles is suggested, the rationale and relevant evidence in support of such interpretation should be provided. Where composite scores are developed, the basis and rationale for arriving at the composites should be given* (Standard 1.12, p. 20)."

> *"When previous research indicates that irrelevant variance could confound the domain definition underlying the test, then to the extent feasible, the test developer should investigate sources of irrelevant variance. Where possible, such sources of irrelevant variance should be removed or reduced by the test developer* (Standard 3.17, p. 46)."

Number of Latent Traits from an Imperfect Model Perspective

In an exploratory framework, an early and critical step is to decide the number of latent traits that influence the variation in item response data. There are two conceptualizations about the number of latent traits underlying item response data. In the first conceptualization, it is implicitly assumed that the number of latent traits is considerably less than the number of variables. Accordingly, it is common to see the expression of "identifying correct/true number of common latent traits" in the factor analysis literature. This conceptualization is the *taxonomic view*. From a different perspective, the question of identifying correct/true number of latent traits is a "fictional question (Cattell, 1966)" or an "unfortunate choice of words (Preacher, Zhang, Kim, & Mels, 2013)," because the number of latent traits that impact the variation in observed data is not necessarily less than the number of variables; moreover, there may be a large number of latent traits that influence any variable. The second conceptualization is the *explanatory view* (Hakstian & Muller, 1972).

One criticism of the taxonomic view is that it ignores model error, and implies that the statistical model perfectly holds at the population level (Cattell, 1958; MacCallum & Tucker, 1991; MacCallum, 2003). In the explanatory view, it is argued that no model can fit real-world data perfectly, because we never know the true model that generates the real-world data, but only approximate the true model. The true model that holds in the population should be a combination of major latent traits (systematic factors), minor latent traits (incidental factors), and unique latent traits (Tucker, Koopman, & Linn, 1969; MacCallum & Tucker, 1991). While the major latent traits are the main research interest and considerably smaller in number, the number of minor latent traits may be very large and may not have practical importance. In this imperfect model perspective, the variance accounted for by minor latent traits but not explicitly

modeled by the statistical model (e.g., common factor model) represents the model error at the population level, and it is acknowledged that the fitted statistical model is always misspecified to some degree in practice.

From the imperfect model perspective, finding the number of latent traits underlying item response data is not a search for truth; instead, it is a search for finding an optimal solution that provides a balance between model parsimony and lack of fit (MacCallum, 2003; Preacher et al., 2013). Therefore, a more appropriate goal in dimensionality assessment is to find an "optimal number of latent traits" rather than to find a "correct/true number of latent traits" that explains the variation in item response data. The optimal number of latent traits, termed *quasi-true* in the literature, is expected to be the number of latent traits that have major influences on the item response data.

Although the imperfect model perspective is more realistic, the literature, as briefly summarized in the next section, has been dominated by the taxonomic view. Most simulation studies have addressed the performance of dimensionality assessment criteria under the assumption that the fitted models include the true statistical model that generates the observed data at the population level.

Dimensionality Assessment Studies

Determining the number of latent traits to model the item response data in an exploratory framework has been extensively discussed in two separate but related contexts in the literature: exploratory factor analysis (EFA) and item response theory (IRT). While *factor* is a commonly used word in the EFA literature to refer to an unobserved latent trait that influences the variation in item responses, *dimension* is the preference in the IRT literature[1].

In the EFA literature, choosing the number of latent traits in the common factor model has been debated in numerous papers under the heading of "factor retention criteria" (Akaike, 1987; Buja & Eyuboglu, 1992; Cattel, 1966; Cattell & Vogelmann, 1977; Crawford, 1975; Crawford & Coopman, 1979; Crawford, Green, Levy, Lo, Scott,

[1] These two words, "factor" and "dimension," will be used interchangeably hereafter through the document.

Svetina, & Thompson, 2010; Dinno, 2009; Glorfeld, 1995; Green, Levy, Thompson, Lu, & Lo, 2012; Hakstian, Rogers, & Cattell, 1982; Hayton, Allen, & Scarpello, 2004; Humphreys & Ilgen, 1969; Humphreys & Montanelli, 1975; Horn, 1965; Joreskog, 1962; Kaiser, 1960; Kaiser & Hunka, 1973; Mumford, Ferron, Hines, Hogarty, & Kromney, 2003; Nasser, Benson, & Wisenbaker, 2002; Picone, 2009; Preacher et al., 2013; Revella & Rocklin, 1979; Velicer, 1976; Zeng, 2010; Zoski & Jurs, 1996; Zwick & Velicer, 1982, 1986). Most of the studies in the EFA literature considered continuous measurement outcomes; however, they also provided some insight regarding the dimensionality assessment of latent structures underlying dichotomous data. Some of these criteria have already been modified for use with dichotomous data.

The issue has been mostly addressed in the IRT literature under the heading of "assessing the assumption of unidimensionality" when the response outcome is dichotomous (Breithaupt, 1996; Drasgow & Lissak, 1983; De Champlain & Gessaroli, 1998; Finch & Habing, 2007; Finch & Monahan, 2008; Froelich, 2000; Hattie, 1984; Hattie, Krakowski, Rogers, & Swaminathan, 1996; Hambleton & Rovinelli, 1986; Nandakumar, 1991; Nandakumar & Stout, 1993; Nandakumar & Yu, 1996; Seraphine, 1994, 2000; Stout, 1987; Tate, 2003; Tran & Formann, 2009; Walker et al., 2006; Weng & Cheng, 2005). The unidimensionality assumption implies that there is a single dominant latent trait that accounts for a significant amount of variance in item responses. This assumption is crucial for commonly used unidimensional IRT models, and some of the previous studies focused on developing/assessing the statistical criteria to test the null hypothesis of unidimensionality. Acknowledging that the assumption of unidimensionality is always violated to some degree in practice, some other studies also investigated the effects of ignoring the multidimensional data structure on the unidimensional IRT item and person parameter estimates (Ackerman, 1989; Ansley & Forsyth, 1985; Drasgow & Parsons, 1983; Harrison, 1986; Kirisci & Hsu, 1995; Kirisci, Hsu, & Yu, 2001; Reckase, 1979; Oshima & Miller, 1990; Wang, 1986; Way, Ansley, & Forsyth, 1988) and on the unidimensional IRT applications such as test equating (Bolt, 1999; Camilli, Wang, & Fesq, 1995; Cook, Dorans, Eignor, & Petersen, 1983; Cook, Eignor, & Taft, 1988; De Champlain, 1996; Dorans & Kingston, 1985; Stocking &

Eignor, 1986), computerized adaptive testing (De Ayala, 1992; Folk & Green, 1989; Lau, 1997), and differential item functioning (Ackerman, 1988; Linn & Harnisch, 1981).

Although studies for testing the unidimensionality assumption have dominated the IRT literature, the contribution was limited because they did not provide information beyond the first dimension. As the use of multidimensional item response theory (MIRT) models with dichotomous response outcomes is increasing, a critical component is to determine the number of latent dimensions in the model a priori, which is a decision process very similar to the multiple common factor analysis (Reckase, 2009). There are different approaches developed and proposed in the literature to determine the number of latent dimensions to model item response data in the context of MIRT (Bock, Gibbons, Muraki, 1988; Gessaroli & De Champlain, 1996; Gessaroli, De Champlain, & Folske, 1997; Maydeu-Olivares, 2001; Schilling & Bock, 2005; Zhang, 1996). In addition, a few studies focused on the empirical performance of different approaches in determining the number of latent dimensions (Berger & Knol, 1990; Cho, Li, & Bandalos, 2009; Finch & Habing, 2005; Finch, Stage, & Monahan, 2008; Garrido, Abad, & Ponsoda, 2011; Nandakumar, Yu, & Zhang, 2011; Nichol, 2011; Roussos & Ozbek, 2006; Svetina, 2011; Zhang & Stout, 1999).

Research Purposes and Study Overview

Although the literature related to the dimensionality assessment of latent structures underlying item response data is very broad, and many studies have addressed the issue in several aspects, all of these studies, in either the factor analysis or IRT framework, share a major weakness. The previous simulation studies assumed that the true model that the item responses follow at the population level is among the fitted models. Therefore, the estimation and analysis procedures only dealt with sampling error, not with model error. In his presidential address to the Society of Multivariate Experimental Psychology, MacCallum (2003) criticized the dominating approach in the literature as follows:

> "*Although studies based on this general approach may provide some interesting information, I would argue that they are of limited value. Although most Monte Carlo studies can be criticized for some lack of realism, the approach just described is especially problematic for one major reason: It ignores the fact that our models are always wrong to some degree, even in the population. This approach addresses the question: How do our methods behave and perform when the model in question is exactly correct in the population? Although answers to this question might be of interest for theorists, they are of only limited value to users of the methods. A more realistic and relevant question is: How do our methods behave and perform when the model in question is not correct in the population? Answers to this question could be more relevant and informative regarding the performance of methods in practice* (p. 135)."[2]

A similar criticism was also made of simulation studies in the area of model selection with generalized linear models. Burnham and Anderson (2002) identified three major weaknesses of the simulation studies in the area of model selection. In these studies,

i. the true generating model was always a simple model with no tapering effects,

ii. the set of fitted models considered in the analysis always included the true generating model, and

iii. the model selection goal was usually to select the true generating model.

In the traditional common factor model, which is typically used in simulation studies for dimensionality assessment, the measured variables are modeled as a linear combination of a small number of major common factors (e.g., 1, 2, 3, 4) and unique factors. The common and unique factors account for all variances and covariances at the population level. A slightly different factor model was proposed by Tucker, Koopman, and Linn (1969) and MacCallum and Tucker (1991) as an alternative to generate item response data for simulation studies. In the Tucker-Koopman-Linn framework, observed variables are modeled as a linear combination of a small number of major factors (e.g., 1,

[2] Although the argument is mainly discussed in factor analytic literature, the same argument applies to IRT models. Researchers that generate data using IRT models with a known dimensional structure, either unidimensional or multidimensional, always implicitly assume that the model perfectly holds at the population level.

2, 3, 4), a large number of minor factors (e.g., 50), and unique factors. In this framework, a fitted common factor model is always imperfect to some degree in the population, because variance due to the minor factors is not modeled. Therefore, it is possible to examine how a dimensionality assessment criterion reacts to model error in addition to sampling error.

Inappropriately, all previous simulation studies investigated the performance of proposed dimensionality assessment criteria in such a condition that there is no model error at the population level. As highly encouraged by MacCallum (2003), a more relevant and informative study should incorporate both model error and sampling error into the simulation process to mimic a more realistic scenario. So far, a few studies in the factor analytic literature have followed an imperfect model perspective in their research design (de Winter, Dodou, & Wieringa, 2009; Hakstian et al., 1982; Lorenzo-Seva, Timmerman, & Kiers, 2011; MacCallum, Widaman, Preacher, & Hong, 2001; MacCallum, Tucker, & Briggs, 2001; Preacher & MacCallum, 2002; Preacher et al., 2013). While most of them have addressed the issue of sample size in factor analysis, only three studies considered the behavior of dimensionality assessment criteria when the true generating model has major factors and many additional minor factors. These studies by Hakstian et al. (1982), Lorenzo-Seva et al. (2011), and Preacher et al. (2013) also had some limitations. First, they only considered continuous measurement outcomes. Second, they assessed the performance of a limited number of criteria or fit indices. The main motivation of the current study is to contribute to the existing literature by assessing the performance of several dimensionality assessment approaches proposed in the literature when the response outcome is dichotomous and the latent structure underlying dichotomous response outcomes has many minor latent factors in addition to the major latent factors.

The current study follows the guidelines recommended by MacCallum (2003) to achieve its goal. According to MacCallum (2003), there are two ways to design such a study. The first design involves finding a large real dataset and treating this dataset as a population to conduct a sampling study by drawing samples of the desired sample size from that population. It is expected that the real dataset, with a large sample of

observations, contains model error to some degree and very little sampling error. The drawback of this approach is that the researcher never knows and controls the nature and amount of model error. The second design recommended by MacCallum (2003) is to use the common factor model proposed by Tucker, Koopman, and Linn (1969), which includes a smaller number of major latent factors (e.g., 1, 2, 3, 4), a large number of minor factors (e.g., 50), and unique factors when simulating data. So, the results obtained from a common factor analysis always include model error in addition to sampling error.

The current research is composed of two studies that incorporate both MacCallum's recommendations (2003) into their research design.

- In the first study, dimensionality assessment criteria are applied to data from the Minnesota Basic Skills Test (BST) taken by eighth-grade students in mathematics and reading in 2005. Both tests were not timed, so students were given all the time they needed. There were 75 common items in the mathematics test and 40 common items in the reading test administered to all students. The mathematics test had eight content areas (whole numbers, percentage and ratio, number sense, estimation, measurement, tables and graphs, chance and data, space and shape) and the reading test had two content areas (literal comprehension and inferential comprehension). Also, the items in the reading test were asked in the contexts of five different reading passages. For research purposes, two subtests with 20 items and 40 items from each subject area were created, yielding a total number of four different subtests. A sample of students (N=500, 1000) was repeatedly drawn from the large dataset (N=67,510) for each of the subtests, and different dimensionality assessment criteria were implemented for each sample data for studying their performance.
- In the second study, simulated datasets were generated by manipulating the number of major factors, variance accounted for by each major factor, inter-factor correlations, variance accounted for by minor factors, sample size, and number of items. Then, similar to the first study, different dimensionality

assessment criteria were implemented to each simulated sample data for studying their performance under various simulation conditions.

The performance of dimensionality assessment criteria was investigated in two different aspects: the recovery of the number of major latent traits and the sampling variability of the decisions regarding the number of latent traits (model selection uncertainty).

CHAPTER 2: LITERATURE REVIEW

There are two frequent uses of the term "dimensionality" in educational and psychological tests (Reckase, 1990). The first use of the term is *psychological dimensionality*, which refers to the number of hypothesized psychological constructs underlying a set of items. Another use of the term is *statistical dimensionality*, which refers to the minimum number of mathematical variables needed to explain the covariation in the item response matrix. In the second conceptualization, the dimensionality of items and the dimensionality of people are viewed as two different entities, and the statistical dimensionality of the item response matrix is defined as the lesser of these two (Reckase, 1990, 2009; Tate, 2009). According to Reckase (2009), no continuum can be detected in item response data if people do not vary on a particular dimension.

A very similar conceptualization was also made earlier by Embretson (1983), who divided the construct validation process into two phases: *construct representation* and *nomothetic span*. In the construct representation phase, a researcher is concerned with identifying the theoretical constructs underlying a set of items (dimensionality of items). In the nomothetic span phase, some of the concerns a researcher deals with are individual differences (dimensionality of people). According to Embretson (1983), meaningful dimensions cannot be defined if the subjects do not vary systematically on the component ability identified in the nomothetic span phase.

Consequently, statistical dimensionality is not viewed as a property of the test itself, but as a property of the data matrix that results from the interaction between examinees and the test items. The concept of local independence is thought to be a basis to mathematically define the statistical dimensionality of the item response data (Hattie, 1985; Lord & Novick, 1968; McDonald, 1981). As yet, three different definitions of local independence have been presented in the literature.

Mathematical Definition of Dimensionality

According to Lord and Novick (1968), the dimensionality of the complete latent space is equal to *m* if the conditional distribution of the item scores are all independent of

each other for a fixed value of $\boldsymbol{\theta}$ ($\boldsymbol{\theta} = \theta_1, \theta_2, \ldots \theta_m$) where $\boldsymbol{\theta}$ is a vector indicating a person's location in a multidimensional space under the monotonicity assumption[3]. More formally, local independence implies that

$$\mathrm{P}(\mathbf{x} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} \mathrm{P}(x_i \mid \boldsymbol{\theta}), \tag{1}$$

where $\mathbf{x}$ is the observed response vector of a person with an ability vector of $\boldsymbol{\theta}$, x_i is the observed response for item *i,* and *n* is the number of items. This form of local independence is labeled as *strong local independence* (McDonald, 1981) and implies that no further relationships remain between items once the complete latent space is accounted for. In other words, the probability of getting the correct response for any item is independent of the outcome obtained from any other item once the ability parameters on the latent space are controlled (Embretson & Reise, 2000).

A more flexible definition of dimensionality is given based on a different definition of local independence. McDonald (1981) defined dimensionality based on *weak local independence*. Weak local independence assumes that the dimensionality of an item response matrix is the minimum number of latent traits that account only for the pairwise covariances among the items. Formally, this is defined as

$$\mathrm{cov}(x_i, x_j \mid \boldsymbol{\theta}) = 0 \tag{2}$$

for each pair of items. While strong local independence takes into consideration the item dependencies in the third- and all higher-order marginals, weak local independence takes into consideration only the item dependencies for the first- and second-order marginals. Another way to express weak local independence is to write the following equation:

$$\mathrm{P}(x_i, x_j \mid \boldsymbol{\theta}) = \mathrm{P}(x_i \mid \boldsymbol{\theta})\mathrm{P}(x_j \mid \boldsymbol{\theta}). \tag{3}$$

Weak local independence implies that the equation above holds for each pair of items.

---

[3] Monotonicity assumption requires that the probability of getting an item correct is a non-decreasing function of the underlying latent constructs.

Stout (1990) proposed a third and even weaker form of weak local independence, *essential local independence*. Stout (1990) criticizes both strong and weak local independence because they are very strict, impractical, and do not differentiate among latent traits with major and minor influences. Formally, essential dimensionality is defined as the minimum number of latent traits that meet the criteria of essential local independence as it is mathematically defined

$$\frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} |\operatorname{cov}(x_i, x_j \,|\, \boldsymbol{\theta})|}{n(n-1)/2} \to 0 \qquad \text{as} \qquad n \to \infty . \tag{4}$$

So, the average covariance across all possible item pairs is equal to zero, as the number of items goes to infinity once the latent space is fully accounted for. In a basic sense, Stout (1990) defined the dimensionality as the number of dominant traits underlying the test responses. However, this approach is also criticized because it assumes an infinite number of items, and the mathematical limit used in the definition does not provide any practical guidelines regarding how closely the average covariances must approach zero when defining dimensionality (Seraphine, 1994).

Effects of Model Misspecification

One of the critical steps when modeling dichotomous response data using psychometric models is to decide the dimensionality of the item response data a priori. From one perspective, it is not possible to determine the true dimensionality of the latent structures underlying item response data due to its complexity (Cattell, 1966; Humphreys, 1964; MacCallum & Tucker, 1991; MacCallum, 2003), and the goal is to find the number of major dimensions, the *quasi-true* number of dimensions (Burnham & Anderson, 2002; Preacher et al., 2013). From this perspective, any psychometric model is misspecified to some degree due to minor factors, and misspecification due to minor factors is inevitable. There are only a few studies in the literature that investigate the effects of misspecification due to minor factors, indicating the hope that, if not ignorance, the misspecification due to minor factors is tolerable. Most of the studies in the literature have focused on the effects of misspecification due to major factors and ignored the

misspecification due to minor factors. These studies, as they appeared in the factor analytic and item response theory literatures, are summarized in this section.

Factor Analysis

Identifying the number of factors to retain is thought of as one of the most important decisions when fitting a linear common factor model, because it subsequently impacts the rotated factor loadings, factor score estimates, and the interpretability of the factors (Cattell & Vogelmann, 1977; Coovert & Kathleen, 1988; Glorfeld, 1995; Hakstian & Muller, 1972; Mumford, Ferron, Hines, Hogarty, & Kromney, 2003; Preacher & MacCallum, 2002; Zwick & Velicer, 1986). The interpretation of the rotated factor solution depends on the assumption that the "correct" number of factors has been extracted (Turner, 1998).

The possible effects of extracting too many or too few factors are mostly based on intuitive judgments in the factor analytic literature, and there are not many empirical studies that have made systematic investigations into possible consequences. Some scholars argue that extracting too many factors is less problematic than extracting too few factors (Cattell, 1958; Mumford et al., 2003; Velicer, Eaton, & Fava, 2000; Zwick & Velicer, 1986). This is intuitively correct. For instance, a two-dimensional object can be represented in a three-dimensional space without any information loss. But, one of the dimensions has to be ignored if we are forced to represent a three-dimensional object in a two-dimensional space. On the other hand, extracting too many factors can also lead to problems. Crawford (1975) indicated that rotating too many oblique factors may produce high inter-factor correlations or cause the factor space to collapse. In addition, extracting too many factors can lead to factor splitting, particularly when varimax rotation is used, that produces imaginary factors and corrupts the "true" factors (Comrey, 1978).

Empirical studies that investigate the effects of extracting too many or too few factors support the intuition that extracting too many is less problematic than extracting too few (Fava & Velicer, 1992, 1996; Wood, Tataryn, & Gorsuch, 1996). In one of these studies, Wood et al. (1996) ran a simulation study to empirically address the effects of under- and over-factoring. After generating the data based on the multiple common factor

model with a known factor structure, they under- and over-extracted (± 5) the number of common factors and rotated the initial factor pattern solution using the varimax criterion. They computed the root mean squared error between the estimated and the true values of the loadings in the core factors[4]. They found that under-extraction always resulted in greater error than over-extraction when the number of factors in the population was held constant. Under-extraction by two factors resulted in greater error than under-extraction by one factor, and under-extraction by three factors resulted in greater error than under-extraction by two. In contrast, as the degree of over-extraction increased, the average error of core loadings increased slightly. The over-extraction by two or three factors produced about the same error in true factor loadings compared to the over-extraction by one factor. They also reported that the variables that should load on the unextracted factors may incorrectly load on the extracted factors when under-extraction occurs.

Item Response Theory

Although multidimensional IRT models have appeared in the literature since the 1980s, efficient estimation algorithms and computer software, such as TESTFACT and NOHARM, were not widely known and used by practitioners until the early 1990s. Therefore, the use of multidimensional IRT models was very rare in practice, and unidimensional IRT models were most commonly used. However, researchers were also well aware that the educational and psychological data did not meet the assumption of unidimensionality in most instances. Therefore, a trend in the IRT research literature appeared in the late 1970s and continued until the mid-1990s. These studies generally fitted unidimensional models to multidimensional data with known dimensional structures, and examined the effects of model misspecification (due to major dimensions) on the unidimensional model parameters as well as on the results of different IRT applications.

The research in this field can be summarized through the direct effects of multidimensionality on the unidimensional IRT parameter estimates, and the indirect

[4] They define the "core factor" as the number of factors extracted in the analysis (for under-extraction) or the number of factors in the population (for over-extraction), whichever was smaller.

effects of multidimensionality on IRT applications such as test equating, computerized adaptive testing (CAT), and differential item functioning (DIF) through inaccurate unidimensional IRT parameter estimates.

The Effects of Multidimensionality on Unidimensional Parameter Estimates

Wang (1986) derived equations that linked the compensatory multidimensional two-parameter logistic (M2PL) IRT model item parameters to their unidimensional estimates. These equations are as follows:

$$\hat{a}_i = \frac{\mathbf{E_1}^T \mathbf{A}\boldsymbol{\varphi}\mathbf{a}_i}{\sqrt{k_1\left(2.89 + \mathbf{a}_i^T\boldsymbol{\varphi}(\mathbf{A}^T\mathbf{E}_2\mathbf{D}_2^{-2}\mathbf{E}_2^T\mathbf{A})\boldsymbol{\varphi}\mathbf{a}_i\right)}} \text{ and} \tag{5}$$

$$\hat{b}_i = \frac{(\mathbf{a}_i\boldsymbol{\mu} - d_i)}{\mathbf{E_1}^T\mathbf{A}\boldsymbol{\varphi}\mathbf{a}_i}, \tag{6}$$

where $\hat{a}_i$ and $\hat{b}_i$ are the unidimensional item discrimination and difficulty estimates, $\mathbf{A}$ is a matrix of true discrimination parameters for n items on m latent traits (n x m), $\boldsymbol{\varphi}$ is a variance-covariance matrix of the latent traits (m x m), $\mathbf{a}_i^T$ is a vector of the true discrimination parameters for item i (1 x m), d_i is a scalar representing the multidimensional item location parameter for item i (1 x 1), $\boldsymbol{\mu}^T$ is a vector of the latent factor means (1 x m), $\mathbf{E}_1^T$ (1 x n) and k_1 (1 x 1) are the first eigenvector and eigenvalue of $\mathbf{A\Phi A}^T$, $\mathbf{D}_2$ is a diagonal matrix (n-1 x n-1) of the remaining eigenvalues (after removing k_1) of $\mathbf{A\Phi A}^T$, and $\mathbf{E}_2$ is the matrix (n x n-1) of the remaining eigenvectors of $\mathbf{A\Phi A}^T$.

In a simulation study, Oshima & Miller (1990) directly compared the unidimensional item parameter estimates derived from Wang's equations with the empirical unidimensional item parameters estimated from BILOG based on the two-dimensional data, and reported that the values obtained from analytical equations successfully approximated the empirical estimates.

Other studies, as given in Table 1, also empirically studied the characteristics of the unidimensional item parameter estimates obtained from multidimensional data (Ackerman, 1989; Ansley & Forsyth, 1985; Drasgow & Parsons, 1983; Harrison, 1986; Kirisci & Hsu, 1995; Kirisci, Hsu, & Yu, 2001; Reckase, 1979; Way, Ansley, & Forsyth,

1988). However, it is hard to formulate conclusive arguments based on the results of these studies for several reasons. First, most of these simulation studies had either no or insufficient replications within experimental conditions. Second, most of them used LOGIST, which is outdated and has a different estimation algorithm than today's software (e.g., IRTPRO), and it is unknown how much of their results apply to the model parameters estimated by other software commonly used today. Third, each study used a different dimensional structure and did not have any control over the amount of dimensionality generated. This makes it difficult to interpret and compare the results across studies. However, the results can still provide some insight in the light of Wang's theoretical derivation.

For instance, Drasgow and Parsons (1983) simulated five-dimensional data with the dimensions accounting for 12%, 8%, 8%, 8%, and 4% of the variance respectively. When they fit a unidimensional model to five dimensional data, the unidimensional item discrimination parameter estimates were more closely related to the true item discrimination parameters of the first factor (strongest), especially when the correlation among traits is below 0.4. As the correlation among the traits increased to around 0.8, the unidimensional item discrimination parameters were equally related to the true item discrimination parameters of each factor. Also, as the correlations among the traits increased from moderate to high, the root mean squared error (RMSE) between the unidimensional item discrimination parameter estimates and the true item discrimination parameters of each factor decreased from around 0.6 to around 0.3. In a similar design, Ansley and Forsyth (1985) generated two-dimensional data with a dominant first factor and a minor second factor using a non-compensatory multidimensional IRT model. The ratio of the first eigenvalue to the second eigenvalue ranged from 8.06 to 13.76 for 60-item data. They found that the correlation between the unidimensional item discrimination estimates and the true item discriminations of the dominant factor ranged from 0.47 to 0.64, and increased as the correlation between the two dimensions increased. The correlation between the unidimensional item discrimination estimates and the true item discriminations of the minor factor, however, was very close to zero regardless of the magnitude of the correlation between traits. Other studies reported very similar

Table 1. *Simulation Studies for the Effects of Multidimensionality on Unidimensional Item Parameter Estimates*

Study	Sample Size	Number of Items	Generating Model	Estimation Model	Number of True Factors	Inter-factor correlation	Factor Structure	Number of Conditions	Replication
Reckase (1979)	1000	50	LCF	U1PL, U3PL	1, 2, 9	0	Simple, Complex	4	1
Drasgow & Parsons (1983)	1000	50	M2PO	U2PL	1, 5	Small, Medium, High	Simple	10	1
Ansley & Forsyth (1985)	1000, 2000	30, 60	N-M3PL	U2PL	1, 2	0, .3, .6, .9, .95	Complex	20	5
Harrison (1986)	1000	30, 50, 70	M2PO	U2PL	1, 4, 8	Medium, High	Simple	27	5
Way et al. (1988)	2000	60	M3PL, N-M3PL	U3PL	1, 2	0, .3, .6, .9, .95	Complex	16	5
Ackerman (1989)	1000	40	M2PL, N-M3PL	U2PL	1, 2	0, .3, .6, .9	Complex	8	1
Kirisci & Hsu (1995)	1000	40	M3PL	U3PL	1, 3	0, .6	-	12	10
Kirisci et al. (2001)	1000	40	M3PL	U3PL	1, 3	.6	Complex	6	10

Note. U1PL, U2PL, and U3PL represent unidimensional one-, two-, and three-parameter logistic models. M1PL, M2PL, and M3PL represent compensatory multidimensional one-,two-, and three-parameter logistic models. M2PO represents the multidimensional two-parameter normal ogive model, and N-M3PL represents the Sympson's non-compensatory multidimensional three-parameter logistic model. LCF represents the linear common factor model.

patterns regarding the item discrimination parameter (Harrison, 1986; Kirisci & Hsu, 1995; Kirisci, Hsu, & Yu, 2001; Way, Ansley & Forsyth, 1988).

The average of the estimated unidimensional item difficulty parameter was greater than the average of the true item difficulty parameters for both dimensions when the generating model had a non-compensatory nature (Ansley & Forsyth, 1985; Way, Ansley & Forsyth, 1988). On the other hand, when the generating model had a compensatory nature, the average of the estimated unidimensional item difficulty parameter was between the average of the true item difficulty parameters for the first dimension and the average of the true item difficulty parameters for the second dimension (Way, Ansley, & Forsyth, 1988). Also, the estimation error in unidimensional item difficulty parameters decreased as the correlations among the traits increased (Ansley & Forsyth, 1985; Drasgow & Parsons, 1983; Harrison, 1986).

Reckase (1979) was first to recognize the inappropriate interpretation of the unidimensional person parameter estimates when the data was multidimensional. He found that LOGIST estimated only one of the factors, ignoring the other factors when the inter-factor correlations are zero and the unidimensional three-parameter logistic model was fitted to multidimensional data. LOGIST unidimensional ability estimates were highly correlated (r =0.93) with the true factor scores on one factor and nearly uncorrelated (r =*0.29*) with true factor scores on the second factor. A similar observation was also made in other studies when the correlation among the latent traits was weak, $r <$ 0.4 (Ansley & Forsyth, 1985; Drasgow & Parsons, 1983). LOGIST seemed to ignore other factors and be drawn to the factor with the strongest dimensional strength. However, when the traits are correlated from moderate to high, which implies a second-order prepotent general factor, the unidimensional person parameter estimates are found as a weighted linear combination of the underlying traits (Ackerman, 1989; Ansley & Forsyth, 1985; Drasgow & Parsons, 1983; Kim, 1995; Way et al., 1988).

To summarize, it appears that the unidimensional estimates of the model parameters in the presence of multidimensionality are a weighted composite of the underlying traits, and these weights are primarily a function of the discrimination and difficulty parameters, and the correlations among the latent traits. When multiple

dimensions with major influences exist, unidimensional analysis is expected to produce an estimate of ability that is a weighted average of abilities on multiple latent traits. Therefore, it becomes difficult to interpret the unidimensional item and person parameter estimates without any reference to the latent factor structure, and any interpretation should be made with extreme caution.

A final note on the effect of multidimensionality should also be addressed regarding item parameter invariance. Parameter invariance is one of the most important and useful characteristics in IRT. In a simulation study, Oshima and Miller (1990) concluded that if the test has different dimensional structures across different groups — for example, if the test functions as unidimensional for one group and two-dimensional for another group — then a large number of items may not be invariant across samples. Also, when the correlations among the traits may vary across the groups (e.g., due to the instructions, anxiety, or reading ability), the unidimensional item parameter estimates are affected differently from different inter-trait correlations across groups. Therefore, an important property of unidimensional IRT parameters may not be valid anymore when multidimensionality is present in the data.

The Effects of Multidimensionality on Unidimensional IRT Applications

The procedures in most IRT applications operate on calibrated item and person parameters. The multidimensionality presence of the data may impact the results of test equating, CAT administrations, and DIF analysis through inaccurate unidimensional model parameters, and invalidate any inference made from these applications. The multidimensional extensions of these applications are more complicated and still in development (Reckase, 2009). Therefore, it is important to examine the consequences of violating the unidimensionality assumption in widely used IRT applications.

***Test Equating*.** The effects of multidimensionality on test equating with the unidimensional IRT models are mostly studied using real data (Camilli, Wang, & Fesq, 1995; Cook, Dorans, Eignor, & Petersen, 1983; Cook, Eignor, & Taft, 1988; De Champlain, 1996; Dorans & Kingston, 1985). Cook et al. (1983) examined the quality of

the IRT true-score equating procedure using the concept of scale drift[5]. They equated an SAT mathematics form to itself through an equating chain. The test forms in the equating chain were different SAT mathematics forms administered in different years, with slightly different specifications, and were expected to have a multidimensional nature. The same procedure was replicated for an SAT verbal test form. They compared the initial scales of the math and verbal sections with the final scales at the end of the equating chain for the same form. They observed that the equating procedure underestimated both the mean and standard deviation of the initial scale scores for the mathematics form, and the bias accounted for 58% of the total equating error. On the other hand, the equating procedure overestimated both the mean and standard deviation of the initial scale scores for the verbal form, and the bias accounted for 86% of the total equating error.

In another study, Dorans and Kingston (1985) reported that the presence of multidimensionality worsened the symmetry property[6] of the equating through its effect on the magnitude of item discrimination parameter estimates, although they concluded that the effect was not considerably large. Camilli et al. (1985) did not find a substantial impact of multidimensionality on true-score equating results. The error due to multidimensionality was less than two points in all conditions on a scale between zero and 98. De Champlain (1996) studied the invariance property of unidimensional true-score equating procedures with multidimensional data. The results indicated that the differences in conversion lines obtained from the African American, Hispanic, and Caucasian groups were negligible throughout the entire raw-score scale. The conversion lines only differed at the low end of the scale. He concluded that the differences in the underlying latent traits between groups yielded conversions that did not differ substantially for most examinees, and invariance of IRT true-score equating functions

[5] "Scale drift" is said to have occurred if the results of equating test form A directly to test form D are not the same as the results of equating test form A to test form D through intervening forms B and C, and indicates the accumulation of equating error in the chain.
[6] Symmetry property implies that the function used to equate Form X to Form Y is equal to the inverse of the function used to equate Form Y to Form X.

across subgroups of examinees was not degraded markedly in the presence of multidimensionality.

The studies with real data, as summarized above, agreed that the effect of multidimensionality on test equating results was negligible. However, these studies did not control the dimensional structure, strength, and the amount of multidimensionality in the data. The correlation among the sub-tests in verbal or mathematics forms used to examine the effect of multidimensionality was high ($r > 0.7$). Also, some studies reported relatively large first eigenvalues (the ratio of first eigenvalue to second eigenvalue is larger than 5.5), indicating that the amount of multidimensionality in the real data might be very low. Therefore, it would be very optimistic to conclude that researchers can ignore the effects of multidimensionality on test equating based on the results of these real-data studies.

A few studies also addressed possible consequences of multidimensionality on test equating through simulation (Stocking & Eignor, 1986; Bolt, 1999). Stocking and Eignor (1986) simulated non-compensatory, two-dimensional 60-item data with equally dominant dimensions and used concurrent calibration with LOGIST from three independent samples of hypothetical examinees. They reported that the mean of true ability scores was 25 points less than the mean of scaled ability scores (standard deviation of the scale was about 113). They attributed this significant impact to poorly estimated item parameters, especially the overestimation of item difficulty. In another study, Bolt (1999) compared the robustness of three equating methods (linear, equipercentile, and IRT true-score equating) in terms of several equity-based criteria[7] in the presence of multidimensionality. The performance of the IRT method appears to be slightly superior to the conventional methods, especially when the correlations between the dimensions are high (>0.7). However, this does not suggest that unidimensional IRT equating is appropriate for multidimensional data. The difference in ability levels between two administered test forms approximated 0.75, 0.60, and 0.50 [8] as the correlations between

[7] Equity criteria implies that it does not make any difference whether Form A or Form B is taken by the examinees at every given ability level.
[8] The exact numbers are not given in the article, but determined from the figures provided in the article.

traits are 0, 0.3, and 0.7 respectively[9]. The effect of multidimensionality is not ignorable, especially when the inter-trait correlations are low.

Computerized Adaptive Testing. The impact of ignoring multidimensionality becomes more complicated in adaptive testing, because each examinee takes a different set of items with different test length. A few researchers have studied the effects of multidimensionality on unidimensional CAT administrations due to the fact that multidimensional extension of CAT is in its infancy (Reckase, 2009). These studies focused on the ability level estimates from a CAT administration using a unidimensional model to estimate the ability level after administering each item, when the simulees' responses follow a multidimensional model and the items are drawn from a bank with unidimensionally calibrated parameters of multidimensional items (De Ayala, 1992; Folk & Green, 1989). One study also focused on the diagnostic accuracy of the classification decisions in a computerized mastery testing (Lau, 1997).

The results for the ability parameter estimates were very similar to the studies described in the previous section, which examined the effects of multidimensionality on parameter estimation (De Ayala, 1992; Folk & Green, 1989). Both studies reported that the estimated ability level reflected an optimal linear combination of the underlying traits based on their dimensional strength. However, it was pointed out that the optimal linear combination would be different from person to person, because each person takes a different set of items. If the items have different multidimensional structures, then it would be hard to compare ability levels across examinees.

In another study, Lau (1997) reported that computerized adaptive mastery testing was fairly robust against the violation of unidimensionality in terms of classification accuracy. The actual average false positive[10] and false negative[11] rates were about .0164 and .0237 in classification decisions. However, it must be noted that Lau (1997) simulated two-dimensional data with the first dimension accounting for about 77%–80%

---

[9] These numbers are in the theta scale, and the baseline to compare these differences was about .37 in the unidimensional case.
[10] An unqualified examinee is classified as qualified.
[11] A qualified examinee is classified as unqualified.

and the second dimension accounting for about 9%–12% of the variance. It is debatable whether the data simulated in this study has significant amount of multidimensionality.

Differential Item Functioning. Multidimensionality is generally seen as the leading cause of DIF observed after fitting unidimensional IRT models (Embretson & Reise, 2000; Linn & Harnisch, 1981). Linn and Harnisch (1981) stated that

> *"…Differences like those reported above might be labeled item bias. But to note that questions about the metric system are "biased" against black students, and that story problems involving money are "biased" in their favor is not very helpful. This observation implies the items are at fault. But it is at least as plausible that the model is at fault and/or that the "bias" is due to instructional differences. The assumption of unidimensionality is clearly violated for this set of items* (p.116)."

Ackerman (1988) mimicked a very realistic scenario in a simulation study to investigate this argument and examined how different trait distributions produced DIF when the multidimensional nature of the data was ignored. He simulated 40-item multidimensional data based on the compensatory multidimensional two-parameter logistic model using the item parameters derived from an ACT math test. Three hypothetical groups with different means for both abilities were generated. The first group had means of zeros in both dimensions as a reference group, while the second and third groups had vector of means (1, -0.5) and (-0.5, 1) for both dimensions respectively as the focal groups. The unidimensional 2PL item parameters were estimated with LOGIST. DIF analysis revealed that items that primarily load on the first dimension discriminate better for the second group, while the items that primarily load on the second dimension discriminate better for the third group. The results of the unidimensional DIF analysis should be interpreted cautiously in the presence of multidimensionality, especially if the groups are suspected to have different distributions in the underlying traits. It is very likely that a very good multidimensional item will be removed from the test after a unidimensional DIF analysis because of model misspecification due to major factors.

Dimensionality Assessment Procedures

Dimensionality assessment of item response data is a critical process that needs extra attention from both test developers and test users. Because of its importance, Lord (1980) declared that there is a great need for statistical procedures to assess unidimensionality (cited in Hattie, 1984), and this call has recently been extended for such methods to assess multidimensionality (Levy & Svetina, 2010). In this section, several approaches to dimensionality assessment are discussed. These approaches are discussed in different subsections. First, the methods that rely on eigenvalue examination are discussed. Second, dimensionality assessment is discussed from a model selection perspective. Third, the DETECT procedure is described.

Eigenvalue Examination

Eigenvalue examination is the most common practice in published studies, particularly in the factor analytic literature. Three methods in eigenvalue examination are very frequently used: the Kaiser-Guttman rule (Guttman, 1954; Kaiser, 1960), the subjective scree test (Cattell, 1966), and parallel analysis (Horn, 1965). In a review of 152 articles published in three psychology journals during the period of 1975 to 1984, Ford, MacCallum, and Tait (1986) found that the Kaiser-Guttman (KG) rule was the most widely used method (21.7%), followed by the subjective scree test (11.2%), in determining the number of factors in factor analytic studies. In another review of 217 factor analytic studies published in psychology journals between 1991 and 1995, Fabrigar, Wegener, MacCallum, and Strahan (1999) reported that the KG rule was again at the top (16.5%), followed by the scree test (15.2). In a similar review, the KG criteria was again found to be the most widely used method in factor analytic studies (56.7%), followed by the scree test (35%), and parallel analysis was used in only 6.7% of the published articles (Henson & Roberts, 2001).

Kaiser-Guttman Rule

The KG rule, also known as the "eigenvalue-greater-than-one" rule, is the easiest to implement with the least consideration. The original idea is related to Guttmann's weaker lower bound (1954), but it is also attributed to Kaiser (1960). The rule is

originally applied only to component analysis, and suggests retaining the components with an eigenvalue greater than one on a psychometric basis. Kaiser (1960) argues that it is necessary and sufficient for a component to have an associated eigenvalue greater than one for a positive reliability coefficient. But this statement is based on the assumption that there is no error in eigenvalues due to sampling. Therefore, the KG rule is likely to suffer from fluctuations in the sample data, a fact that is acknowledged by empirical research. The number of components retained by the KG rule is expected to be between 1/3 and 1/6 (Kaiser, 1960) or between 1/3 and 1/5 (Gorsuch, 1983) of the number of variables in the dataset regardless of the number of dimensions in the data. The number of components retained based on the KG rule is a function of the number of variables, which is not surprising since the KG rule has a psychometric basis and coefficient alpha is a function of the number of variables. Empirical studies have also confirmed that the number of components retained by the KG rule often falls in a range from 1/3 to 1/5 of the number of variables, especially at low factor saturation (Zwick & Velicer, 1986; Zwick & Velicer, 1982).

The KG rule is not appropriate for common factor analysis, but an adaptation of the KG rule for common factor analysis is related to Guttman's strongest lower bound. In this procedure, the unities on the diagonal of the correlation matrix are replaced by communality estimates (e.g., squared multiple correlations), and eigenvalues are computed from the reduced correlation matrix. The number of positive eigenvalues is suggested as the number of factors to retain in common factor analysis. The same assumption that the correlation coefficients are the population parameters is also made in this application. In an application of Guttman's strongest lower bound for the number of common factors to 64 sample correlation matrices, Kaiser and Hunka (1973) reported that the correlation between the number of observed variables and Guttman's strongest lower bound is .98. They concluded that the strongest lower bound is similarly not an efficient tool for determining the number of common factors in the world of real data. In another study, Humphreys (1964) also provided an example to suggest that the KG rule might be too conservative. He provided an empirical example to show that 10 factors could be extracted and interpreted, while the KG rule suggested only five.

Scree Test

The scree test was first proposed by Cattell (1966) to determine the number of factors to retain. Although its name includes the word "test," it has a descriptive nature and is a highly subjective procedure. In scree examination, the eigenvalues are arranged in descending order and plotted against their order number. A researcher always observes a curve with a big decrease at the beginning, and random fluctuations from right to left around a straight line through the end. The researcher looks for a break in the curve before the random fluctuations start to decide the number of factors to extract. Although this is a very simple and easy-to-implement procedure, it is very subjective and requires some sort of "art" to apply.

The reliability of the decisions regarding the number of factors is one of the concerns for the scree test. It is likely to yield different results for different people regardless of their level of experience (Weiss, 1971). To examine the inter-rater reliability for the scree test, Crawford and Koopman (1979) simulated 100 sample correlation matrices, and the scree plot of each correlation matrix was examined by five raters. They reported that the inter-rater reliability was very low, and inexperienced factor analysts should be cautious when using the scree plot to make decisions. On the other hand, Cattell and Vogelmann (1977) gave full instructions to 12 people (six experienced and six novice) about how to interpret the scree plot, and presented 15 different simulated sample correlation matrices after the instruction. They reported high inter-rater reliability regardless of the amount of experience the subjects had a priori. So far, there is no research with real data, and the inter-rater reliability is likely to be lower with real data even when the practitioners are well trained, because when the number of variables per factor and the sample size are small, definite breaks in the scree plot are less likely to occur, especially with complex factor structures (Hayton, Allen, & Scarpello, 2004). In those cases, it is more likely that the scree test will suffer from subjectivity and ambiguity.

Other researchers attempted to objectify the scree plot examination using a regression approach (Gorsuch, 1983; Jurs, Zoski, & Mueller, 1993; Zoski & Jurs, 1996). These procedures are based on the regression of the magnitude of the eigenvalues on their

ordinal positions. The $\text{Scree}_{\text{CNG}}$ approach uses six eigenvalues at a time, and compares the slope of the first three eigenvalues with the slope of next three eigenvalues. Then, the slope of the second, third, and fourth eigenvalues is compared to the slope of the fifth, sixth, and seventh eigenvalues. This process continues until all sets of consecutive six eigenvalues have been compared. The number of factors to retain is the point where the absolute difference between two slopes is the greatest.

Scree_{MR} is very similar to $\text{Scree}_{\text{CNG}}$, but it uses more information. First, the slope from the first three eigenvalues is compared to the slope from the fourth eigenvalue to the n^{th} eigenvalue. Then, the slope of the second, third, and fourth eigenvalues is compared to the slope of the fifth to the n^{th} eigenvalues, and so on. The number of factors to retain is again the point where the absolute difference between two slopes is the greatest. Also, Nasser, Benson, and Wisenbaker (2002) proposed a *t* value index to compare the slopes obtained from the Scree_{MR} procedure. At each step, the *t* value index is computed for the difference between two slopes, and the number of factors to retain corresponds to the point where the largest absolute value of the *t* index is found.

A slightly different regression approach, Scree_{SE}, which evaluates the fit of successive regression lines, was proposed by Zoski and Jurs (1996). In this approach, the first regression line is fitted from the first eigenvalue to the n^{th} eigenvalue. Then, the second regression line is fitted from the second eigenvalue to the n^{th} eigenvalue, and this continues until the final regression line is fitted to the last three eigenvalues. The model fit for each of *n-3* regression lines is evaluated through the standard error of the estimate computed as follows:

$$s_{Y,X} = \sqrt{\frac{\sum_{i=1}^{n}(k_i - \hat{k}_i)^2}{n-2}}, \tag{7}$$

where k_i is the sample eigenvalue and $\hat{k}_i$ is the corresponding regression estimate. Zoski and Jurs (1996) proposed $1/n$ as a criterion, and each standard error that exceeds $1/n$ is accepted as a non-trivial factor to retain in factor analysis.

As a result of a simulation study that compared four different regression approaches to the subjective scree test, $Scree_{SE}$ was found to be the best procedure. $Scree_{CNG}$ consistently indicated three factors regardless of data characteristics when the factors were correlated. $Scree_{MR}$ consistently overestimated the number of factors when the factors were uncorrelated, and over-extraction increased as a function of sample size. When factors were correlated, $Scree_{MR}$ behaved exactly the same as $Scree_{CNG}$ by predicting three factors regardless of any condition. $Scree_{SE}$ was the most consistent in terms of its performance for correlated and uncorrelated factors. Except for the conditions where the factor loadings are .5 and the number of variables per factor is low, the $Scree_{SE}$ procedure was almost completely accurate (Nasser, Benson, & Wisenbaker, 2002).

Parallel Analysis

Parallel analysis was originally proposed by Horn (1965) for component analysis and can be conceptualized as a more sophisticated way of implementing the KG rule. The eigenvalues from a correlation matrix for the variables uncorrelated in the population would all be equal to one. In other words, the eigenvalues from an identity matrix would create a horizontal straight line at y=1 in a scree plot. Therefore, what KG rule suggests is to keep any component that accounts for more than chance assuming that the correlation matrix is obtained from population. However, the first sample eigenvalues are expected to be greater than one, and later sample eigenvalues are expected to be less than one due to sampling error (Hayton et al., 2004), although there is zero correlation among all variables at the population level. Parallel analysis similarly suggests retaining the components that have an eigenvalue greater than what would be expected due to chance, but it acknowledges sampling fluctuations in eigenvalues. Parallel analysis replaces the "one" in the KG rule with a cut-off criterion derived from the empirical sampling distribution of the eigenvalues from a correlation matrix for variables uncorrelated in the population.

The magnitudes of the eigenvalues from a sample correlation matrix depend on the sample size, the number of variables, and the magnitude of the previous eigenvalues. Unfortunately, there still has not been an analytical solution for the sampling distribution of eigenvalues related to the multivariate random variables uncorrelated in the

population, and analytical solutions appear to be an intractable problem (Glorfeld, 1995). Hence, the parallel analysis procedure empirically derives the sampling distribution of eigenvalues from random data through simulation.

To be able to generate the empirical sampling distribution of eigenvalues for random data, a large number of datasets with zero correlations among variables are generated using the same number of variables and observations in the sample data under examination. Second, a correlation matrix and its eigenvalues are computed for each synthetic dataset. Then, empirical sampling distributions for the eigenvalues at each rank position are obtained. Horn (1965) originally proposed comparing each sample data eigenvalue to the average of the empirical eigenvalue sampling distribution for the corresponding rank position and retaining the components that have larger sample eigenvalues than the average random data eigenvalue. The original procedure was also extended to common factor analysis by creating empirical eigenvalue sampling distributions from the random sample correlation matrices after squared multiple correlations are placed on the diagonal (Humphreys & Ilgen, 1969).

Since it was proposed, some adjustments have been suggested for the original parallel analysis procedure to improve its efficiency. One practical concern was using the average of the empirical eigenvalue sampling distribution. The use of the average eigenvalue as a criterion implies that the original parallel analysis procedure performs at the significance level of 0.5, which is very generous in terms of the conventional hypothesis testing approach (Buja & Eyuboglu, 1992). This would tend to increase the probability of making a Type I error (extracting a factor that actually should not be extracted) and makes the parallel analysis procedure tend to over-factor (Glorfeld, 1995). Therefore, it is suggested that a large number of datasets be generated and that the 95^{th} or 99^{th} percentiles of the empirical eigenvalue sampling distribution be used. As a result of a simulation study, Glorfeld (1995) reported that using the 95^{th} percentile instead of the average reduced the over-extraction error by about 15%.

Another concern was related to the sensitivity of the eigenvalue sampling distribution to the distributional form used to generate multivariate data with uncorrelated random variables. In most applications, the empirical eigenvalue distributions were

generated from uncorrelated random variables with a multivariate normal distribution. Whether these empirical eigenvalue sampling distributions are sensitive to non-normality was an open question. Simulation studies consistently showed that none of the distributional forms overestimate or underestimate the mean or quantiles of the random data eigenvalue sampling distributions. Both mean and centile estimates were stable across various distributional forms (Buja & Eyuboglu, 1992; Dinno, 2009; Glorfeld, 1995).

Recently, researchers argued that the 95th percentile eigenvalues generated from random data provide an appropriate null hypothesis only for the first eigenvalue, because the size of the later noise eigenvalues are influenced by the presence of the prior significant factors (Green et al., 2012; Lautenschlager, 1989; Turner, 1998). Beyond the first eigenvalue, the sampling distribution of the random data eigenvalues is not directly comparable to the sample eigenvalue estimates unless the previous significant factors have been modeled into the data generation process. Therefore, this would suggest a separate simulation to test each eigenvalue rather than only one simulation as in the current practice to test all eigenvalues at once. Green et al. (2012) proposed a revised version of parallel analysis that relies on successive simulations to test each eigenvalue independently by taking the magnitude of previous significant eigenvalues into account.

Modifications of the parallel analysis procedure using tetrachoric correlations have also appeared in the literature and reported encouraging results in terms of determining the necessary number of latent traits to model dichotomously scored data (Cho et al., 2009; Crawford, Green, Levy, Lo, Scott, Svetina, & Thompson, 2010; Drasgow & Lissak, 1983; Finch & Monahan, 2008; Tran & Formann, 2009; Weng & Cheng, 2005). These procedures first simulate multivariate continuous data with uncorrelated variables, and then transform these continuous data to binary variables using thresholds. Then, the eigenvalues extracted from the reduced tetrachoric correlation matrix, and sampling distribution of the eigenvalue at each position, are obtained. Finally, the eigenvalue estimates of the sample tetrachoric correlation matrix are compared to the empirical sampling distribution of eigenvalues from random data.

***A drawback of parallel analysis in complex factor structures*.** Parallel analysis has become more accessible to practitioners as the computational tools have been made available in recent years. In factor analytic studies, even editorial recommendations were made to use parallel analysis rather than KG rule or scree plot when making a decision regarding the number of factors to retain (Thompson & Daniel, 1996). Recently, a significant number of papers appeared that suggested the application of parallel analysis on tetrachoric correlation matrices for the dimensionality assessment of dichotomous item response data. All these studies argued that parallel analysis is a promising procedure in dimensionality assessment of latent structures (Cho et al., 2009; Crawford et al., 2010; Dinno, 2009; Drasgow & Lissak, 1983; Finch & Monahan, 2008; Tran & Formann, 2009; Weng & Cheng, 2005).

In most of these studies, however, a simple factor structure was used in data generation. This may cause a highly misleading reliance on parallel analysis because the interpretation of the first few eigenvalues from reduced correlation matrices may be ambiguous when the underlying factor structure is complex. For demonstration purposes, Table 2 presents five different factor structures. Structures 1 and 2 represent a simple and complex factor structure respectively with two common factors. Structures 3, 4, and 5 represent a simple, semi-complex, and complex factor structure respectively with three common factors. At the bottom of the table, the sums of the squared loadings are reported to show the relative strength of each common factor at the population level in the five structures. In addition, for each of the five factor structures, 1000 sample correlation matrices for a sample size of 1000 were generated using a common factor model, and the eigenvalues were computed after squared multiple correlations were placed on the diagonal. At the bottom of the table, the average value for the first few eigenvalues across 1000 samples are also reported. Finally, the 95^{th} percentiles of the empirical sampling distribution for the corresponding eigenvalues obtained from 1000 random datasets are reported.

For Structure 1 and Structure 3, which have simple factor structures, average sample eigenvalue estimates reflect the sum of squared loadings at the population level for the associated common factor. When parallel analysis is applied to sample data with

these generating factor structures, it is expected that parallel analysis will find the correct solution almost every time, as the sample eigenvalue estimates will be much higher than the corresponding random data eigenvalues. On the other hand, the interpretation of the magnitude of the first few eigenvalues is not the same when the underlying factor structure is semi-complex or complex. For instance, the average sample estimate for the first eigenvalue in Structure 2 is 7.27, which seems to reflect the sum of the first two eigenvalues (5.59 + 1.82 = 7.41) at the population level. The average sample estimate for the second eigenvalue, however, is even smaller than the corresponding random data eigenvalue. If parallel analysis is applied to sample data coming from a population with Structure 2, it is very likely that parallel analysis will select a one-factor model almost every time. Similarly, in Structure 5, the average sample estimate for the first eigenvalue seems to reflect the sum of the first three eigenvalues (5.88 + 2.93 + 0.73 = 9.54) at the population level, and the average sample estimates for the second and third eigenvalues are smaller than the corresponding random data eigenvalues. Again, if parallel analysis is applied to sample data coming from a population with Structure 5, parallel analysis will favor a one-factor model almost every time. Finally, in Structure 4, average sample estimates for the first two eigenvalues are higher than the population values, but the average sample estimate for the third eigenvalue is smaller than the random data eigenvalue. Parallel analysis would select a two-factor model for sample data coming from a population with Structure 4, although there are three common factors underlying the data.

The main point of this demonstration is that the interpretations of the first few eigenvalues are ambiguous when the underlying factor structure is not perfectly simple. Therefore, it is always the case that parallel analysis may give misleading results for non-simple structures. Interestingly, this fact was realized and known almost a century ago by Thomson (1916) when he published “A hierarchy without a general factor” and provided the following conclusions after a sort of simulation study with dice:

Table 2. *Demonstration for the Interpretation of First Eigenvalues with Complex Structures*

	Factor Structure												
	1		2		3			4			5		
Item	F1	F2	F1	F2	F1	F2	F3	F1	F2	F3	F1	F2	F3
1	0.6	0	0.6	0.4	0.6	0	0	0.6	0	0	0.6	0.3	0.3
2	0.5	0	0.5	0.3	0.6	0	0	0.7	0	0	0.7	0.3	0.1
3	0.7	0	0.7	0.5	0.7	0	0	0.7	0	0	0.6	0.5	0.1
4	0.5	0	0.6	0.2	0.6	0	0	0.6	0	0	0.4	0.5	0.2
5	0.6	0	0.6	0.2	0.6	0	0	0.7	0	0	0.5	0.5	0.3
6	0.7	0	0.7	0.4	0	0.4	0	0.5	0.2	0	0.5	0.4	0.2
7	0.6	0	0.7	0.3	0	0.5	0	0.6	0.2	0	0.7	0.6	0.2
8	0.6	0	0.5	0.2	0	0.5	0	0.6	0.3	0	0.4	0.4	0.2
9	0	0.3	0.6	0.4	0	0.5	0	0.4	0.2	0	0.7	0.5	0.1
10	0	0.3	0.5	0.4	0	0.4	0	0.4	0.4	0	0.7	0.5	0.3
11	0	0.4	0.6	0.3	0	0	0.3	0	0.7	0.4	0.8	0.4	0.2
12	0	0.5	0.7	0.3	0	0	0.7	0	0.6	0.4	0.7	0.4	0.3
13	0	0.3	0.6	0.5	0	0	0.3	0	0.5	0.5	0.6	0.3	0.3
14	0	0.3	0.6	0.2	0	0	0.3	0	0.6	0.4	0.8	0.5	0.2
15	0	0.3	0.6	0.4	0	0	0.5	0	0.7	0.4	0.5	0.4	0.1
Sum of Squared Loadings	2.92	0.86	5.59	1.82	1.93	1.07	1.01	3.48	2.32	0.98	5.88	2.93	0.73
Average Eigen-value[1]	2.87	0.80	7.27	0.12	1.85	1.01	0.84	3.85	2.69	0.09	9.27	0.14	0.09
Random Eigen-value[2]	0.28	0.22	0.28	0.22	0.28	0.22	0.18	0.28	0.22	0.18	0.28	0.22	0.18

Note 1. Average sample eigenvalue estimates were obtained from 1000 sample correlation matrices generated based on the associated factor structure. Sample size was 1000. Squared multiple correlations were replaced with unities on the diagonal before extracting eigenvalues.

Note 2. Random eigenvalue is the 95th percentile of the empirical sampling distribution for the corresponding eigenvalue based on 1000 random datasets. Sample size was 1000 and number of variables was 15 when generating random data. Squared multiple correlations were replaced with unities on the diagonal before extracting eigenvalues.

> " *1. Since correlation does actually exist between items, there must be either Group factors or a General factor present, or both.*
> *2. If there is no general factor, then it is probable that group factors* ***overlap*** *in a complicated fashion; for otherwise there would be no hierarchy.*
> *3. There is not the slightest mathematical evidence so far forthcoming which will enable us to distinguish between* ***overlapping*** *Group Factors and a General Factor* (p. 280 – 281)."

By "overlapping Group Factors," Thomson (1916) implies that there are items loading on more than one common factor. As he realized, when overlapping occurs in a complicated fashion, the resulting correlational structure suggests an imaginary general factor, as implied by a very large first eigenvalue of sample data generated based on Structure 2 and Structure 5 above. To conclude, researchers should use any method that relies on eigenvalue interpretation to assess dimensionality of item response data with caution.

Model Selection Approach

Model selection is another approach to determine the dimensionality of item response data. In this approach, several competing models with different numbers of latent traits are fit to data, and one of the models is selected based on a criterion. In case of dichotomous item response data, psychometric models such as the compensatory multidimensional two-parameter and three-parameter normal ogive models (Reckase & McKinley, 1982; McDonald, 1999) are commonly used. The compensatory multidimensional two-parameter normal-ogive model (M2PO) is mathematically equivalent to the multiple common factor model within a parameter transformation and a nonlinear link function between manifest and latent variables. In the multiple common factor model, the uniqueness is defined as the unexplained variance in item score by the hypothesized factor structure. The uniqueness is equal to

$$\psi_i^{\,2} = 1 - \boldsymbol{\lambda}_i^{\,T} \boldsymbol{\varphi} \boldsymbol{\lambda}_i \,, \tag{8}$$

(Mulaik, 2010, p. 133) where ψ_i is the uniqueness for the *i*th item, $\boldsymbol{\lambda}_i$ is an *m* x 1 vector of factor loadings for the *i*th item on *m* latent dimensions, and $\boldsymbol{\varphi}$ is an *m x m* correlation

matrix among *m* latent dimensions. Then, the probability to give the correct response to item *i* for person *s* is defined as

$$\mathrm{P}(x_{is}=1\,|\,\boldsymbol{\theta}_s)=UVN\left(\frac{-\tau_i+\boldsymbol{\lambda}_i^{T}\boldsymbol{\theta}_s}{\psi_i}\right), \tag{9}$$

where $\boldsymbol{\theta}_s$ is a vector of ability parameters on *m* dimensions for person *s*, τ_i is the threshold for item *i*, and *UVN*(.) is the cumulative distribution function for the univariate standard normal distribution. In Equation 9, the new parameters are defined as

$$d_i=\frac{-\tau_i}{\psi_i}, \text{ and} \tag{10}$$

$$\mathbf{a}_i=\frac{\boldsymbol{\lambda}_i}{\psi_i}, \tag{11}$$

where d_i is the item location parameter, and $\mathbf{a}_i$ is a vector of item discrimination parameters. Reckase (2009) defined

$$A_i=\sqrt{\mathbf{a}_i^{T}\mathbf{a}_i} \text{ and} \tag{12}$$

$$B_i=\frac{-d_i}{A_i}, \tag{13}$$

where A_i is the multidimensional item discrimination parameter and B_i is the multidimensional item difficulty parameter.

The compensatory multidimensional three-parameter normal ogive (M3PO) model, an extension of the compensatory M2PO model, is given by McDonald (1999) as

$$\mathrm{P}(x_{is}=1\,|\,\boldsymbol{\theta}_s)=c_i+\left[(1-c_i)*UVN\left(\frac{-\tau_i+\boldsymbol{\lambda}_i^{T}\boldsymbol{\theta}_s}{\psi_i}\right)\right], \tag{14}$$

where c_i is the guessing parameter for item *i*.

After a series of models with different numbers of latent traits is fit to the data, there are several criteria in the literature proposed to select a model. The most commonly

known and used criteria can be classified in different categories based on the discrepancy function they minimize. In Cudeck and Henley's framework (1991), four matrices are defined to distinguish these discrepancy functions: population covariance matrix ($\boldsymbol{\Sigma}$), model-implied covariance matrix in population ($\tilde{\boldsymbol{\Sigma}}$), sample covariance matrix (**S**), and estimated model-implied covariance matrix in sample ($\hat{\boldsymbol{\Sigma}}$). The *discrepancy at the sample level (DS)* is defined as the difference between **S** and $\hat{\boldsymbol{\Sigma}}$, and represents the degree of misfit between a specified model and a true model at the sample level. The *discrepancy due to approximation (DA)* is defined as the difference between $\boldsymbol{\Sigma}$ and $\tilde{\boldsymbol{\Sigma}}$, and represents the degree of misfit between the specified model and the true model at the population level. *The discrepancy due to estimation (DE)* is the difference between $\hat{\boldsymbol{\Sigma}}$ and $\tilde{\boldsymbol{\Sigma}}$, and represents the sampling variability. Finally, the *overall discrepancy (DO)* is defined as the difference between $\boldsymbol{\Sigma}$ and $\hat{\boldsymbol{\Sigma}}$, and it represents the sum of the *discrepancy due to estimation* and *discrepancy due to approximation* (Cudeck & Henley, 1991; Preacher et al., 2013).

The chi-square statistics and root mean squared residual (RMSR) statistics are concerned with minimizing DS under the assumptions that the specified model has a perfect fit in the population (DA ≈ 0) and the sample size is large enough (DE ≈ 0). Other fit indices (e.g., RMSEA, CFI) are concerned with minimizing DA under the assumption that the model is fit at the population level (DE ≈ 0). Information-theoretic approaches (e.g., AIC and BIC), on the other hand, aim to minimize the overall discrepancy (DO ≈ DA + DE) by establishing a balance between approximation error and estimation error (Kline, 2005; Preacher et al., 2013).

Criteria based on sample discrepancy

A common way of selecting a model with a number of latent traits is null hypothesis testing through chi-square test statistics. In general, a psychometric model (e.g., M2PO, M3PO) with a certain number of latent traits is fit to data; a chi-square statistic is computed based on the discrepancy between model prediction and sample data; and a probability value associated with the chi-square is used to make a decision about whether or not the model provides an adequate fit. The number of latent

dimensions in the model is increased one at a time until the specified model fits to sample data at a certain significance level.

In dimensionality assessment of dichotomous item response data, several chi-square statistics are proposed in the literature in the context of different estimation approaches to fit the M2PO and M3PO models. Bolt (2005) summarized two different approaches to fit compensatory multidimensional IRT models. These are *limited-information* and *full-information* approaches. One type of limited-information approach is to fit a linear common factor model to the tetrachoric correlation matrix to approximate the M2PO model. Once the tetrachoric correlation matrix is obtained among dichotomous items, the common factors can be extracted using iterative principal factor analysis (IPFA), unweighted least squares factor analysis (ULS), generalized least squares factor analysis (GLS), maximum likelihood factor analysis (MLFA), alpha factor analysis (Alpha), or minimum residual factor analysis (MINRES) (Bolt, 2005; Knol & Berger, 1991; Mulaik, 2010). A useful recommendation is to correct the tetrachoric correlation estimates for any guessing effect (Carroll, 1945). In addition, tetrachoric correlation matrices are generally not positive-definite, and smoothing the tetrachoric correlation matrix is recommended before common factor analysis to make them positive-definite. (Wothke, 1993). Software such as MicroFACT (Waller, 2001) and Mplus (Muthén & Muthén, 1998 – 2010) are available for researchers to fit a linear common factor model to the tetrachoric correlation matrix.

A second type of limited-information approach is the Normal Ogive Harmonic Analysis Robust Method (NOHARM) developed by Fraser and McDonald (1988). NOHARM employs an unweighted least squares approach to fit a polynomial factor model up to a third degree using Hermite-Tchebycheff polynomials to approximate the nonlinear multidimensional IRT models (Maydeu-Olivares, 2001; McDonald, 1967).

As an alternative to the limited-information approaches, Bock, Gibbons, and Muraki (1988) introduced a full-information approach to fit the M2PO model directly to the frequencies of all distinct response patterns. This approach is recommended due to the limitations of computing tetrachoric correlations in some cases, especially when the items have extreme difficulties. The proposed algorithm employs the marginal maximum

likelihood estimation based on the EM algorithm (Bock & Aitkin, 1981; Bock et al., 1988; Dempster, Laird, & Rubin, 1977). The MLR estimator in Mplus (Muthén & Muthén, 1998 – 2010) and IRTPRO (Cai, Thissen, & du Toit, 2011) implements the full information approach to fit the multidimensional IRT models.

McDonald's Unweighted Least squares Estimation. McDonald's ULS fits a polynomial approximation to the M2PO and M3PO models by using the first- and second-order marginal probabilities. The details of the estimation procedure are described in Maydeu-Olivares (2001) and Bolt (2005). Let π_i and π_{ij} be the proportion correct for the *i*th item and joint-proportion correct for the *i*th an *j*th item, respectively; and p_i and p_{ij} be the corresponding sample estimates with the following relationship:

$$\pi_i = p_i + \varepsilon_i \text{ and} \tag{15}$$

$$\pi_{ij} = p_{ij} + \varepsilon_{ij} \tag{16}$$

where ε_i and ε_{ij} are error terms. The model-predicted first- and second-order marginals can be written using the estimated parameters of M2PO as $\pi_i = UVN(-\tau_i)$ and $\pi_{ij} = BVN(-\tau_i, -\tau_j, \rho_{ij})$, where $UVN(.)$ and $BVN(.)$ are the cumulative distribution functions for the univariate and bivariate normal distributions, respectively; and ρ_{ij} is the tetrachoric correlation between item *i* and *j* as defined $\rho_{ij} = \boldsymbol{\lambda}_i^T \boldsymbol{\varphi} \boldsymbol{\lambda}_j$.

The goal of the estimation is to find the set of model parameters $\boldsymbol{\tau}$, $\boldsymbol{\Lambda}$, $\boldsymbol{\varphi}$ that minimize the following fit function:

$$\mathrm{F} = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = (\mathbf{P} - \boldsymbol{\pi})^T (\mathbf{P} - \boldsymbol{\pi}), \tag{17}$$

where $\boldsymbol{\tau}$ is a vector of threshold parameters, $\boldsymbol{\Lambda}$ is the factor loading matrix, $\boldsymbol{\varphi}$ is inter-factor correlation matrix, $\boldsymbol{\pi} = (\boldsymbol{\pi}_1, \boldsymbol{\pi}_2)$ is a vector of model-predicted first-order ($\boldsymbol{\pi}_1$ = $\pi_1, \pi_2, \ldots, \pi_n$) and second-order ($\boldsymbol{\pi}_2 = \pi_{21}, \pi_{31}, \ldots \pi_{n,n-1}$) marginal proportions, $\mathbf{P} = (\mathbf{P}_1, \mathbf{P}_2)$ is the corresponding sample estimates of $\boldsymbol{\pi}$, and $\boldsymbol{\varepsilon}$ is a vector of corresponding residuals ($\mathbf{P} - \boldsymbol{\pi}$).

A unique characteristic of McDonald's ULS estimation as implemented in NOHARM is using Hermite-Tchebycheff polynomials when computing π_{ij} rather than dealing with a double integration in a bivariate normal distribution cumulative distribution function. A fourth degree polynomial approximation for the second-order marginal proportion for item *i* and *j* is

$$\pi_{ij} = UVN(-\tau_i) * UVN(-\tau_j) + f(-\tau_i)\, f(-\tau_j) \sum_{k=1}^{4} \frac{(\boldsymbol{\lambda}_i{}^T \boldsymbol{\varphi} \boldsymbol{\lambda}_j)^k}{k!} H_{k-1}(\tau_i)\, H_{k-1}(\tau_j), \quad (18)$$

where $f(.)$ is the probability density function for the standard normal distribution and H_k is a Hermite polynomial of degree *k* as defined for the first four terms:

$$H_k(x) = \begin{cases} 1 & \text{if} \quad k = 0 \\ x & \text{if} \quad k = 1 \\ x^2 - 1 & \text{if} \quad k = 2 \\ x^3 - 3x & \text{if} \quad k = 3 \end{cases}. \quad (19)$$

In NOHARM, Equation 17 is minimized using either a quasi-newton or a conjugate gradients minimization algorithm. The iterations continue until the F value continues to decrease and the magnitude of the largest gradient is smaller than some small value set by the user.

Once the polynomial approximation of the M2PO or M3PO model with a certain number of latent traits is fit to data using NOHARM, a chi-square fit statistic can be computed. So far, four different chi-square statistics based on NOHARM estimation have been proposed in the literature (De Champlain, 1993; Gessaroli, De Champlain, & Folske, 1997; Maydeu-Olivares, 2001). The first one is the approximate chi-square statistic (AChi, De Champlain, 1993). In the approximate chi-square statistic, the residual correlations are obtained from the proportion corrects and model-predicted residual joint-proportion corrects:

$$\hat{r}_{ij} = \frac{\varepsilon_{ij}}{\sqrt{p_i(1-p_i)p_j(1-p_j)}}. \quad (20)$$

Then, each residual correlation is transformed to a Fisher's z using the formula

$$z_{ij} = \frac{1}{2}\log_e\left(\frac{1+\hat{r}_{ij}}{1+\hat{r}_{ij}}\right), \tag{21}$$

and the approximate chi-square statistic is computed as

$$AChi = (N-3)\sum_{i-2}^{n}\sum_{j=1}^{i=1}(z_{ij})^2 . \tag{22}$$

It is argued that this statistic is approximately distributed as a chi-square distribution with degrees of freedom equal to $\frac{n(n-1)}{2} - t$, where t is the number of estimated parameters in the model.

The second alternative is the approximate likelihood ratio chi-square statistic (ALR, Gessaroli et al., 1997). In this procedure, the contribution of each item pair to the likelihood is first computed as

$$G_{ij}^{\ 2} = -2*(p_{11}\log_e\frac{\hat{p}_{11}}{p_{11}} + p_{01}\log_e\frac{\hat{p}_{01}}{p_{01}} + p_{10}\log_e\frac{\hat{p}_{10}}{p_{10}} + p_{00}\log_e\frac{\hat{p}_{00}}{p_{00}}), \tag{23}$$

where $p_{11}, p_{01}, p_{10}, p_{00}$ are the observed proportions of examinees in score combinations for item i and j, and $\hat{p}_{11}, \hat{p}_{01}, \hat{p}_{10}, \hat{p}_{00}$ are the corresponding model-based estimates. The $G_{ij}^{\ 2}$ values are assumed to be independent of each other, so the approximate likelihood ratio statistic is the sum of $G_{ij}^{\ 2}$ values over all item pairs:

$$ALR = \sum_{i=2}^{n}\sum_{j=1}^{i} G_{ij}^{\ 2} . \tag{24}$$

The resulting statistic is compared to a chi-square distribution with the same degree of freedom as AChi.

Both Achi and ALR statistics are criticized by Maydeu-Olivares (2001), because they lack theoretical justification for why those statistics should be distributed as a chi-square distribution. Maydeu-Olivares (2001) provided a theoretical derivation for a computationally more exhaustive chi-square fit statistic while developing the large

sample asymptotic standard errors for NOHARM parameter estimates. Four matrices have to be defined to give the computational details of this chi-square statistic (Maydeu-Olivares, personal communication, 2012):

1) $\mathbf{\Delta_{11}} = \dfrac{\partial \boldsymbol{\pi}_1}{\partial \boldsymbol{\tau}}$, derivatives of first-order probabilities with respect to thresholds. This is an n x n diagonal matrix with elements $-f(\tau_i)$.

2) $\mathbf{\Delta_{21}} = \dfrac{\partial \boldsymbol{\pi}_2}{\partial \boldsymbol{\tau}}$, derivatives of second-order probabilities with respect to thresholds. This is an $\tilde{n}$ x n matrix with elements

$$\frac{\partial \boldsymbol{\pi}_2}{\partial \tau_r} = \begin{cases} -f(\tau_i)\ UVN\left(\dfrac{-\tau_j + \rho_{ij}\tau_i}{\sqrt{1-\rho_{ij}^{\ 2}}}\right) & \text{if } r = i \\ -f(\tau_j)\ UVN\left(\dfrac{-\tau_i + \rho_{ij}\tau_j}{\sqrt{1-\rho_{ij}^{\ 2}}}\right) & \text{if } r = j \\ 0, & \text{otherwise} \end{cases} \tag{25}$$

where $\tilde{n} = n(n\text{-}1)/2$.

3) $\mathbf{\Delta_{22}} = \dfrac{\partial \boldsymbol{\pi}_2}{\partial \boldsymbol{\Lambda}}$, $\tilde{n}$ x q matrix of derivatives of second-order probabilities with respect to factor loadings in the factor structure where q is the number of factor loadings in the structure.

4) $\mathbf{\Gamma}$ is a square matrix of dimensions $\tilde{n} + n = n\,(n+1)/2$. It is the asymptotic covariance matrix of the first- and second-order proportions. Let $\mathbf{Y}$ be the N x n data matrix. The cross-products matrix, what NOHARM uses as input, is $\mathbf{C} = \dfrac{\mathbf{Y}^T\mathbf{Y}}{N}$. The diagonals of $\mathbf{C}$ are the first-order proportions ($\mathbf{P_1}$), and the below

diagonal elements are the second-order proportions ($\mathbf{P_2}$) observed in the sample. Let $\mathbf{P}$ is a vector including both $\mathbf{P = (P_1, P_2)}$.

$$\mathbf{d}_j = \left(\mathbf{y}_j, vecr(\mathbf{y}_j \mathbf{y}_j^T)\right)^T \tag{26}$$

is a $n + \tilde{n}$ vector for each respondent, where $\mathbf{y}_j$ is the response vector for the jth respondent, and *vecr(.)* represents an operation that takes the lower triangular part of a matrix excluding the diagonal. Then,

$$\mathbf{\Gamma} = \left(\frac{1}{N} \sum_{j=1}^{N} \mathbf{d}_j \mathbf{d}_j^T \right) - \mathbf{P}\mathbf{P}^T . \tag{27}$$

Given that the above matrices are defined, two more matrices are computed as follows:

$$\mathbf{\Omega} = (-\mathbf{\Delta_{21}}\mathbf{\Delta_{11}}^{-1} \mid \mathbf{I}_{\tilde{n}})\, \mathbf{\Gamma}\, (-\mathbf{\Delta_{21}}\mathbf{\Delta_{11}}^{-1} \mid \mathbf{I}_{\tilde{n}})^T \text{ and} \tag{28}$$

$$\mathbf{H} = \mathbf{I}_{\tilde{\mathbf{n}}} - \mathbf{\Delta_{22}}(\mathbf{\Delta_{22}}^T \mathbf{\Delta_{22}})^{-1} \mathbf{\Delta_{22}}^T , \tag{29}$$

where $\mathbf{I}_{\tilde{n}}$ is the identity matrix with a dimension $\tilde{n}$ and "|" presents an operation that puts two matrices together. Maydeu-Olivares (2001) proposed a test statistic, $T = N\,\mathrm{F} = N\,\boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon}$, which is asymptotically distributed as a weighted sum of independent chi-squared distributions with one degree of freedom each, where the weights are the eigenvalues of $\mathbf{H\Omega}$. Two scaled chi-square test statistics,

$$T_M = \frac{r}{Tr(\mathbf{H\Omega})} T \text{ and} \tag{30}$$

$$T_{MV} = \frac{Tr(\mathbf{H\Omega})}{Tr((\mathbf{H\Omega})^2)} T , \tag{31}$$

as mean-adjusted (T_{M}) and mean-and-variance adjusted (T_{MV}), are proposed to assess the model fit for NOHARM estimation. The mean-adjusted T statistic is suggested to compare a chi-square distribution with r, the degrees of freedom of the model; and mean-

and-variance adjusted T statistic is suggested to compare a chi-square distribution with an adjusted degrees of freedom, $\frac{(Tr(\mathbf{H\Omega}))^2}{Tr((\mathbf{H\Omega})^2)}$.

Muthen's Weighted Least squares Estimation. Similar scaled chi-square test statistics can be computed using the model parameter estimates obtained from fitting a linear common factor model to the tetrachoric correlation matrix through the Muthen's weighted least squares estimation. Let the tetrachoric correlation estimate between item *i* and *j* be

$$\hat{\rho}_{ij} = BVN^{-1}(p_{ij} \mid -\hat{\tau}_i, -\hat{\tau}_j) \tag{32}$$

and δ_{ij} be the discrepancy between and ρ_{ij} and $\hat{\rho}_{ij}$. Then, the goal of the estimation is to find the set of model parameters $\boldsymbol{\tau}$, $\boldsymbol{\Lambda}$, $\boldsymbol{\varphi}$ that minimize the following fit function

$$\mathrm{F} = \boldsymbol{\delta}^T \boldsymbol{\delta} = (\hat{\boldsymbol{\rho}} - \boldsymbol{\rho})^T \mathbf{W} (\hat{\boldsymbol{\rho}} - \boldsymbol{\rho})^T , \tag{33}$$

where $\boldsymbol{\rho}$ is a vector of model-predicted tetrachoric correlations and $\hat{\boldsymbol{\rho}}$ is the corresponding sample estimate, and **W** is the weight matrix, which is the inverse of the asymptotic covariance matrix of the estimated tetrachoric correlations. In Mplus, WLSM and WLSMV estimators use only the diagonal elements of **W** rather than the full **W** matrix to simplify the estimation procedure (Asparouhov & Muthen, 2007). The WLSM and WLSMV estimators in Mplus are identical to each other in estimation, but they differ in computing the chi-square statistic. The WLSM estimator computes mean-adjusted chi-square statistics based on the Mplus estimated model parameters by using the computational procedure explained above for the T_{M} statistic. The WLSMV estimator computes a mean-and-variance adjusted chi-square statistic that differs slightly from $\mathrm{T_{MV}}$ above (Asparouhov & Muthen, 2010). Let's define *T* again as $T = N\,\mathrm{F} = N\,\boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{\delta}$, which is asymptotically distributed as a weighted sum of independent chi-squared distributions with one degree of freedom each, where the weights are the eigenvalues of $\mathbf{H\Omega}$. Asparouhov & Muthen (2010) defined $T_{\mathrm{MV}}^{\mathrm{Mplus}} = a*T + b$, where *a* and *b* are equal to

$$a = \sqrt{\frac{r}{Tr((\mathbf{H\Omega})^2)}} \text{ and} \tag{34}$$

$$b = r - \sqrt{\frac{r\left(Tr(\mathbf{H\Omega})\right)^2}{Tr((\mathbf{H\Omega})^2)}}, \tag{35}$$

respectively. The mean-and-variance adjusted chi-square statistic based on the Mplus model parameter estimates is distributed as a chi-square distribution with the degrees of freedom *r*.

Full Information Maximum Likelihood Estimation (FIML). A final type of chi-square statistic is based on full information factor analysis through marginal maximum likelihood estimation based on the EM algorithm (Bock & Aitkin, 1981; Bock, Gibbons, & Muraki, 1988; Dempster, Laird, & Rubin, 1977). First, a linear common factor model is fitted to the tetrachoric correlation matrix to get starting values. After starting parameters are obtained, an expected log-likelihood function is computed after multiple integrals are approximated in the expectation step of the EM algorithm. Then, new item parameters that maximize the likelihood function are found in the maximization step. An iterative procedure continues until the estimated item parameters obtained in the maximization step are stabilized.

A likelihood ratio statistic after fitting the M2PO or M3PO model using the full information factor analysis is given as

$$G^2 = 2 * \left[\sum_{l=1}^{s} w_l \log_e\left(\frac{w_l}{N\widetilde{P}_l}\right)\right], \tag{36}$$

where w_l is the observed frequency of response pattern *l*, $\widetilde{P}$ is the model-predicted marginal probability for the response pattern *l*, and *s* is the number of observed distinct response patterns in the item response data (Bock, Gibbons, & Muraki, 1988). It's argued that G^2 is distributed as a chi-square distribution with a degree of freedom

$$(s-1) - \left[n(m+1) - \frac{m(m-1)}{2}\right],$$

where *m* is the number of latent traits in the model and *n* is the number of items (Wood et al., 2003). But, this chi-square test assumes that all 2^n possible response patterns have expected values greater than one or two (Bock et al., 1988). If the number of possible distinct response patterns is much larger than the sample size and the table of frequencies is sparse, the results of this procedure are neither very reliable nor powerful (Bock et al., 1988; Bolt, 2005). As an alternative to overcome this limitation, a likelihood ratio chi-square difference test (G^2_{diff}) is a recommended practice. In this procedure, G^2_m and G^2_{m+1} are computed after fitting the *m* and *m*+1 dimensional models, and then the difference of these two statistics is compared to a chi-square distribution with a degree of freedom equal to the difference in the degrees of freedom of two successive models. This process is continued until adding another dimension does not significantly contribute to explaining the variance in the item response data (Bock et al., 1988; Wood et al., 2003).

Standardized Root Mean Squared Residual. The standardized root mean squared residual (SRMR, Bentler, 2006) is a descriptive measure of discrepancy between model-predicted correlations and observed sample correlations. The SRMR is computed using the following formula:

$$SRMR = \sqrt{\frac{\sum_{i=2}^{n}\sum_{j=1}^{i}(\hat{r}_{ij} - r_{ij})^2}{n(n-1)/2}}, \tag{37}$$

where $\hat{r}_{ij}$ is the model-predicted correlation and r_{ij} is the observed sample correlation between item *i* and *j*. In case of dichotomous data, Mplus output reports the SRMR based on the residual tetrachoric correlations. NOHARM does not report the SRMR statistic, but users can compute it after transforming the residual joint proportion correct statistics to correlations using Equation 20. Since SRMR is a descriptive measure, there are rules of thumb to evaluate the magnitude of the SRMR statistic. Different cut-off values are given in the literature. For instance, Hu and Bentler (1999) recommend values less than 0.08, while Kline (2005) suggests values less than 0.10. As a result of a simulation study, Yu (2002) has recommended a cut-off value of 0.07 for categorical outcomes.

Criteria based on approximation discrepancy

There are several limitations of the criteria based on the discrepancy observed at the sample level, and the use of chi-square statistics in model selection is criticized in the literature. One criticism from a theoretical point of view is that it is not very interesting to minimize the discrepancy between **S** and $\hat{\boldsymbol{\Sigma}}$ assuming that the specified model perfectly holds at the population level, because scholars agree that any model is misspecified to some degree, and it is unrealistic to assume that the fitted model is correct at the population level: $\boldsymbol{\Sigma} - \widetilde{\boldsymbol{\Sigma}} \approx \mathbf{0}$ (Brown & Cudeck, 1993; Cudeck & Henley, 1991; MacCallum, 2003; Preacher et al., in 2013). In addition, the sample size in practice is generally not large enough to assume that the estimation error is small; $\hat{\boldsymbol{\Sigma}} - \widetilde{\boldsymbol{\Sigma}} \approx \mathbf{0}$ or $\mathbf{S} - \boldsymbol{\Sigma} \approx \mathbf{0}$. To overcome the limitations of the criteria based on the discrepancy observed at the sample level, some other fit indices are proposed for model selection. The key characteristic of these fit indices is that their goal is to minimize the error of approximation, the discrepancy between the specified model and the true model ($\boldsymbol{\Sigma}$ and $\widetilde{\boldsymbol{\Sigma}}$), at the population level (Cudeck & Henley, 1991; Kline, 2005; Preacher et al., 2013); and they do not depend on the sample size.

Root Mean Square Error Approximation. The root mean square error approximation (RMSEA, Steiger & Lind, 1980) can be thought of as "*an estimate of model misfit per degree of freedom in the population,*" (Preacher et al., 2013) and is a measure for the lack of fit of a certain specified model to the population correlation matrix. The RMSEA statistic is computed using the following formula:

$$RMSEA = \sqrt{\frac{\max\left(\chi_{\mathrm{M}}{}^{2} - df_{\mathrm{M}}, 0\right)}{df_{\mathrm{M}}(N-1)}}, \tag{38}$$

where $\chi_{\mathrm{M}}{}^{2}$ is the model chi-square statistic and df_{M} is the model's degree of freedom. A rule of thumb is that values smaller than 0.05 indicate close approximate fit, values between 0.05 and 0.08 are reasonable error of approximation, and values larger than 0.10 suggest poor fit (Brown & Cudeck, 1993; Kline, 2005). Hu and Bentler (1999) suggest a cut-off value of 0.06 for the RMSEA statistic, while Yu suggests using 0.05 when the

sample size is larger than 250 and increasing the cutoff criteria for smaller samples. In a different implementation, Preacher et al. (2013) found that it is better to use the confidence interval for the RMSEA statistic instead of using a point sample estimate. In their implementation, they retained the model when the lower bound of the RMSEA statistic dropped below 0.05.

Comparative Fit Index. Another population-based index is comparative fit index (CFI, Bentler, 1990). CFI is a measure of relative improvement in model fit for the specified model compared with a baseline model. The baseline model is generally the independence model, which hypothesizes no correlation among the variables at the population level, although a different baseline model can be specified (Kline, 2005). The CFI statistic is computed as

$$CFI = 1 - \frac{\max\left(\chi_{\mathrm{M}}^{2} - df_{\mathrm{M}}, 0\right)}{\max\left(\chi_{\mathrm{M}}^{2} - df_{\mathrm{M}}, \chi_{\mathrm{B}}^{2} - df_{\mathrm{B}}, 0\right)}, \tag{39}$$

where χ_{B}^{2} is the chi-square statistic and df_{B} is the degree of freedom for the baseline model. Typically, a value greater than .95 is suggested as a cut-off criterion for good fit of the specified model (Hu & Bentler, 1999; Yu, 2002).

Criteria based on the overall discrepancy

The criteria based on the discrepancy due to approximation also have limitations, because they are population-based and ignore estimation error. The overall discrepancy is approximately equal to the sum of discrepancy due to approximation and discrepancy due to estimation, or DO ≈ DA + DE. Increasing model complexity decreases the approximation error by being closer to the operating model at the population level, but also increases the estimation error because more parameters are estimated with the same amount of data (Browne, 2000; Zucchini, 2000). If the increase in estimation error is more than the decrease in approximation error, then fitting a better approximating more complex model to sample data may not be beneficial, because the gain in approximation error is not more than the loss in estimation error, and overall error increase. Therefore, finding the best approximating model at the population level by minimizing DA does not guarantee that the same model is the best fitting model across samples. A simpler model

may have less overall error than a more complex model at the sample level, although it is less accurate. The indices such as Akaike Information Criterion and Bayesian Information Criterion are frequently used in practice, and they minimize the overall discrepancy.

Akaike Information Criterion. Akaike information criterion (AIC, Akaike, 1973) is shown to be related to the Kullback-Leibler distance in information theory. Let *f(x)* be the true probability density function underlying the data, and *g (x|η)* be the model-implied probability density function given the true parameters of a specified model (*η*). The Kullback-Leibler distance, or Kullback-Leibler information, between two density functions *f(x)* and *g(x|η)* is given as

$$I(f,g) = \int f(x) \log\left(\frac{f(x)}{g(x \mid \eta)}\right) dx . \tag{40}$$

The notation *I (f, g)* denotes "the information lost when *g* is used to approximate *f*." The above equation is mathematically equivalent to

$$I(f,g) = \int f(x) \log\left(f(x)\right) dx - \int f(x) \log\left(g(x|\eta)\right) dx , \tag{41}$$

and each term can be written as a statistical expectation with respect to *f*,

$$I(f,g) = \mathrm{E}_f\left[\log\left(f(x)\right)\right] - \mathrm{E}_f\left[\log\left(g(x|\eta)\right)\right] . \tag{42}$$

The first expectation is a constant since the truth does not change, but it is unknown. The second expectation changes from model to model, and is also unknown to us because it depends on the true model parameters. In general, we estimate the second expectation using the estimated model parameters from a sample. Therefore, the estimated Kullback-Leibler information is

$$I(f,\hat{g}) = \mathrm{E}_f\left[\log\left(f(x)\right)\right] - \mathrm{E}_f\left[\log\left(g(x|\hat{\eta})\right)\right] . \tag{43}$$

Since the first expectation is a constant and does not change from model to model,

$$I(f,\hat{g}) - \mathrm{C} = -\mathrm{E}_f\left[\log\left(g(x|\hat{\eta})\right)\right] \tag{44}$$

provides an estimated relative distance of a specific model from the truth (Burnham & Anderson, 2002). In model selection, a set of models can be ranked based on their estimated relative distance from the truth, and the best approximating model can be chosen.

Akaike (1973) found a relationship between the maximized log-likelihood and the estimated relative Kullback-Leibler information. The maximized log-likelihood is a biased estimator of the estimated relative distance of a specific model from the truth, and this bias is approximately equal to the number of estimated parameters in the model

$$I(f,\hat{g})\text{-}\mathrm{C} = \log\left(\mathrm{L}(\hat{\eta}\,|\,\text{sample data})\right) - K\,, \tag{45}$$

where K is the number of estimated parameters in the model. Finally, AIC is defined as

$$AIC = -2\left[\log\left(\mathrm{L}(\hat{\eta}\,|\,\text{sample data})\right) - K\right] = -2\log\left(\mathrm{L}(\hat{\eta}\,|\,\text{sample data})\right) + 2K\,. \tag{46}$$

The choice of -2 for multiplication is arbitrary and for historical reasons (Burnham & Anderson, 2002). Any number can be chosen instead of -2, since it does not really change the rank order of the competing fitted models as long as both log-likelihood and K are multiplied by the same constant.

As noted above, the overall discrepancy is a sum of two components: discrepancy due to approximation and discrepancy due to estimation. From a different perspective, the first term of the AIC, $-2\log\left(\mathrm{L}(\hat{\eta}\,|\,\text{sample data})\right)$, reflects the approximation error, and it decreases as the model gets more complex and closer to the truth. The second term of the AIC (2K) reflects the estimation error and uncertainty in the parameter estimation, and it increases as the model gets more complex, because a larger number of parameters is estimated from the same amount of data. Therefore, AIC minimizes the overall error by establishing a balance between the approximation error and estimation error.

In small samples, Burnham and Anderson (2002) recommended using a corrected AIC,

$$AIC_c = AIC + \frac{2K(K+1)}{N-K-1}\,, \tag{47}$$

when the ratio of sample size (N) to the number of estimated parameters (K) is smaller than 40. In practical applications, a set of candidate models is ranked from the lowest AIC to the highest AIC, and the model with the lowest AIC is chosen as it is the best approximating model to the truth.

Bayesian Information Criterion. Bayesian information criterion (BIC, Schwarz, 1978) has a very similar structure to AIC and refers to information theory as well, but this is a "misnomer" (Burnham & Anderson, 2002). The underlying concept for BIC is the Bayes factor (BF), which is the ratio of likelihoods of the data for a given two models as defined

$$\mathrm{BF} = \frac{\mathrm{P}(\mathrm{data} \mid \mathrm{Model\,A})}{\mathrm{P}(\mathrm{data} \mid \mathrm{Model\,B})}. \tag{48}$$

While a value of BF greater than one favors Model A, a value of BF less than one favors Model B. It's argued that -2log(BF) between two models can be approximated by the difference in BICs for the two models (Ghosh & Samanta, 2001; Raftery, 1995; Western, 1999), where BICs are defined as

$$\mathrm{BIC_A} = -2\log\left(\mathrm{L}(\hat{\eta}_A \mid \mathrm{sample\,data})\right) + K_A \log(N) \text{ and} \tag{49}$$

$$\mathrm{BIC_B} = -2\log\left(\mathrm{L}(\hat{\eta}_B \mid \mathrm{sample\,data})\right) + K_B \log(N). \tag{50}$$

If $\mathrm{BIC_A}$ is equal to $\mathrm{BIC_B}$, then BF is equal to 1 and indicates that both models are equally likely to be the true model. If the difference between $\mathrm{BIC_A}$ and $\mathrm{BIC_B}$ ($\mathrm{BIC_A}$ - $\mathrm{BIC_B}$) is greater than zero, then BF is less than 1 and Model B is more likely to be the true model; if the difference is less than zero, then BF is greater than 1 and Model A is more likely to be the true model. So, a model with smaller BIC is thought to be more likely the true model. In practical applications, a set of candidate models is ranked from the lowest BIC to the highest BIC, and the model with the lowest BIC is chosen. Although BIC is very similar to AIC in terms of structure and usage, it is not an information criterion, because it is not an estimate of Kullback-Leibler information (Burnham & Anderson, 2002).

One of the important assumptions for BIC is that a true model exists, and this model is in the set of candidate models. The probability that BIC selects the true model approaches one as the sample size goes to infinity under the assumption that the true

model is in the set of candidate models. BIC is highly criticized by Burnham and Anderson (2002) and not recommended for use due to the unrealistic assumptions. In contrast, AIC does not make such an assumption that the true model is in the set of candidate models.

DETECT

In addition to the eigenvalue examination methods and a wide variety of model selection approaches, a non-parametric method is proposed to assess the number of latent traits underlying item response data. DETECT, Dimensionality Evaluation to Enumerate Contributing Traits, is a conditional covariance-based nonparametric method to assess multidimensionality (Kim, 1995; Zhang, 1996). DETECT is based on the optimal partitioning of a set of items in such a way that the items with positive conditional covariances are grouped in the same clusters while the items with negative conditional covariances are grouped in different clusters. The goal is to find the partition that maximizes the DETECT value. The number of clusters in the optimum partition gives the estimated number of traits underlying the data. Kim (1995) proposed the following quantity for a pre-specified partitioning of a set of items:

$$\hat{D}(\mathrm{P}) = \frac{2}{n\ (n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \upsilon_{ij} \hat{C}(i, j \mid \boldsymbol{\theta}), \tag{51}$$

where υ_{ij} equals 1 if the *i*th and *j*th items are in the same cluster and -1 otherwise, and $\hat{C}(i, j \mid \boldsymbol{\theta})$ is the conditional covariance estimate between the *i*th and *j*th items, defined as

$$\hat{C}^1 = \sum_{k=0}^{n} \frac{J_k}{N} \hat{C}\,(i, j \mid S = k), \tag{52}$$

$$\hat{C}^2 = \sum_{k=0}^{n} \frac{J_k}{N} \hat{C}\,(i, j \mid S_{i,j} = k), \text{ and} \tag{53}$$

$$\hat{C}(i, j \mid \boldsymbol{\theta}) = \frac{\hat{C}^1 + \hat{C}^2}{2}, \tag{54}$$

where S is the total score with all items, $S_{i,j}$ is the rest-score excluding items i and j, and J_k is the number of students with a score of k.

Based on theory, DETECT is expected to reach a maximum value for the true partitioning of items when the data is multidimensional. Several cut-off criteria are proposed to evaluate the magnitude of the DETECT index. For instance, Kim (1996) classified the DETECT indices from 0 to 0.19 as an indicator of unidimensionality, 0.20 to 0.39 as an indicator of weak multidimensionality, 0.40 to 0.79 as an indicator of moderate multidimensionality, and above 0.80 as an indicator of strong multidimensionality. Stout, Nandakumar, and Habing (1996) proposed a slightly different classification by assigning the intervals (0, 0.10), (0.10, 0.50), (0.50, 1), (1, 1.50), and above 1.50 respectively to unidimensionality, weak, moderate, strong, and very strong multidimensionality. Based on a simulation study, Roussos and Ozbek (2006) recommended using 0.20, 0.40, and 1 as cut-off criteria for very weak, weak, moderate, and strong multidimensionality. Even though these suggestions are helpful in estimating the amount of multidimensionality in the data, researchers are mostly interested in finding the number of traits or correct partitioning of the items based on the latent traits.

In the explanatory framework, the total number of partitions for n-element[12] is equal to a Bell number in mathematics and increases incredibly as the number of elements increases. For instance, the number of possible partitions reaches 115,975 for 10 items, and finding the number of traits underlying the data becomes an optimization problem by finding the correct partitioning of *n* items with the highest DETECT value. Kim (1995) originally proposed using some prior judgments with the help of cluster analysis to begin, but no solution was given to find the $DETECT_{max}$ until a scientifically sound solution was developed (Zhang, 1996; Zhang & Stout, 1999). Zhang (1996) first developed the theoretical justification for DETECT, and then transferred the idea of *genetic algorithm* from biostatistics for an optimization search of the maximum DETECT value among all possible partitions of a set of items. In this optimization process, an informed choice of a partition is specified by the user (e.g., based on cluster analysis) to start, and then the genetic algorithm is used to find the optimum partitioning that

[12] Five possible partitions for three items are {(1), (2), (3)}, {(1), (2,3)}, {(2), (1,3)}, {(3), (1,2)}, {(1,2,3)}.

maximizes the DETECT value. The number of partitions is expected to be the number of dimensions underlying the data.

One of the assumptions when deriving the theoretical justification for the DETECT index is that the items have an *approximate simple structure*. In the approximate simple structure, items are expected to load primarily on one of the dimensions and to load relatively less on the other dimensions. Zhang (1996) also showed that the ratio of the maximum DETECT value to the observed DETECT value can be used to assess whether or not the assumption of approximate simple structure holds. Observed DETECT value is computed by assuming that a set of items is unidimensional. This ratio ranges from 0 to 1; higher values are an indicator of simpler structure and .8 is recommended as a cut-off for approximate simple structure (Zhang & Stout, 1999). However, it has been found in a simulation study that the ratio index is not very effective at differentiating between simple and semi-complex structures, and it is hard to find a cut-off point that applies to all conditions (Finch, Stage, & Monahan, 2008).

Dimensionality Assessment Studies

Given that there is a wide variety of different methods proposed in the literature to determine the number of latent traits, researchers have been studying the performance of these procedures under different conditions and developing guidelines for practitioners. The empirical studies that examine the performance of the wide variety of methods in dimensionality assessment can be summarized in three groups: dimensionality assessment of multivariate continuous data, testing the assumption of unidimensionality for multivariate dichotomous data, and multidimensionality assessment for multivariate dichotomous data. The studies focusing on determining the necessary number of latent traits to model continuous outcomes mostly appear in the factor analytic literature. Then, we see an extensive number of studies that only focus on testing the assumption of unidimensionality for multivariate dichotomous data, and these studies appear in the IRT literature. However, they do not provide information beyond the first dimension once the unidimensionality assumption is rejected. Finally, we see a small amount of work that focuses on determining the necessary number of latent traits for multivariate dichotomous data, again in the IRT literature.

The main characteristics of the studies on dimensionality assessment of continuous data are given in Table 3. The likelihood ratio chi-square test was mostly found to extract too many factors and is not recommended. The KG rule was consistently found to be the worst procedure in deciding the number of factors to retain. The most promising approach was consistently found to be parallel analysis (PA), followed by minimum average partial correlations (MAP), in detecting the number of factors to retain. Among the objective scree tests, $Scree_{SE}$ was found to be useful, while $Scree_{CNG}$ and $Scree_{MR}$ are not recommended for any use. PA, MAP, and $Scree_{SE}$ are highly recommended by most researchers.

The second group of studies attempted to address the assumption of unidimensionality. These studies simulated unidimensional and two-dimensional data to examine the empirical Type I error rates and the power of the analytical procedures under a variety of conditions. In these studies, the empirical power was the proportion of simulated two-dimensional datasets in which the null hypothesis of unidimensionality was rejected, and the empirical Type I error rate was the proportion of simulated unidimensional datasets in which the null hypothesis of unidimensionality was rejected. The main characteristics of these studies are given in Table 4 below.

Most studies were published by Stout and his students regarding the DIMTEST procedure developed to test the null hypothesis of essential unidimensionality for dichotomously scored items. After the DIMTEST was originally proposed (Stout, 1987), subsequent revisions were made to improve its efficiency (Nandakumar & Stout, 1993, Froelich, 2000, Froelich & Stout, 2003). In most studies, DIMTEST was found to be very powerful even for conditions in which two dimensions highly correlate while holding the Type I error rate below its nominal level. For instance, Froelich (2000) reported that the DIMTEST rejected the null hypothesis almost all of the time for both levels of inter-trait correlations (0.3 and 0.7) when the data was two dimensional with a perfect simple structure. When the data were two dimensional with approximate simple structure, DIMTEST rejected the null hypothesis 99% of the time for the correlation of .3, but the

Table 3. *Simulation Research on the Performance of Analytical Procedures in Assessing Dimensionality of Continuous Outcomes*

Study	Sample Size	Number of Items	Number of Factors	Generating Model	Inter-factor correlation	Factor Structure	Assessment Method	Number of Conditions	Replication
Humphreys & Ilgen (1969)	215,286, 710,437	7,9,11	-	CFM	-	-	PA	8	1
Humphreys & Montanelli (1975)	100,500	20,40	7	CFM	-	-	PA, G^2_{dif}	6	50
Cattell & Vogelmann (1977)	-	8,20,40	2, 5, 8, 10, 12, 20	CFM	-	S, C	KG, Scree	15	1
Revelle & Rocklin (1979)	50,100, 150,200	24	1, 2, 3, 4	CFM	-	S	PA, G^2_{dif}, KG, VSS	16	2
Zwick & Velicer (1982)	75, 150, 450	36,72,144	3, 6, 12	PCM	-	S	KG, Scree, MAP	48	10
Hakstian & Rogers (1982)	150, 400	12, 30,50	3, 6, 12, 8, 20	CFM	-	S, C	KG, Scree, G^2	288	3
Zwick & Velicer (1986)	72,180, 144,360	36,72	3, 6, 9	PCM	-	S, C	KG, MAP, Scree, PA	96	5
Mumford et al. (2003)	-	-	3, 5, 7	CFM	0,.3,.5	-	KG, MAP, $Scree_{SE}$, $Scree_{CNG}$, $Scree_{MR}$, PA	540	10,000
Piccone (2009)	250,500, 1000	-	1,2,3,4,5,8,10	CFM	0, .2, .4	C	MAP, PA, $Scree_{SE}$, KG	-	-

Note1. CFM: Common Factor Model; PCM: Principal Component Model
Note2. S: simple structure; APS: approximate simple structure; C: complex structure
Note3. PA: Parallel Analysis; G^2: Likelihood ratio test; G^2_{dif}: Likelihood ratio chi-square difference test, MAP: minimum average partial correlations; KG: Kaiser-Guttman criteria, VSS: Very Simple Structure criterion

power decreased to 79% for the correlation of 0.7. DIMTEST held the Type I error rate below the nominal level (α=0.05) for all but two of the 36 conditions related to unidimensional data. The average rejection rate was 3.6%. So, the findings suggested that DIMTEST was very conservative in terms of its Type I error rate.

Nandakumar and Yu (1996) examined the degree of robustness of the DIMTEST procedure against six different types of non-normal ability distributions and reported that the power of the DIMTEST procedure was not affected by the type of ability distribution. The average power across all conditions was above 0.9 for all types of ability distribution except when the correlation among the factors was 0.8. The power decreased to .66 for the 30-item test and to 0.8 for the 50-item test when the correlation between two dimensions was 0.8. Hattie et al. (1996) reported very similar results regarding the power of DIMTEST (> 0.9) for two- and three-dimensional data when the responses followed a compensatory model. However, the power of the DIMTEST procedure decreased significantly to an average of 39% for two-dimensional data and 3% for three-dimensional data when the responses followed a noncompensatory model. They also found that the presence of guessing did not impact the power for compensatory data, but reduced the power for non-compensatory data. In contrast to the other studies, they reported that the Type I error rate was inflated and was on average 15% across all conditions.

A few studies compared the NOHARM-based statistics to the DIMTEST procedure. In one of these studies, Breithaupt (1996) found that the NOHARM-based AChi slightly outperformed the DIMTEST procedure in terms of power when the correlation among the traits was 0 and 0.5., about 95% versus 85% on average. But, when the correlation among the traits was increased to 0.7, DIMTEST clearly outperformed AChi, about 42% versus 16% on average. The low power of both methods was mainly due to the presence of guessing. Both methods performed poorly in rejecting the null hypothesis for two-dimensional data when guessing was present in the model (below 25% on average), and held the Type I error rate below its nominal level across all conditions. In a similar study, Finch and Habing (2007) compared the performance of the three NOHARM-based procedures (AChi, ALR, T_M) to DIMTEST.

Table 4. *Simulation Research on the Performance of Analytical Procedures in Testing the Assumption of Unidimensionality for Dichotomous Outcomes*

Study	Sample Size	Number of Items	Generating Model	Number of Factors	Inter-factor correlation	Factor Structure	Assessment Method	Number of Conditions	Replication
Hambleton & Rovinelli (1986)	1500	40	U3PL, N-M3PL	1, 2	.1, .6	S	PA, Scree, NOHARM	5	1
Stout (1987)	750, 2000, 20000	25, 30, 40, 50	U3PL, M3PL	1, 2	0	APS	DIMTEST, HR	12	100
Nandakumar (1991)	2000	25, 40, 50	U3PL, M3PL	1, 2	.3, .7	APS	DIMTEST	7	1
Nandakumar & Stout (1993)	750, 2000	50	U3PL, M3PL	1, 2	.5, .7	APS	DIMTEST	30	100
Seraphine (1994)	500, 1000, 1500	25, 50	M2PL	2	.3, .7	APS	DIMTEST, HR, G^2_{dif}	24	100
Breithaupt (1996)	1000	30, 45	U3PL, M3PL	1, 2	0, .5, .7	S	AChi, DIMTEST	16	100
Hattie et al. (1996)	1000	35	U3PL, M3PL, N-M3PL	1,2	.1, .3, .5	S	DIMTEST	28	15
Nandakumar & Yu (1996)	1000, 1500	30, 50	U3PL, M3PL	1, 2	.3, .6, .8	APS	DIMTEST	98	100
De Champlain & Gessaroli (1998)	250, 500, 1000	20, 40	U2PL, M2PL	1, 2	0, .7	C	AChi, G^2_{dif}	24	100
Froelich (2000)	750, 1000, 1500, 2000	25, 40, 80	U3PL, M2PL	1, 2	.3, .7	S, APS	DIMTEST	252	100

Table 4 (Continued)

Study	Sample Size	Number of Items	Generating Model	Number of Factors	Inter-factor correlation	Factor Structure	Assessment Method	Number of Conditions	Replication
Seraphine (2000)	1500	50	U2PL, M2PL	1, 2	.3	C	DIMTEST	56	200
Tate (2003)	2000	60	U1PL, M1PL	1, 2	.6	S	AChi,G^2_{dif}, DIMTEST, DETECT	9	1
Weng & Cheng (2005)	50,100,200,500,1000	8, 20	U2PL	1	-	-	MPA	150	500
Walker et al. (2006)	500, 1500, 2500	40	M2PL	2	.3, .6, .9	APS	DIMTEST	630	100
Finch & Habing (2007)	1000, 2000	15, 30, 60	U3PL, M2PL, M3PL	1, 2	0, .3, .8, .95	S	DIMTEST, ALR, AChi, T_M	270	500
Finch & Monahan (2008)	250, 500, 100, 2000	15, 30, 60	U3PL, M2PL,M3PL	1, 2	0, .3, .8, .95	S	DIMTEST, MPA	108	500
Tran & Forman (2009)	250, 500, 100	10	U2PL	1	-	-	MPA	9	1000

Note 1. U1PL, U2PL, and U3PL represent unidimensional one-, two-, and three-parameter logistic models; M1PL, M2PL, and M3PL represent compensatory multidimensional one-, two-, and three-parameter logistic models; M2PO represents the multidimensional two-parameter normal ogive model; N-M3PL represents the Sympson's noncompensatory multidimensional three-parameter logistic model.

Note 2. S: simple structure; APS: approximate simple structure; C: complex structure

Note 3. PA: Parallel Analysis; MPA: Modified Parallel Analysis; HR: Holland and Rosenbaum Procedure; G^2_{dif}: TESTFACT-based approximate likelihood ratio chi-square difference test; NOHARM: residual examination based on NOHARM output; AChi: NOHARM-based approximate chi-square test; ALR: NOHARM-based approximate likelihood ratio test; T_M is NOHARM-based mean-adjusted chi-square statistics

They found that the NOHARM-based procedures outperformed DIMTEST in terms of power when guessing is not included in the response model, while they had lower Type I error rates than DIMTEST. Among the NOHARM-based procedures, the AChi was the most efficient in terms of power and Type I error rate. On the other hand, when the guessing is present in the model, NOHARM-based statistics had extremely inflated Type I error-rates ($0.1 < \alpha < 0.4$), while DIMTEST still held it at a reasonable level ($.007 < \alpha < .071$) for the nominal alpha level of 0.05. The higher error rates of NOHARM-based statistics were observed regardless of the magnitude of the guessing parameter provided to the NOHARM program in estimation. Their results supported the previous studies regarding the conclusion that both NOHARM-based statistics and DIMTEST were very powerful even for inter-trait correlations as high as 0.8 (> 0.9 for two-parameter models and 0.7 for three-parameter models[13]).

Parallel analysis was shown to be a promising method for assessing the dimensionality of continuous outcomes in the factor analytic literature, and it was modified for dichotomous outcomes to test the unidimensionality assumption (Drasgow & Lissak, 1983). Later, Weng and Cheng (2005) and Tran and Forman (2009) generated one-dimensional dichotomous data and assessed the performance of modified parallel analysis (MPA) by applying the original PA procedure to principal component eigenvalues of tetrachoric correlations. Both studies reported that when the principal component eigenvalues of sample tetrachoric correlations were compared to the 95^{th} or 99^{th} percentile of the principal component eigenvalues of random tetrachoric correlations, MPA accurately identified one factor above 95% in almost all conditions unless the sample size was below 250 and the items had extreme difficulties (> 0.9 or <0.1). In those cases, the difficulty in estimating tetrachoric correlations led to the poor performance of MPA. Finch and Monahan (2008) used a slightly different approach to parallel analysis and compared its performance to DIMTEST in testing the null hypothesis of unidimensionality. They first fitted a unidimensional model to the sample data under examination, and then generated 100 datasets using the same item and person

[13] After adjusting the power of NOHARM-based statistics for inflated Type I error rates.

parameter estimates. Then, they extracted the principal axis eigenvalues from the tetrachoric correlations of each synthetic dataset, and the sampling distribution was obtained for the second eigenvalue in the presence of the first real dimension. Finally, if the 95th percentile of the sampling distribution for the second eigenvalue was higher than the second sample data eigenvalue, it would be decided that the data was not unidimensional and that there was at least a second significant factor. They found that MPA, as they applied it, had lower Type I error rates than the nominal level in all conditions, while DIMTEST had inflated Type I error rates for smaller sample sizes and shorter tests. Also, MPA had comparable power to DIMTEST in terms of rejecting unidimensionality for two-dimensional data. Across all conditions, the power of DIMTEST was above .95 for 46 conditions, and above .85 for six conditions, while MPA's power was above .95 for 43 conditions, and above .85 for seven conditions out of 60. Power decreased below .7 for both methods when the sample size was smaller than 500, the correlation among the traits was above .7, and the number of items was smaller than 15.

Two studies investigated the performance of the approximate likelihood ratio chi-square difference test based on full information factor analysis (G^2_{diff}) as implemented by TESTFACT, and the results were not encouraging. De Champlain and Gessaroli (1998) reported very high power, but the results were not reliable due to the extremely inflated Type I error rates. G^2_{diff} had a Type I error rate of .17 in the best case (n=1000) and .59 in the worst case (40 items and 250 examinees). Similarly, Seraphine (1994) reported very high power for G^2_{diff}, but this result was not reliable either due to a lack of Type I error rate examination in the study.

Multidimensionality Assessment of Dichotomous Data

There are relatively few studies on the performance of analytical methods in determining the number of dimensions underlying dichotomous data, and the main characteristics of these studies are summarized in Table 5. In one of these studies, DETECT was found to be highly powerful for identifying the number of underlying dimensions for dichotomously scored data. Across all conditions manipulated, the power to detect the "true" number of dimensions used to generate data was above 97% (Zhang

& Stout, 1999). Finch and Habing (2005) compared the performance of DETECT to the performance of the NOHARM-based approximate likelihood ratio chi-square test (ALR). For two-dimensional data, the recovery was very close to perfect for both methods when the correlation between the traits was 0 and .3 for no-guessing conditions. As the correlation increased to .8 and .95, both methods overestimated the number of dimensions, especially in the presence of guessing. The average estimated numbers of dimensions for ALR and DETECT across all conditions were 2.30 and 2.46 with no guessing, and 2.67 and 2.48 with guessing, respectively. For six-dimensional data, the recovery was again very close to perfect for both methods when the correlation between the traits was 0 and .3 for no-guessing conditions. But in contrast to two-dimensional data, both methods underestimated the number of dimensions as the correlation increased to .8 and .95. The average estimated number of dimensions for ALR and DETECT were 5.82 and 5.17 with no guessing, and 5.58 and 5.53 with guessing, respectively. ALR seemed more resistant to the higher inter-trait correlations and the presence of guessing in terms of recovering the true dimensionality.

In a recent study, Svetina (2011) compared NOHARM-based RMSR, ALR, and AChi statistics to DETECT. For compensatory two-dimensional data, DETECT was both more accurate and more consistent than the NOHARM-based methods, especially for the inter-trait correlation below .6 and larger sample size. In conditions where the number of complex items is 30% or less and the inter-trait correlation is .75 or smaller, all methods were successful at identifying the "true" number of dimensions. For compensatory three-dimensional data, the same pattern occurred as the two-dimensional data, but remarkably, the performance of all methods deteriorated as the inter-trait correlations went above .3. In general, as the number of complex items increased, the performance deteriorated for all methods, but more markedly for NOHARM-based statistics. On the other hand, ALR and AChi generally outperformed DETECT for non-compensatory data in identifying the "true" number of dimensions. ALR and AChi accurately identified the "true" number of dimensions, especially for lower correlations when the sample size was 2000. DETECT is not recommended if the researcher is expecting the data to be non-compensatory.

Other researchers studied whether MPA and MAP were useful in recovering the true number of dimensions underlying binary items. The results were very promising, even though they are available for limited conditions only. For instance, Crawford et al. (2010) reported that MPA correctly identified the number of dimensions for one-, two-, and four-dimensional data almost all the time when the sample size was 500 or above, the average factor loadings were 0.5 or above, and there were six items per factor in a simple factor structure. As the sample size, average factor loading, and number of items per factor decreased, the proportion of correctly identifying the number of dimensions decreased. Results reported by Garrido et al. (2011) were also encouraging for MAP as a factor decision rule for categorical data. They reported that the accuracy was close to 100% for the conditions when the number of variables per factor was more than four while average factor loading was .75, or when the number of variables per factor was more than seven while the average factor loading was .5.

Only one study included a likelihood ratio chi-square test based on full-information factor analysis (G^2), RMSR based on iterative principal factor analysis as implemented by MicroFact, and multidimensional scaling (MDS) in order to compare their performance with the NOHARM-based RMSR statistic (Nichol, 2011). For one-dimensional data, NOHARM-based RMSR was highly successful (>95%) in recovering the true number of dimensions for both compensatory and non-compensatory data, while all other methods significantly over-factored. For two-dimensional data, RMSR based on IPFA was the best performer in detecting two dimensions, and the accuracy increased as the number of items increased; none of the other methods performed well. In the worst case, the proportion of correct decisions by IPFA-based RMSR was 47% for compensatory and 73% for non-compensatory data when the number of items was 30 and the ratio of the strength of the first factor to the strength of the second factor was two. In the best case, the proportion of correct decisions by IPFA-based RMSR was 99% for compensatory and 79% for non-compensatory data when the number of items was 60 and the ratio of the strength of the first factor to the strength of the second factor was 1.3. For three-dimensional data, the proportion of correct decisions for any method never

Table 5. *Simulation Research on the Performance of Analytical Procedures in Assessing Multidimensionality of Dichotomous Outcomes*

Study	Sample Size	Number of Items	Number of Factors	Generating Model	Inter-factor correlation	Factor Structure	Assessment Method	Number of Conditions	Replication
Zhang & Stout (1999)	400, 800	20, 40	2, 3, 4	M2PL	-	APS	DETECT	12	100
Finch & Habing (2005)	1000, 2000	15, 30, 60	2, 6	M2PL M3PL	0, .3, .8, .95	APS	ALR, DETECT	60	500
Cho et. al. (2009)	200, 800	-	3, 8	M2PL	.3, .7	S	MPA	64	100
Crawford et. al. (2010)	100, 500	-	1, 2, 4	M2PL	0, .4, .6, .7, .8	-	MPA	138	1000
Garrido et al. (2011)	250, 500, 1000	4, 8, 12*	2, 4, 6	M2PL	0, .3, .5	S	MAP	4,374	100
Nichol (2011)	2000	30, 60	1, 2, 3	M2PL N-M2PL	0	C	IPFA, NOHARM, G^2, MDS	34	75
Svetina (2011)	500, 1000, 2000	10, 20*	2, 3	M3PL N-M3PL	0, .3, .6, .75, .9	S, C	DETECT, ALR, AChi, NOHARM	480	500

* These numbers are variable per dimension

Note 1. M2PL and M3PL represent compensatory multidimensional two- and three-parameter logistic models. N-M3PL and N-M2PL represent the Sympson's noncompensatory multidimensional two- and three-parameter logistic models.

Note 2. S: simple structure; APS: approximate simple structure; C: complex structure

Note 3. PA: Parallel Analysis; G^2: TESTFACT-based likelihood ratio test; NOHARM: residual examination based on the NOHARM output; AChi: NOHARM-based approximate chi-square test; ALR: NOHARM-based approximate likelihood ratio test; MDS: multidimensional scaling; IPFA: residual examination based on iterative principal factor analysis

exceeded .4 for any condition, except the MDS-based RMSR index, when the number of items was 30 in non-compensatory conditions.

AIC and BIC. Although AIC and BIC have recently gained attention in IRT model selection, there is no study that has investigated the model selection behavior of AIC and BIC in the context of dimensionality assessment of dichotomous data. The relevant research focused on the performance of AIC and BIC in choosing the best fitting model among the unidimensional 1-PL, 2-PL, and 3-PL dichotomous IRT models (Kang & Cohen, 2007), among the unidimensional polytomous IRT models (Kang, Cohen, & Sung, 2009), the combination of unidimensional dichotomous and polytomous IRT models for mixed-format tests (Whittaker, Chang, & Dodd, 2012), in identifying the number of latent classes for mixture IRT models (Li, Cohen, Kim, & Cho, 2009; Preinerstorfer & Formann, 2012), and in identifying cross-level two-way differential item functioning (Patarapichayatham, Kamata, & Kanjanawasee, 2012). These studies consistently found that BIC tends to select simpler underparameterized models, whereas AIC tends to select more complex overparameterized models. This finding is also consistent with the findings of the studies in other contexts. For instance, Burnham and Anderson (2002) simulated a chain binomial model with seven parameters. As a result of the simulation study, they found that AIC selected a model with 7.6 parameters on average, whereas BIC selected a model with 5.1 parameters on average. Also, 95% of the models selected by BIC contained 4, 5, or 6 parameters, while 95% of the models selected by AIC contained between 5 and 13 parameters.

Research Questions

Although the literature related to the dimensionality assessment of latent structures underlying item response data is very broad, and many studies have addressed the issue in several aspects, all these studies share a major weakness. The previous simulation studies assumed that a simpler true model that the item responses follow at the population level exists, and this model is among the fitted models under examination. MacCallum (2003) criticized this dominating approach in the literature and stated that it is not very interesting to learn how methods perform when the models under examination are correct at the population level (no model misspecification). From a practical point of

view, it would be more interesting to investigate how the methods perform when the true model at the population level is very complex in many aspects and none of the models under examination reflect the truth (imperfect models). While MacCallum (2003) made this criticism in the context of latent trait models, a similar criticism was also made by Burnham and Anderson (2002) in the area of model selection with generalized linear models in biological sciences.

> "*People have often (mis)used Monte Carlo methods to study the various criteria, and this has been the source of confusion in some cases. In Monte Carlo studies, one knows the generating model and often considers it to be "truth." The generating model is often quite simple, and it is included in the set of candidate models. In the analysis of the simulated data, attention is (mistakenly) focused on what criterion most often finds this true model. Under this objective, we would suggest the use of the dimension consistent criteria in this artificial situation, especially if the order of the true model was quite low (e.g., K = 3–5), or the residual variation (σ^2) was quite small, or the sample size was quite large. However, this contrived situation is far from that confronted in the analysis of empirical data in the biological sciences. Monte Carlo studies to evaluate model selection approaches to the analysis of real data must employ generating models with a range of tapering effect sizes and substantial complexity. Such evaluations should then focus on selection of a best approximating model and ranking of the candidate models; the notion that the true (in this case, the generating) model is in the set should be discarded* (p. 287)."

Given the wide variety of methods proposed in the literature for determining the number of latent traits and the criticism made by MacCallum (2003) and Burnham and Anderson (2002) for previous simulation research on the performance of these methods, the goal of the current study is to discuss the following research questions:

1) How does parallel analysis perform in the dimensionality assessment of dichotomous outcomes when the generating model has a complex factor structure and includes many minor tapering latent factors in addition to major latent factors?

   Hypothesis: As discussed before, the magnitude of the first principal-axis eigenvalue reflects the total variance accounted for by the structure and implies an imaginary general factor when the underlying factor structure is complex. In those cases, the eigenvalues beyond the first eigenvalue do not account for more than

chance. Therefore, it is hypothesized that parallel analysis will favor one-dimensional models in most cases regardless of the underlying true structure when the true factor structure is complex.

2) How do the chi-square statistics obtained from different estimation approaches perform in the dimensionality assessment of dichotomous outcomes when the generating model has a complex factor structure and includes many minor tapering latent factors in addition to major latent factors?

   Hypothesis: The chi-square statistics test whether or not a specified model exactly fits to sample data assuming the specified model is correct in population and the sample size is large enough. Because they test exact fit, they are known to be too powerful for minor discrepancies between the specified model and "true" model. It is not surprising that they select unnecessarily complex models due to the statistical power. Therefore, it is hypothesized that the chi-square statistics will identify at least the latent factors with major influences, but they may select unnecessarily complex models due to the presence of latent factors with minor effects.

3) How do the alternative fit indices such as SRMR, RMSEA, CFI, AIC, and BIC perform in the dimensionality assessment of dichotomous outcomes when the generating model has a complex factor structure and includes many minor tapering latent factors in addition to major latent factors?

   Hypothesis: The goal of alternative fit indices is to find the best approximating model rather than to find the "true model." They acknowledge the model misspecification to some degree and intend to avoid including unnecessary minor influences in the model by incorporating a penalty term for model complexity in selecting the best approximating model. Therefore, it is hypothesized that alternative fit indices will identify the number of latent traits with major influences.

CHAPTER 3: METHODOLOGY

There are two studies in the current research. Each study has its own design, but the dimensionality assessment criteria used for the analyses are all the same. The criteria considered in the current study are parallel analysis, revised parallel analysis, DETECT, adjusted and unadjusted chi-square statistics (AChi, ALR, T_M, and T_{MV}) obtained from NOHARM estimation, adjusted chi-square statistics (T_M^2, T_{MV}^2) obtained from the WLSM and WLSMV estimators available in Mplus 6.12 (Muthén & Muthén, 1998 – 2010), chi-square difference test, AIC, AIC_c, and BIC based on marginal maximum likelihood (MML) estimation as implemented by the MLR estimator in Mplus 6.12 (Muthén & Muthén, 1998 – 2010), and the RMSEA, SRMR, and CFI fit indices obtained from the WLSM and WLSMV estimators available in Mplus.

Study 1

Dataset. Data for the first study were the Minnesota Basic Skills Test taken by 67,510 eighth-grade students in mathematics and reading in 2005. The datasets comprised 75 and 40 common items administered to all students in mathematics and reading, respectively. The mathematics test had eight content strands and the reading test had two content strands. Also, the items in the reading test were asked in the contexts of five different reading passages. The distribution for the number of items in both tests across different content areas is given in Table 6. Both the reading and mathematics tests were not timed.

Both the reading and mathematics tests were very easy. The average item difficulty was 0.82 with a standard deviation of 0.10 for the mathematics test, and was 0.86 with a standard deviation of 0.09 for the reading test. The most difficult items had proportion corrects of 0.560 and 0.597 for the mathematics and reading tests, respectively. Overall, both tests had moderately discriminating items. The average point biserial correlation was 0.42 with a standard deviation of 0.11 for the mathematics test, and was 0.44 with a standard deviation of 0.07 for the reading test. The minimum point biserial correlations were .12 and .23, and the maximum point biserial correlations were .56 and .60 for the mathematics and reading tests, respectively. Table 7 summarizes the

item statistics for all items in the reading and mathematics tests. The number-correct scores were negatively skewed, and the students had generally high scores on both tests. The average number-correct score was 61.39 (82% correct) with a standard deviation of 11.64 for the mathematics test, and was 34.49 (86% correct) with a standard deviation of 5.59 for the reading test. Figure 1 shows the density plot for the distribution of number-correct scores.

Study Design. Subtests with 20 and 40 items were created and labeled RS, RL, MS, and ML with the first letter indicating mathematics (M) or reading (R) and the second letter indicating short (S) or long (L). The RL test included all 40 items in the original reading test. The RS test had 20 items purposefully selected from 40 items in the original reading test. Five items were selected from each reading passage, excluding the fourth reading passage. The fourth reading passage was excluded because most of the items in the fourth reading passage were very easy with proportions correct above .95. Also, the items were selected such that each passage had a mix of inferential and literal comprehension items.

Table 6. *Distribution of Number of Items in the Mathematics and Reading Tests Across Content Areas*

Mathematics Test		Reading Test		
Content Area	Number of Items		Literal Comprehension	Inferential Comprehension
Whole Number	15	Passage 1	5	2
Percentage and Ratio	10	Passage 2	5	3
Number Sense	7	Passage 3	5	4
Estimation	8	Passage 4	7	2
Measurement	9	Passage 5	4	3
Tables and Graphs	11			
Chance and Data	8			
Space and Shape	7			
Total	75	Total	26	14

Table 7. *Summary Item Statistics for 2005 Minnesota Basic Skills Reading and Mathematics Tests*

	Mathematics							Reading			
Item No	N	Item Diff.	P.Bis. Cor.	Item No	N	Item Diffi.	P.Bis. Cor.	Item No	N	Item Diff.	P.Bis. Cor.
1	67481	0.93	0.24	41	67476	0.93	0.43	1	67890	0.98	0.23
2	67455	0.73	0.31	42	67454	0.66	0.53	2	67884	0.83	0.49
3	67478	0.83	0.31	43	67468	0.96	0.45	3	67847	0.88	0.46
4	67495	0.91	0.29	44	67459	0.79	0.35	4	67870	0.90	0.41
5	67478	0.82	0.48	45	67471	0.78	0.38	5	67873	0.82	0.50
6	67484	0.77	0.45	46	67459	0.91	0.50	6	67878	0.96	0.38
7	67481	0.79	0.34	47	67472	0.93	0.36	7	67865	0.89	0.52
8	67490	0.95	0.40	48	67468	0.84	0.39	8	67787	0.95	0.30
9	67494	0.91	0.39	49	67465	0.56	0.39	9	67850	0.96	0.48
10	67480	0.85	0.47	50	67476	0.71	0.48	10	67838	0.88	0.36
11	67452	0.73	0.40	51	67464	0.93	0.37	11	67857	0.83	0.38
12	67486	0.65	0.51	52	67460	0.79	0.55	12	67814	0.68	0.48
13	67470	0.75	0.46	53	67425	0.68	0.56	13	67886	0.81	0.37
14	67473	0.88	0.31	54	67471	0.89	0.49	14	67877	0.94	0.43
15	67491	0.95	0.33	55	67461	0.82	0.47	15	67865	0.90	0.42
16	67498	0.99	0.12	56	67427	0.71	0.50	16	67814	0.82	0.49
17	67470	0.87	0.43	57	67472	0.87	0.45	17	67785	0.86	0.40
18	67466	0.87	0.23	58	67460	0.81	0.41	18	67813	0.89	0.50
19	67503	0.98	0.25	59	67410	0.65	0.58	19	67835	0.93	0.43
20	67499	0.84	0.51	60	67477	0.88	0.22	20	67843	0.85	0.32
21	67459	0.76	0.49	61	67441	0.78	0.50	21	67847	0.93	0.52
22	67471	0.95	0.41	62	67459	0.67	0.40	22	67825	0.91	0.34
23	67499	0.87	0.36	63	67457	0.87	0.38	23	67847	0.87	0.48
24	67490	0.77	0.54	64	67459	0.77	0.59	24	67840	0.72	0.45
25	67486	0.97	0.31	65	67464	0.61	0.46	25	67821	0.98	0.36
26	67489	0.91	0.28	66	67458	0.75	0.26	26	67850	0.97	0.46
27	67448	0.71	0.42	67	67459	0.83	0.46	27	67626	0.92	0.54
28	67479	0.98	0.21	68	67460	0.67	0.51	28	67839	0.94	0.44
29	67483	0.85	0.55	69	67438	0.76	0.24	29	67839	0.76	0.40
30	67480	0.76	0.40	70	67470	0.74	0.32	30	67857	0.91	0.51
31	67493	0.96	0.40	71	67426	0.65	0.49	31	67831	0.89	0.49
32	67467	0.85	0.52	72	67465	0.84	0.43	32	67778	0.89	0.50
33	67475	0.74	0.49	73	67387	0.77	0.40	33	67838	0.97	0.46
34	67485	0.95	0.31	74	67452	0.75	0.57	34	67723	0.82	0.51
35	67482	0.84	0.54	75	67440	0.85	0.60	35	67726	0.68	0.44
36	67464	0.75	0.43					36	67810	0.89	0.56
37	67483	0.97	0.37					37	67801	0.85	0.44
38	67446	0.83	0.55					38	67712	0.60	0.44
39	67492	0.72	0.54					39	67798	0.83	0.33
40	67491	0.93	0.49					40	67828	0.64	0.38

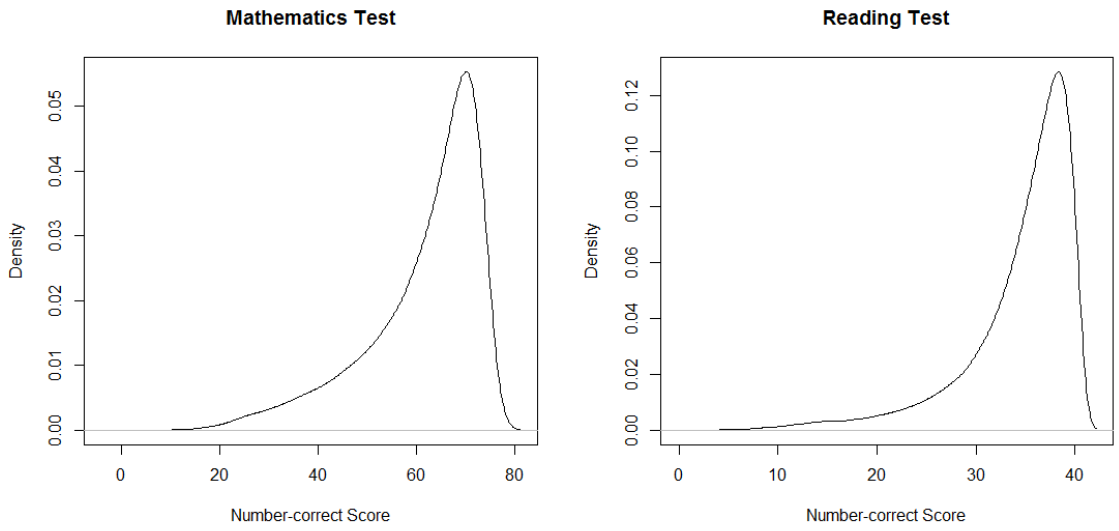


Figure 1. The Density of Distributions for the Number-Correct Scores in the Mathematics and Reading Tests

The original mathematics test had 75 items from eight different content strands. The ML test with 40 items was created by selecting items from only four content strands with approximately an equal number of items. Then, the MS test with 20 items was created by selecting five items from each of the four content strands included in the ML test. Table 8 shows the distribution for the number of items among the content areas in the 20- and 40-item subtests used for the current research, and Table 9 shows the summary statistics for the items included in the ML, MS, RL, and RS subtests. Figure 2 shows the density of the distributions for the number-correct scores in each subtest. A sample of students with two different sample sizes (N=500, 1000) was repeatedly drawn 500 times from the large dataset (N=67,510) for each of the subtests. In total, 4000 sample datasets were obtained as a result of the sampling procedure. To ensure that 500 replications were enough, a running average for each statistic used in the study was computed (see Appendix A for graphical representation of these running averages). In most instances, 100 replications were enough to get stable estimates; but in some instances, around 200 replications were needed for stable estimates.

Dimensionality Assessment. Different dimensionality assessment criteria were implemented for each of the 4000 data samples and the number of latent traits suggested by these criteria was recorded for studying their performance. R code to implement the necessary procedures is given in Appendix B. The procedures are briefly summarized below.

Parallel Analysis (PA): A hundred random datasets with uncorrelated continuous variables were first generated using the same number of items and sample size for each of the 4000 sample datasets. Then, continuous variables in the random datasets were dichotomized by using the threshold estimates obtained from the item proportion-correct statistics in the associated sample dataset. So, each of the 100 random datasets has the same number of items (n), same number of people (N), and approximately the same item difficulty levels with the associated sample dataset under investigation. The eigenvalues were obtained from the tetrachoric correlation matrix for each of 100 random datasets using TESTFACT (Wood, Wilson, Gibbons, Schilling, Muraki, & Bock, 2003). Then, the empirical eigenvalue sampling distribution at each rank position was obtained for an associated sample dataset under investigation. In a similar procedure, the sample dataset under investigation was also analyzed by TESTFACT and sample eigenvalues were extracted from the sample tetrachoric correlation matrix. Finally, each sample eigenvalue was compared to the 95^{th} percentile of the empirical eigenvalue distribution, and the decision was made as the number of eigenvalues greater than the 95^{th} percentile of the eigenvalue distribution in the corresponding rank position.

Revised Parallel Analysis (RPA). In a slightly different approach, the RPA procedure runs a series of simulations to test each eigenvalue sequentially with a separate simulation. A stepwise procedure as recommended by Green et al. (2012) was applied to each of the 4000 sample datasets.

Table 8. *Distribution of Number of Items in the Mathematics and Reading Subtests Across Content Areas*

Mathematics Long Test (ML)		Reading Long Test (RL)		
Content Area	Number of Items		Literal Comprehension	Inferential Comprehension
Whole Number	12	Passage 1	5	2
Percentage and Ratio	10	Passage 2	5	3
Measurement	8	Passage 3	5	4
Tables and Graphs	10	Passage 4	7	2
		Passage 5	4	3
Total	40	Total	26	14

Mathematics Short Test (MS)		Reading Short Test (RS)		
Content Area	Number of Items		Literal Comprehension	Inferential Comprehension
Whole Number	5	Passage 1	3	2
Percentage and Ratio	5	Passage 2	2	3
Measurement	5	Passage 3	3	2
Tables and Graphs	5	Passage 5	2	3
Total	20	Total	10	10

Table 9. *Summary of Item Difficulty and Biserial Correlation Statistics for the Subtests*

	Number of Items	Average Item Diff.	Average Point Biserial Corr.	Minimum Item Diff.	Maximum Item Difficulty	Minimum Point Biserial Corr.	Maximum Point Biserial Corr.
Mathematics Long Test	40	0.799 (0.104)	0.457 (0.095)	0.560	0.967	0.227	0.610
Mathematics Short Test	20	0.748 (0.091)	0.516 (0.063)	0.560	0.870	0.433	0.627
Reading Long Test	40	0.863 (.094)	0.435 (.072)	0.597	0.983	0.232	0.561
Reading Short Test	20	0.808 (.093)	0.456 (.055)	0.597	0.904	0.345	0.524

* The standard deviations are in parentheses.

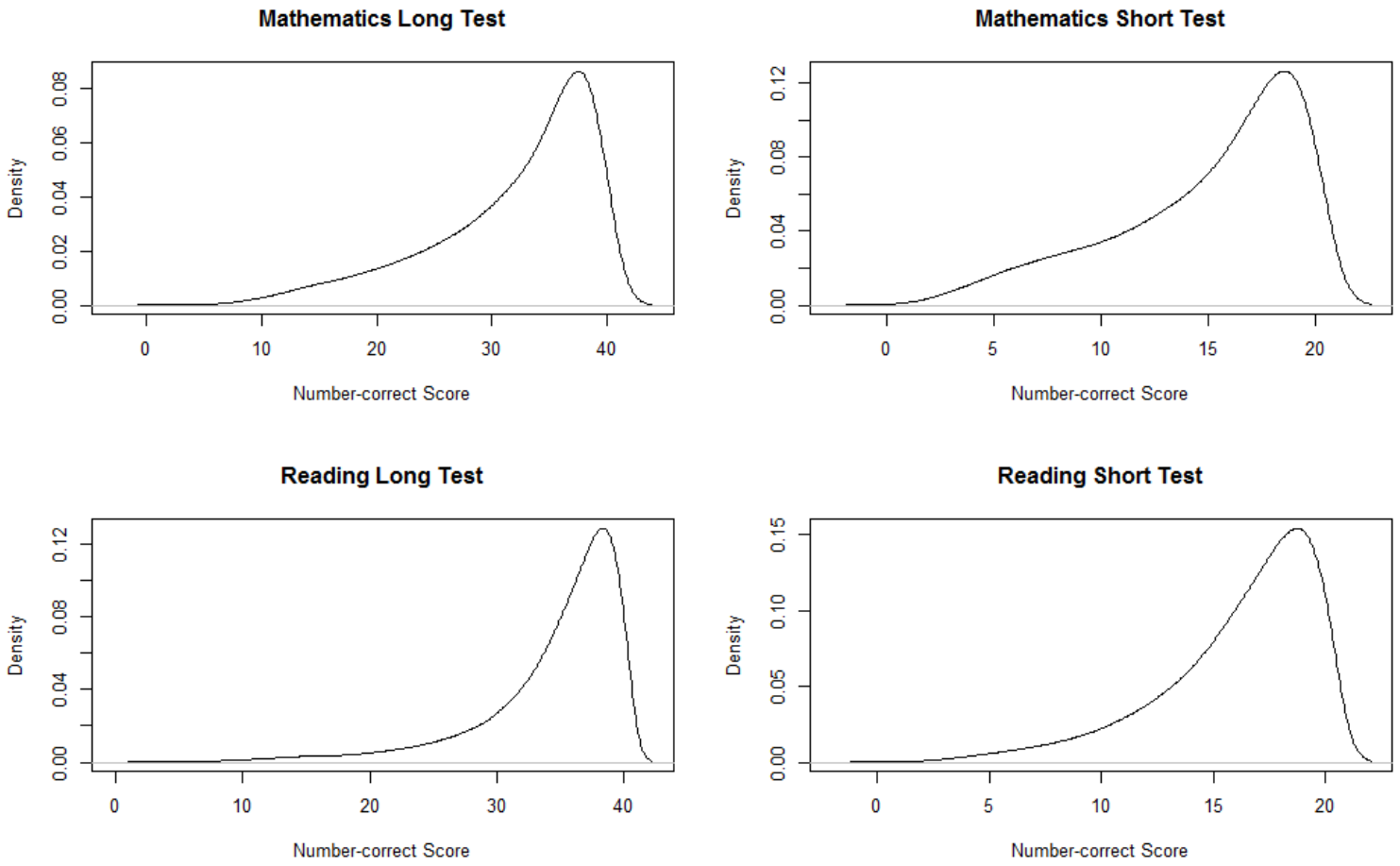


Figure 2. The Density of Distributions for the Number-Correct Scores in the Subtest

Step 1. 100 random datasets are generated and empirical sampling distribution of the first eigenvalue is derived as described in the PA procedure above. The first sample eigenvalue is compared to the 95th percentile of the empirical sampling distribution of the first eigenvalue from random datasets to test whether the first sample eigenvalue is significantly higher than the first eigenvalue from random datasets.

Step 2. If the first eigenvalue is significant, one factor is extracted from the sample dataset under investigation using MINRES, and the factor-loading matrix is obtained. A hundred datasets with continuous variables are generated with the same number of variables (n) and same number of observations (N) using the common factor model

$$\mathbf{X} = \mathbf{F}\mathbf{\Lambda}^{T} + \mathbf{U}\mathbf{E},$$

where **X** is an *N x n* latent continuous score matrix, **F** is an *N x 1* common factor score matrix with a standard normal distribution, **Λ** is an *n x 1* factor loading matrix obtained from MINRES analysis, **U** is an *N x n* unique factor score matrix and has a multivariate normal distribution with a mean vector of zeros and correlation matrix of identity, and **E** is an *n x n* diagonal matrix with item uniqueness on the diagonal. Using the thresholds estimated from item proportion-correct statistics in the sample dataset under investigation, the variables in the generated datasets are dichotomized. Each generated dataset is analyzed using TESTFACT and the eigenvalues from the tetrachoric correlation matrix are extracted to derive the empirical sampling distribution for the second eigenvalue conditioning on the magnitude of the first eigenvalue. The second sample eigenvalue is compared to the 95th percentile of the empirical sampling distribution for the second eigenvalue from generated datasets to test the significance of the second sample eigenvalue when the magnitude of the significant first eigenvalue is taken into account.

Step 3. If the second eigenvalue is significant, two-factor solutions are obtained from the sample dataset under investigation using MINRES. Using a similar procedure to Step 2, 100 datasets are generated with an *N x 2* common factor score matrix (**F**) with zero correlation among the factors, an *n x 2* unrotated factor loading matrix (**Λ**) from MINRES analysis, an *N x n* unique factor score matrix (**U**), and an *n x n* diagonal matrix with item uniqueness on the diagonal (**E**). The continuous variables are again dichotomized using the item threshold estimates from the sample dataset under investigation. Each generated dataset is analyzed using TESTFACT and the eigenvalues are extracted from the tetrachoric correlation matrix to derive the empirical sampling distribution for the third eigenvalue conditioning on the magnitude of the significant first two eigenvalues. The third sample eigenvalue is compared to the 95^{th} percentile of the empirical sampling distribution of the third eigenvalue from generated datasets to test the significance.

This sequential process continues until the *k*th step, where the *k*th sample eigenvalue is not found higher than the corresponding eigenvalue from the generated random datasets.

Unweighted Least squares estimation with NOHARM. Item thresholds and factor loading estimates for each of the 4000 sample datasets were obtained by fitting the approximate compensatory M2PO model up to the eight-dimensional solution with default options in NOHARM. The guessing parameters were fixed to zero during the model-fitting process. After obtaining the threshold and loading estimates and residual joint-proportion corrects from NOHARM, the decision regarding the number of latent traits was made using the *p*-values associated with the four proposed chi-square statistics computed from the NOHARM output. These are approximate chi-square (Achi; De Champlain, 1993), approximate likelihood ratio chi-square (ALR; Gessaroli et al., 1997), and mean-adjusted and mean-and-variance adjusted chi-square statistics (T_M^1 and T_{MV}^1; Maydeu-Olivares, 2001).

Weighted Least squares estimation with Mplus. Item thresholds and factor loading estimates for each of the 4000 sample datasets were obtained by fitting the linear common factor model to the tetrachoric correlation matrix up to the eight-dimensional solution using diagonal weighted least-squares estimation as implemented by the WLSM and WLSMV estimators available in Mplus with default options for the exploratory factor analysis of categorical outcomes. Mplus returned a *p*-value associated with mean-adjusted (T_M^2) or mean-and-variance adjusted (T_{MV}^2) chi-square statistics, a 90% confidence interval for the RMSEA fit index, and the SRMR and CFI fit indices for each factor solution from one to eight. The optimal number of latent traits was chosen based on each criterion by identifying

- the smallest number of factor solutions with a *p*-value greater than 0.05 (T_M^2 and T_{MV}^2),
- the smallest number of factor solutions with a lower bound of the RMSEA value smaller than 0.05,
- the smallest number of factor solutions with an SRMR value smaller than 0.07, and
- the smallest number of factor solutions with a CFI value greater than 0.95.

Full Information Factor Analysis in Mplus. The MLR estimator in Mplus is available to conduct the full information factor analysis with marginal maximum likelihood (MML) and the EM algorithm for categorical variables. A series of exploratory factor analyses were run using the MLR estimator in Mplus to obtain up to the six-dimensional solution for each of the 4000 sample datasets. In the model-fitting process, the number of maximum EM cycles was fixed to 250, and the EM iterations were stopped when the log-likelihood difference between two successive EM cycles was .01. Seven Gauss-Hermite quadrature points when fitting the one-, two-, three- and four-dimensional models, and four Gauss-Hermite quadrature points when fitting the five- and six-dimensional models were used to approximate the integration at the E step. The defaults were used for the convergence criterion at the M step. These settings were fixed across all runs in

the study. Mplus returned the log-likelihood value after convergence, a scaling correction factor for the log-likelihood value, AIC, and BIC values. Corrected AIC values were computed based on the information provided in the output. The necessary number of latent traits was chosen based on each criterion by identifying

- the smallest number of factor solutions with a *p*-value (>.05) associated with the log-likelihood ratio chi-square difference test with unadjusted log-likelihoods ($FIFA_1\ \chi^2$),
- the smallest number of factor solutions with a *p*-value (>.05) associated with the log-likelihood ratio chi-square difference test based on the adjusted log-likelihoods ($FIFA_2\ \chi^2$, described in http://www.statmodel.com/chidiff.shtml),
- the factor solution that provided the minimum AIC value,
- the factor solution that provided the minimum corrected AIC value, and
- the factor solution that provided the minimum BIC value.

DETECT. The DETECT analysis was run in an exploratory mode for each of the 4000 sample datasets by setting the MINCELL option to two, and the MUTATIONS option to four for the 20-item tests and to eight for the 40-item tests. The MINCELL option indicates the minimum number of examinees required to be present in any one cell when calculating the conditional covariances, and the MUTATIONS option indicates the number of vectors mutated in the genetic algorithm when maximizing the *D* value to find the optimal cluster solution. The maximum number of dimensions to be found was set to 12 for each run. DETECT returned the number of dimensions that maximized the D value as the optimal solution.

Analysis. As a result of implementing each dimensionality assessment criterion, the outcome variable was the number of latent traits recommended by each criterion for each of the 4000 sample datasets generated in eight different conditions (four subtests x two sample size). The average, standard deviation, and range of the number of latent traits

suggested by each criterion within each condition were tabulated to examine the model-selection behavior across different criteria.

Study 2

The second study was a simulation study motivated by the first. The first study was not able to give a clear idea about the performance of the dimensionality assessment criteria because the true latent structure was not known for the real datasets. The second study aimed to provide more information about the performance of the dimensionality assessment criteria in a controlled simulation environment.

Study Design. The following variables were manipulated in the simulation study: number of major latent traits and variance accounted for by the major latent traits (40 different variations), number of items (n=20, 40), sample size (N=500, 1000), inter-factor correlations (r=0, .5), and the amount of variance accounted for by minor latent traits (V=10%, 20%). A snapshot of 40 different variations for the number of major latent traits and the variance accounted for by these traits is given in Table 10. All manipulated variables were fully crossed, yielding a total number of 640 simulation conditions.

Simulation Model. All datasets were generated using the factor model proposed by Tucker, Koopman, and Linn (1969) and MacCallum and Tucker (1991). The model can be expressed as follows:

$$\mathbf{X} = \mathbf{F_M \Lambda_M}^{T} + \mathbf{UE}, \tag{55}$$

where $\mathbf{X}$ was an $N \times n$ latent continuous item score matrix, $\mathbf{F_M}$ was an $N \times (K+k)$ standardized mega factor score matrix for major and minor latent traits with K indicating the number of major latent traits and k indicating the number of minor latent traits, $\mathbf{\Lambda_M}$ was an $n \times (K+k)$ mega factor loading matrix for major and minor latent traits, $\mathbf{U}$ was an $N \times n$ unique factor score matrix, and $\mathbf{E}$ was an $n \times n$ diagonal matrix with item uniqueness on the diagonal.

The mega factor loading matrix can be expressed as follows:

$$\mathbf{\Lambda_M} = \left[\mathbf{\Lambda} \quad \boldsymbol{\lambda}\right],$$

where $\mathbf{\Lambda}$ is an *n x K* factor loading matrix for the major latent traits, and $\boldsymbol{\lambda}$ is an *n x k* factor loading matrix for the minor latent traits. The mega factor score matrix for the major and minor latent traits ($\mathbf{F_M}$) had a multivariate normal distribution with a mean vector of zeros and mega factor correlation matrix $\mathbf{\Phi_M}$ expressed as the following (Hong, 1999):

$$\mathbf{\Phi_M} = \begin{bmatrix} \mathbf{\Phi} & \mathbf{Y} \\ \mathbf{Y}^T & \mathbf{\Gamma} \end{bmatrix},$$

where $\mathbf{\Phi}$ is a *K x K* correlation matrix for the major latent traits, $\mathbf{\Gamma}$ is a *k x k* correlation matrix for the minor latent traits, and $\mathbf{Y}$ is a *K x k* correlation matrix among the major and minor latent traits. Unique factor scores had a multivariate normal distribution with a mean vector of zeros and a correlation matrix of identity. The diagonal elements of the unique factor loading matrix ($\mathbf{E}$) were equal to the diagonal of $\mathbf{I} - \mathbf{\Lambda_M \Lambda_M}^T$.

Data Simulation. The factor loading matrices for the major latent traits ($\mathbf{\Lambda}$) were assumed to have a complex structure, and the factor loadings for a latent trait were assumed to have a uniform distribution between *a* and *b*. This assumption helps to generate reasonable factor loadings for a major latent trait by controlling the dimensional strength. The expected value for a square of a random variable from a uniform distribution with the boundaries *a* and *b* is equal to

$$\mathrm{E}[X^2] = \frac{a^2 + ab + b^2}{3} \quad . \tag{56}$$

If a set of variables is randomly drawn from a uniform distribution with the boundaries *a* and *b*, the above equation provides the expected average of the squared values. This fact can be used to control the dimensional strength for a latent dimension when generating its factor loadings. For instance, if a set of factor loadings for 20 items was generated from a uniform distribution with the boundaries 0.3 and 0.8, then the variance accounted for by the set of factor loadings would be approximately 32.3%. Table 11 indicates the boundaries of uniform distributions used in the current study to generate factor loadings for a major latent trait given the variance accounted for by the major latent trait.

First, given the number of major latent traits and variance accounted for by the major latent traits in a simulation condition, factor loading vectors for each latent trait were generated based on the uniform distributions shown in Table 11, and the factor loading vectors were combined in a matrix to construct an *n x K* factor loading matrix of major latent traits (**Λ**). Then, an *n x 50* factor loading matrix for minor latent traits (**λ**) was constructed given the variance accounted for by minor latent traits in the simulation condition using a similar procedure described in Hong (1999). The factor loading matrix for the minor latent traits (**λ**) was generated by using multivariate random normal deviates. The standard deviation of the first minor latent trait was equal to 1, and the standard deviation of each successive minor latent trait was .9 times the standard deviation of the preceding minor factor. After generating random factor loadings for the minor latent traits, the rows of the factor loading matrix were rescaled such that the desired level of contribution by minor factors was satisfied. Finally, two factor loading

Table 10. *The Amount of Variance Accounted for by the Major Latent Traits in 40 Different Factor Structures Used in the Simulation Study*

	One-Dimensional	Two-Dimensional		Three-Dimensional			Four-Dimensional			
	Dim1	Dim1	Dim2	Dim1	Dim2	Dim3	Dim1	Dim2	Dim3	Dim4
1	20	20	5	20	5	5	20	10	5	5
2	30	30	5	30	5	5	30	10	5	5
3	40	40	5	40	5	5	40	10	5	5
4		20	10	20	10	5	20	20	5	5
5		30	10	30	10	5	20	10	10	5
6		40	10	40	10	5	30	10	10	5
7		20	20	20	20	5	40	10	10	5
8		30	20	30	20	5	20	20	10	5
9		40	20	20	10	10	30	20	10	5
10		30	30	30	10	10	20	20	20	5
11				40	10	10	20	10	10	10
12				20	20	10	30	10	10	10
13				30	20	10	20	20	10	10
14				20	20	20				

matrices were combined to construct the mega factor loading matrix, $\mathbf{\Lambda_M}$. After creating the $\mathbf{\Lambda_M}$ matrix, the communalities were checked for each row, and small adjustments were made if the communality for any item was larger than 1. The diagonal elements of the unique factor loading matrix (**E**) were obtained by the diagonal elements of the matrix ($\mathbf{I - \Lambda_M \Lambda_M}^{\boldsymbol{T}}$).

An *Nx(K+50)* mega factor score matrix for the major and minor latent traits ($\mathbf{F_M}$) were generated from a multivariate normal distribution with a mean vector of zeros and mega factor correlation matrix $\mathbf{\Phi_M}$, where *N* was either 500 or 1000 as determined by the simulation conditions. The factor correlation matrix ($\mathbf{\Phi}$) for the major latent traits was a *K x K* matrix with all off-diagonal elements equal to *r*, where *r* was either 0 or .5 as determined by the simulation condition. The rest of the elements in the mega factor correlation matrix, a *50 x 50* correlation matrix for minor latent traits ($\mathbf{\Gamma}$) and a *K x 50* correlation matrix among the major and minor latent traits (**Y**), were identity matrices. An *N x n* unique factor score matrix (**U**) was generated from a multivariate normal distribution with a mean vector of zeros and a correlation matrix of identity. After

Table 11. *The Boundaries of Uniform Distributions to Generate Factor Loadings for a Major Latent Trait Given the Variance Accounted for by the Major Latent Trait*

Lower Boundary	Upper Boundary	Expected Average of Squared Values (Variance Accounted)
0.14	0.30	0.05 (5%)
0.20	0.42	0.10 (10%)
0.20	0.65	0.20 (20%)
0.32	0.75	0.30 (30%)
0.32	0.90	0.40 (40%)

generating each input matrix as described above, the latent continuous item scores were generated using Equation 57, and the latent continuous item scores were dichotomized using the threshold values generated from a uniform distribution between -2 and 2.

Analysis. As described in the first study**,** different dimensionality assessment criteria were implemented for each of the 500 simulated datasets in 640 conditions, and the

number of latent traits suggested by these criteria was recorded for studying their performance. For parallel analysis, revised parallel analysis, and the criteria based on the FIML estimation, only 100 simulated datasets were analyzed due to time constraints while all 500 simulated datasets were analyzed for other criterion procedures. Another difference in Study 2 for the Mplus FIML estimation was the smaller quadrature points used in the analysis. Seven quadrature points when fitting the one-dimensional models, five quadrature points when fitting the two- and three-dimensional models, four quadrature points when fitting the four-dimensional models, and three quadrature points when fitting the five- and six-dimensional models were used to approximate the integration for the FIML estimation in Study 2. After the analysis was completed, the proportion of replications with the correctly identified quasi-true number of dimensions was computed within each condition. In addition, the bias with respect to the number of major dimensions (quasi-true number of latent dimensions) and root mean squared deviation from the quasi-true number of dimensions were computed within each condition based on the following equations:

$$O_{BIAS} = \frac{\sum_{k=1}^{R} (O_s^{\,k} - O)}{R} \text{ and} \tag{57}$$

$$O_{RMSD} = \sqrt{\frac{\sum_{k=1}^{R} (O_s^{\,k} - O)^2}{R}}, \tag{58}$$

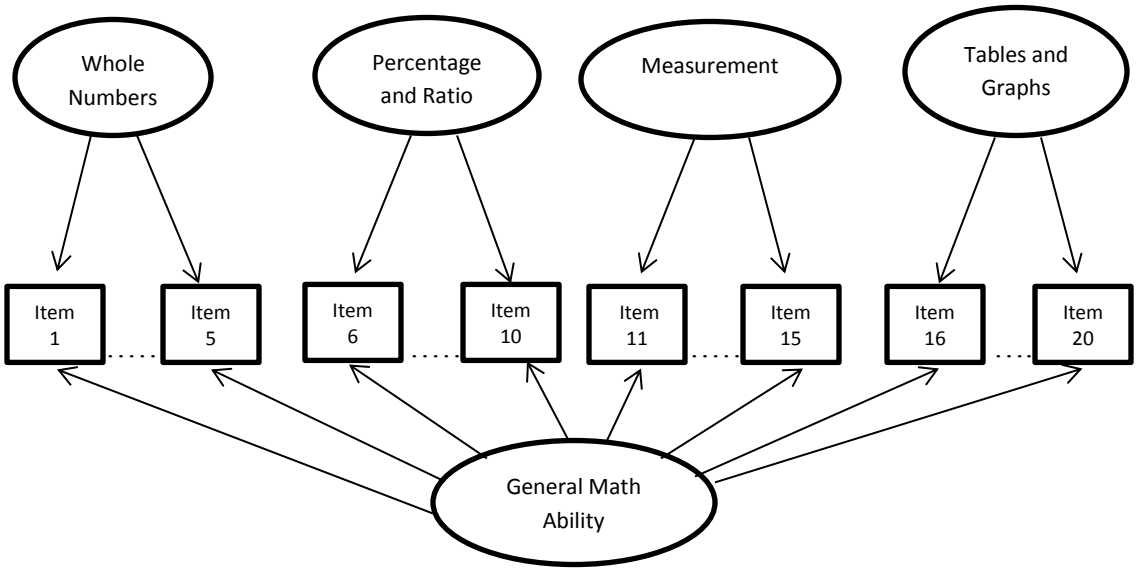
where R was the number of analyzed simulated datasets, $O_s^{\,k}$ was the recommended number of latent traits by a criterion for the kth simulated data, and O was the quasi-true number of dimensions in the true generating model.

CHAPTER 4: RESULTS

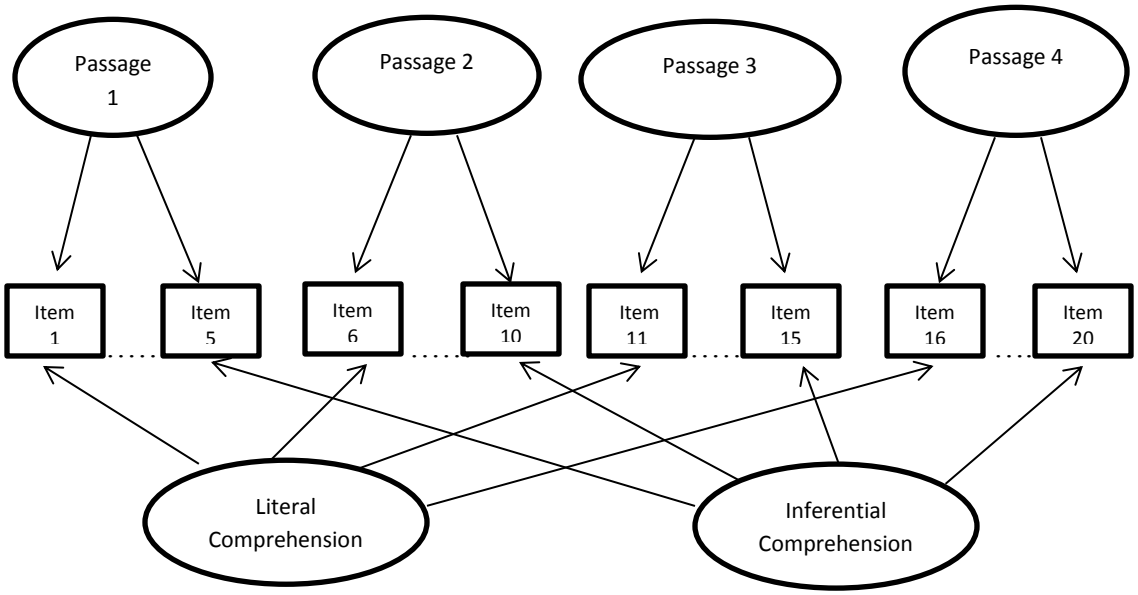
Study 1

In the first phase of Study 1, four datasets with more than 67,000 observations were treated as "population" in the current study and analyzed using parallel analysis, revised parallel analysis, DETECT, NOHARM, WLSM, WLSMV, and MLR estimators in Mplus, and the number of dimensions suggested by several of these criteria was identified at the population level. Then, in the second phase of Study 1, each of the 500 samples drawn from the population datasets was analyzed and the suggested number of dimensions by each procedure/criterion was determined. The outcome variables of interest were the average, standard deviation, and the range of the suggested number of dimensions by each procedure/criterion across 500 replications. In addition, the results of the analysis for each method at the population level are provided.

Since the true factor structures underlying real datasets were not known, it was not easy to interpret the results of Study 1 without making any reference to the true structure. However, the expected structures for these real datasets are presented to help interpretation of the results. In Figure 3, the factor structures for the major latent dimensions expected for the 20-item reading and mathematics tests are depicted based on the information provided in Table 8. In the 20-item mathematics test, four dimensions can be hypothesized due to the content area knowledge in addition to a general ability for mathematics as a typical bifactor model (Gibbons & Hedeker, 1992; Reise, Morizot, & Hays, 2007). Therefore, five major latent dimensions are expected for the 20-item and 40-item mathematics tests, as the structure of the 40-item mathematics test is very similar to the 20-item mathematics test. In the 20-item reading test, six major dimensions can be expected as a result of four dimensions due to the different passages' content and two latent dimensions related to the two main target abilities (literal comprehension and inferential comprehension). In the 40-item reading test, a similar structure with seven major latent dimensions can be expected due to an additional fifth reading passage in the test. Also, a number of minor factors to unknown degrees are expected to be present for all four population datasets.



(a) 20-item Mathematics Test



(a) 20-item Reading Test

Figure 3. Expected Major Latent Dimensions for the 20-item Mathematics and Reading Tests

Full Data Analysis

Parallel Analysis. The first five eigenvalues from four datasets and the 95th percentile of the random data eigenvalue at the corresponding rank position are given in Table 12. Given these values, parallel analysis indicated one dimension for the 20-item mathematics test, two dimensions for the 40-item mathematics test, one dimension for the 20-item reading test, and three dimensions for the 40-item reading test at the population level.

Table 12. *Parallel Analysis Results for the Population Datasets*

	Mathematics		Reading	
	20 item	40 item	20 item	40 item
Eigenvalue				
1	**8.65 (1.07)**	**15.63 (1.13)**	**7.29 (1.08)**	**16.16 (1.20)**
2	0.99 (1.06)	**1.81 (1.11)**	1.03 (1.06)	**1.36 (1.16)**
3	0.85 (1.05)	1.03 (1.10)	0.90 (1.05)	**1.14 (1.13)**
4	0.76 (1.04)	0.98 (1.09)	0.84 (1.05)	0.99 (1.12)
5	0.75 (1.03)	0.94 (1.08)	0.78 (1.04)	0.93 (1.10)

Note. Numbers in parentheses are the 95th percentile of the associated random data eigenvalue distribution based on 100 random datasets.

Revised Parallel Analysis. Revised parallel analysis requires a separate simulation to test the eigenvalue at each rank position. For demonstration purposes, the results of the revised parallel analysis for the 20-item mathematics test are fully given in Table 13. The first round of the revised parallel analysis is an identical procedure to testing the first eigenvalue using parallel analysis. The first eigenvalue was found significant in Round 1 as the sample eigenvalue estimate was larger than the 95th percentile of the random data eigenvalue distribution. Then, one factor was extracted from the dataset, and the associated one-dimensional factor loading matrix was used to generate 100 datasets in Round 2. As seen in Table 13, the average value for the first eigenvalue from the simulated datasets closely matched the first sample eigenvalue estimate. Therefore, the 95th percentile of the second eigenvalue from the simulated datasets in Round 2 was a conditional estimate given the magnitude of the significantly found first eigenvalue. The second sample eigenvalue was also found to be significant. The next step in Round 3 was

to extract two factors from the dataset and to use the estimated two-dimensional factor loading matrix in generating two-dimensional datasets. As seen in the table, the average values for the first and second eigenvalues across 100 simulated datasets closely matched with the real data eigenvalue estimates. The 95th percentile of the third eigenvalue from the simulated datasets in Round 3 was similarly a conditional estimate given the magnitude of the significantly found first and second eigenvalues. As a result, the third eigenvalue was also found to be significant. Finally, three factors were extracted from the dataset and the

Table 13. *Revised Parallel Analysis for the 20-item Mathematics Test*

	Eigenvalue	Sample Estimate	Simulated Data	
			Mean	95 %
Round 1	1	**8.645**	1.057	**1.066**
Round 2	1	8.645	8.658	8.717
	2	**0.988**	0.766	**0.773**
Round 3	1	8.645	8.662	8.724
	2	0.988	0.989	1.004
	3	**0.849**	0.765	**0.772**
Round 4	1	8.645	8.660	8.714
	2	0.988	0.988	1.005
	3	0.849	0.850	0.862
	4	**0.761**	0.756	**0.764**

Note. A hundred datasets were simulated for each round.

estimated three-dimensional factor loading matrix was used to simulate data in Round 4. The fourth sample eigenvalue estimate was not larger than the 95th percentile of the fourth eigenvalue conditioning on the magnitude of the previously found first three significant eigenvalues. Revised parallel analysis suggested three dimensions for the 20-item mathematics test.

The results for the four subtests in the current study are shown in Table 14. Revised parallel analysis suggested 13 dimensions for the 40-item mathematics test, six dimensions for the 20-item reading test, and seven dimensions for the 40-item

mathematics test. The results were substantially different than what regular parallel analysis suggested.

Table 14. *Revised Parallel Analysis Results for the Population Datasets*

	Mathematics		Reading	
	20 item	40 item	20 item	40 item
Eigenvalue				
1	8.65 (1.07)	15.63 (1.13)	7.29 (1.08)	16.16 (1.19)
2	0.99 (0.77)	1.81 (0.93)	1.06 (0.85)	1.36 (0.86)
3	**0.85 (0.77)**	1.03 (0.91)	0.90 (0.81)	1.14 (0.84)
4	0.76 (0.76)	0.98 (0.91)	0.84 (0.79)	0.99 (0.82)
5		0.94 (0.91)	0.78 (0.78)	0.93 (0.81)
6		0.88 (0.86)	**0.78 (0.77)**	0.87 (0.80)
7		0.82 (0.77)	0.75 (0.76)	**0.80 (0.79)**
8		0.80 (0.76)		0.78 (0.78)
9		0.76 (0.75)		
10		0.75 (0.74)		
11		0.74 (0.72)		
12		0.72 (0.71)		
13		**0.71 (0.70)**		
14		0.69 (0.70)		

Note. Numbers in parentheses are the 95th percentile of the eigenvalue distribution from the simulated datasets conditioning on the magnitude of the previously found significant eigenvalues.

DETECT. The results for the DETECT analysis are given in Table 15. DETECT analysis suggested that the DETECT value was maximized for five latent dimensions (clusters) for all population datasets used in the current study. However, if we consider the classifications for the DETECT values recommended by Kim (1996), the maximized DETECT values were very low and an indication of unidimensionality. Based on the recommendations by Stout et al. (1996) and Roussos and Ozbek (2006), the maximized DETECT values indicated weak or very weak multidimensionality present in all four datasets.

Table 15. *DETECT Results for the Population Datasets*

	Mathematics		Reading	
	20 item	40 item	20 item	40 item
Maximized DETECT Value	0.184	0.124	0.158	0.078
Number of Latent Traits (Clusters)	5	5	5	5

Chi-Square Statistics. The current study included several chi-square statistics proposed in the literature. These chi-square statistics are based on the NOHARM ULS, Mplus WLS, and Mplus FIML estimations. For the Mplus WLS and NOHARM ULS estimations, models up to 14 dimensions were fitted to the 20-item tests and models up to 27 dimensions were fitted to the 40-item tests, and proposed chi-square statistics were computed for each solution. For the FIML estimation with the Mplus MLR estimator, models up to eight dimensions were fitted to all original population datasets. An important limitation in the FIML estimation was the number of quadrature points used in the analysis to approximate the multiple integrals at the E-step. Seven quadrature points for the one-, two-, three, and four-dimensional models, five quadrature points for the five-dimensional model, four quadrature points for the six-dimensional model, three quadrature points for the seven dimensional model, and two quadrature points for the eight-dimensional models are used for approximation at the E-step when analyzing the population datasets. Computational sources were not available to fit eight-dimensional models with three or more quadrature points as it required at least 20 GB RAM[14] for analyzing a dataset with more than 67,000 observations.

The results from these analyses for the original population datasets are reported in Table 16, Table 17, Table 18, and Table 19. For the 20-item mathematics test, NOHARM ALR and Achi selected a nine-dimensional solution, while NOHARM mean-adjusted (T_{MChi}) and mean-and-variance adjusted (T_{MVChi}) chi-square statistics selected a 10-dimensional solution and Mplus mean-adjusted (WLS_{MChi}) and mean-and-variance

[14] Regular desktop computers have 4 to 8 GB RAM. The super computers available at the Minnesota Super Computer Institute were used for the analysis, but even the Windows-based machines did not allow more than 16GB usage. Mplus does not work under Linux systems, so better computer resources could not be used.

adjusted (WLS_{MVChi}) chi-square statistics chose a 11-dimensional solution. For the 20-item reading test, NOHARM-based chi-square statistics were all significant even for the 14-dimensional model, which was the largest model fitted. Mplus-based chi-square statistics indicated that a 10-dimensional solution fitted the data. For the 40-item mathematics test, NOHARM-based chi-square statistics were all significant even for the 27-dimensional model, which was the largest model fitted, except NOHARM ALR, which selected the 22-dimensional model. Both Mplus-based chi-square statistics selected the 25-dimensional model. For the 40-item reading test, all NOHARM-based chi-square statistics were significant even for the 27-dimensional model, and both Mplus-based chi-square statistics selected the 26-dimensional model.

The results from the FIML estimation were not conclusive for the population datasets. The p-values associated with the unadjusted and adjusted chi-square differences between the one-dimensional model and two-dimensional model, two-dimensional model and three-dimensional model, three-dimensional model and four-dimensional model, four-dimensional model and five-dimensional model, five-dimensional model and six-dimensional model, and six-dimensional model and seven-dimensional model were all significant at the alpha level of 0.001 for all population datasets. The only exception was the 20-item mathematics test in which the comparison between the six-dimensional model and seven-dimensional model was not possible due to the non-convergence of the seven-dimensional solution.

Comparisons between the seven-dimensional model and eight-dimensional model favored the seven-dimensional model in all population datasets, but were not meaningful, because negative twice log-likelihood (-2LL) statistics increased unexpectedly from the seven-dimensional model to the eight-dimensional model. This may be due to using only two quadrature points, which make the approximated log-likelihood statistics very unreliable for eight-dimensional models. At most, the conclusion for the FIML analysis was that the adjusted and unadjusted chi-square difference tests indicated at least six dimensions for the 20-item mathematics test and seven dimensions for other tests, and reliable information beyond that point was not available due to the lack of the large amount of computational resources required.

Table 16. *Chi-Square Fit Statistics from NOHARM ULS and MPLUS WLS estimations for the 20-item Mathematics and 20-item Reading tests*

Number of Dimensions	df	Noharm Achi	Noharm ALR	Noharm T_{MChi}	Noharm T_{MVChi}[a,b]	Mplus WLS_{MChi}	Mplus WLS_{MVChi}
20-item Mathematics Test							
1	170	5348.6***	5053.1***	7256.2***	6567.7***	7589.9***	7424.3***
2	151	1532.4***	1327.8***	2010.0***	1844.7***	2035.6***	2010.0***
3	133	674.3***	594.1***	948.9***	879.7***	929.4***	920.9***
4	116	489.2***	431.5***	694.8***	646.2***	664.0***	658.1***
5	100	316.7***	276.6***	445.6***	414.7***	434.4***	431.0***
6	85	197.8***	176.9***	295.6***	276.1***	289.5***	287.7***
7	71	127.9***	115.0***	181.1***	169.4***	182.6***	181.6***
8	58	90.3***	80.1*	123.7***	116.6***	127.9***	127.4***
9	**46**	**61.6**	**52.7**	77.3***	73.3***	85.1***	84.9***
10	35			**46.9**	**44.5**	53.4*	53.3*
11	25					**36.5**	**36.5**
20-item Reading Test							
1	170	3530.4***	2799.5***	4278.8***	3774.2***	3899.2***	3847.6***
2	151	1642.8***	1199.6***	1702.5***	1539.4***	1754.6***	1739.0***
3	133	921.8***	667.4***	934.8***	848.0***	966.0***	959.3***
4	116	635.6***	468.2***	675.0***	617.2***	649.2***	645.7***
5	100	482.2***	348.3***	497.1***	458.0***	453.7***	451.7***
6	85	392.4***	286.5***	408.5***	374.2***	291.2***	290.1***
7	71	272.3***	201.3***	284.0***	263.1***	200.4***	199.8***
8	58	184.0***	129.1***	167.3***	155.5***	113.5***	113.3***
9	46	184.3***	140.0***	193.1***	182.0***	78.5***	78.4***
10	35	109.9***	83.5***	108.5***	100.9***	**35.6**	**35.6**
11	25	67.4***	54.4***	70.0***	65.8***		
12	16	43.9***	34.9***	46.6***	44.8***		
13	8	59.1***	48.6***	67.1***	65.9***		
14	1	30.3***	25.8***	16.9***	33.7***		

Note. Noharm-Achi: Noharm based approximate chi-square statistics, Noharm-ALR: Noharm based approximate likelihood ratio chi-square statistics, Noharm – T_{MChi}: Noharm based mean-adjusted chi-square statistics, Noharm – T_{MVChi}: Noharm based mean-and-variance adjusted chi-square statistics, Mplus – WLS_{MChi}: Mplus weighted least squares mean-adjusted chi-square statistics, Mplus – WLS_{MVChi}: Mplus weighted least squares mean-and-variance adjusted chi-square statistics

[a] The adjusted degrees of freedom for the T_{MVChi} statistics are 153.9, 138.6, 123.3, 107.9, 93.1, 79.4, 66.4, 54.7, 43.6, and 33.2 for 20-item mathematics test for the 1-10 dimensional models.

[b] The adjusted degrees of freedom for the T_{MVChi} statistics are 150.0, 136.6, 120.7, 106.1, 92.1, 77.9, 65.8, 53.9, 43.4, and 32.6 for 20-item reading test for the 1-10 dimensional models.

* $p < .05$. ** $p < .01$. *** $p < .001$.

Table 17. *Chi-square Fit Statistics from NOHARM ULS and MPLUS WLS estimations for the 40-item Mathematics Test*

Number of Dimensions	df	Noharm Achi	Noharm ALR	Noharm T_{MChi}	Noharm T_{MVChi}[a]	Mplus WLS_{MChi}	Mplus WLS_{MVChi}
1	740	29802***	25469***	35903***	27236***	35185***	33145***
2	701	8946.0***	7410.5***	12819***	10177***	10714***	10401***
3	663	6332.2***	4653.0***	7100.6***	5712.0***	6717.9***	6547.9***
4	626	4535.5***	3245.1***	4820.2***	3905.1***	4781.2***	4676.7***
5	590	3145.2***	2257.4***	3328.6***	2706.2***	3354.4***	3289.8***
6	555	2133.4***	1616.3***	2460.9***	2004.7***	2428.3***	2387.6***
7	521	1752.4***	1317.2***	2025.3***	1649.4***	2044.8***	2013.4***
8	488	1432.7***	1092.2***	1689.9***	1377.4***	1674.6***	1651.5***
9	456	1269.2***	963.6***	1455.4***	1192.2***	1503.9***	1484.5***
10	425	1121.3***	842.7***	1274.3***	1045.6***	1286.1***	1270.9***
11	395	997.0***	736.0***	1083.3***	897.9***	1124.9***	1112.5***
12	366	855.8***	644.6***	909.4***	755.6***	957.6***	947.9***
13	338	753.0***	559.6***	785.7***	657.0***	834.3***	826.6***
14	311	644.1***	475.0***	662.7***	554.0***	721.6***	715.5***
15	285	619.9***	448.5***	618.2***	518.8***	625.5***	620.8***
16	260	509.6***	366.0***	529.0***	446.1***	549.9***	546.4***
17	236	441.3***	335.6***	458.1***	385.0***	472.9***	470.3***
18	213	383.8***	280.9**	372.5***	316.2***	413.7***	411.6***
19	191	357.5***	254.0**	340.1***	294.3***	362.2***	360.7***
20	170	302.5***	222.6**	294.1***	257.0***	295.6***	294.6***
21	150	265.7***	194.4**	262.6***	225.5***	226.9***	226.3***
22	131	201.4***	**153.5**	192.2***	164.7***	187.2**	186.8**
23	113	185.3***		167.1***	145.1***	154.0**	153.8**
24	96	165.8***		142.3***	122.1***	135.6**	135.4**
25	80	138.0***		124.5***	107.8***	**87.1**	**87.1**
26	65	106.3***		90.3*	79.8*		
27	51	90.2***		77.1*	70.7*		

Note. Noharm-Achi: Noharm based approximate chi-square statistics, Noharm-ALR: Noharm based approximate likelihood ratio chi-square statistics, Noharm – T_{MChi}: Noharm based mean-adjusted chi-square statistics, Noharm – T_{MVChi}: Noharm based mean-and-variance adjusted chi-square statistics, Mplus – WLS_{MChi}: Mplus weighted least squares mean-adjusted chi-square statistics, Mplus – WLS_{MVChi}: Mplus weighted least squares mean-and-variance adjusted chi-square statistics

[a] The adjusted degrees of freedom for the T_{MVChi} statistics are 561.4, 556.5, 533.3, 507.2, 479.7, 452.1, 424.3, 397.8, 373.5, 348.8, 327.4, 304.1, 282.6, 260.0, 239.1, 219.3, 198.3, 180.8, 165.3, 148.5, 128.8, 112.2, 98.1, 82.4, 69.2, 57.4, and 46.7 for the 1-27 dimensional solutions, respectively.

* $p < .05$. ** $p < .01$. *** $p < .001$.

Table 18. *Chi-square Fit Statistics from NOHARM ULS and MPLUS WLS estimations for the 40-item Reading Test*

Number of Dimensions	df	Noharm Achi	Noharm ALR	Noharm T_{MChi}	Noharm T_{MVChi}[a]	Mplus WLS_{MChi}	Mplus WLS_{MVChi}
1	740	16162***	9614.9***	16162***	10972***	13417***	12764***
2	701	10933***	5897.1***	8522.7***	5962.3***	8331.9***	7995.6***
3	663	8050.7***	4140.8***	5561.6***	3946.0***	5643.5***	5461.8***
4	626	5925.3***	3099.9***	4349.4***	3091.3***	4333.7***	4213.0***
5	590	3937.2***	2188.9***	3197.5***	2296.1***	2956.5***	2891.5***
6	555	3517.9***	1809.0***	2458.6***	1772.8***	2325.3***	2278.7***
7	521	2736.2***	1422.1***	1948.9***	1406.6***	1939.1***	1903.5***
8	488	2189.5***	1204.9***	1703.9***	1238.9***	1562.2***	1537.0***
9	456	1786.3***	976.2***	1347.9***	987.2***	1295.1***	1276.7***
10	425	1697.2***	922.8***	1276.5***	937.4***	1093.4***	1079.5***
11	395	1746.1***	895.9***	1142.3***	840.8***	934.2***	923.7***
12	366	1497.5***	804.5***	1067.8***	793.2***	809.4***	801.2***
13	338	1427.4***	753.5***	1010.2***	752.0***	687.8***	681.5v***
14	311	1523.0***	767.4***	910.2***	691.5***	617.6***	612.4***
15	285	1171.2***	584.7***	752.3***	569.6***	537.9***	533.7***
16	260	1137.4***	562.2***	712.9***	551.6***	454.8***	452.0***
17	236	1086.1***	542.0***	658.1***	509.5***	411.6***	409.3***
18	213	951.8***	475.4***	581.0***	458.8***	332.1***	330.5***
19	191	841.4***	401.4***	526.1***	423.9***	276.7***	275.6***
20	170	754.4***	381.7***	458.1***	365.3***	214.8*	214.2*
21	150	777.2***	369.7***	423.5***	349.5***	-	-
22	131	943.1***	395.5***	466.0***	383.5***	197.4***	196.7***
23	113	645.2***	294.2***	362.8***	302.3***	167.2**	166.7**
24	96	687.5***	298.6***	358.1***	300.9***	141.9**	141.5**
25	80	654.0***	293.1***	338.0***	293.9***	115.0**	114.8**
26	65	686.8***	297.1***	356.3***	316.6***	**79.3**	**79.2**
27	51	526.7***	253.9***	298.0***	272.8***		

Note. Noharm-Achi: Noharm based approximate chi-square statistics, Noharm-ALR: Noharm based approximate likelihood ratio chi-square statistics, Noharm – T_{MChi}: Noharm based mean-adjusted chi-square statistics, Noharm – T_{MVChi}: Noharm based mean-and-variance adjusted chi-square statistics, Mplus – WLS_{MChi}: Mplus weighted least squares mean-adjusted chi-square statistics, Mplus – WLS_{MVChi}: Mplus weighted least squares mean-and-variance adjusted chi-square statistics. Dashes indicate no convergence.

[a] The adjusted degrees of freedom for the T_{MVChi} statistics are 502.3, 490.4, 470.4, 444.9, 423.7, 400.2, 376.0, 354.8, 334.0, 312.1, 290.7, 271.9, 251.6, 236.3, 215.8, 201.2, 182.7, 168.2, 153.9, 135.6, 123.8, 107.8, 94.2, 80.7, 69.6, 57.8, and 46.7 for the 1-27 dimensional solutions, respectively.

* $p < .05$. ** $p < .01$. *** $p < .001$.

Table 19. *Full Information Maximum Likelihood Estimation Statistics from Mplus MLR Estimator*

		20-item			
		Mathematics		Reading	
Number of Dimensions	Number of Estimated Parameters	-2 LL	Scaling Factor	-2 LL	Scaling Factor
1	40	1248873	1.02	1117669	1.01
2	59	1244863	1.02	1115864	1.04
3	77	1244167	1.03	1115491	1.03
4	94	1243929	1.09	1115538	0.00
5	110	1243723	0.00	1115395	0.00
6	125	1243147	0.00	1114839	1.11
7	139	-	-	1114665	1.28
8	152	1245034	1.02	1116911	1.07

		40-item			
		Mathematics		Reading	
Number of Dimensions	Number of Estimated Parameters	-2LL	Scaling Factor	-2 LL	Scaling Factor
1	80	2100360	1.04	1650774	1.03
2	119	2089614	1.04	1647049	1.05
3	157	2086110	1.04	1645451	1.08
4	194	2085304	1.05	1644285	1.06
5	230	2083520	0.00	1643751	1.13
6	265	2082486	1.03	1643344	1.11
7	299	2081802	1.09	1643088	1.37
8	332	2099468	-	1644320	1.17

Note. Seven quadrature points for the one-, two-, three, and four-dimensional models, five quadrature points for the five-dimensional model, four quadrature points for the six-dimensional model, three quadrature points for the seven-dimensional models, and two quadrature points for the eight-dimensional models are used for approximation at the E-step.

The large number of dimensions suggested by the chi-square statistics is not surprising. As discussed in Chapter 2, the goal of the chi-square statistics is to minimize the discrepancy between the fitted model and true model by assuming that there is no estimation error. When there is sufficient power, the chi-square statistics will attempt to identify the dimensions in the "true" model regardless of their size, and they will tend to include minor factors. When analyzing population datasets, there is no doubt that we have sufficient power with more than 67,000 observations, and it is expected that the chi-square statistics will select very complex models with many latent dimensions.

Other Model Fit Indices. The RMSEA, CFI, and SRMR indices obtained from the Mplus WLS estimation are reported in Table 20. Considering the standard cut-off value of 0.05 for the RMSEA index, 0.95 for the CFI index, and 0.07 for the SRMR index recommended in the literature, the one-dimensional model was selected for all four population datasets.

The results for the fit indices AIC, AIC_c, and BIC obtained from Mplus FIML estimation with the MLR estimator are reported in Table 21. The results for the AIC and AIC_c were identical due to the large sample size. The results were not conclusive in most occasions. The values from the eight-dimensional solution were not included in the table due to the unreliable -2LL statistics as discussed before. For the 20-item mathematics test, a solution for the seven-dimensional model was not available, and the minimum AIC, AIC_c, and BIC values occurred for the six-dimensional solution. For the 20-item reading and 40-item mathematics tests, minimum AIC, AIC_c, and BIC values occurred for the seven-dimensional solution. For these datasets, there was no decision for the number of dimensions because the information beyond was not available. The exception was the 40-item reading test. Minimum AIC and AIC_c values occurred in the seven-dimensional solution, so there was no decision again. However, minimum BIC values occurred in the six-dimensional solution. Therefore, BIC favored the six-dimensional solution for the 40-item reading test.

Table 20. *Fit Indices from Mplus WLS Estimation for the Population Datasets*

	20-item					
	Mathematics			Reading		
Number of Dimensions	RMSEA	CFI	SRMR	RMSEA	CFI	SRMR
1	**0.025**	**0.991**	**0.029**	**0.018**	**0.990**	**0.026**
2	0.013	0.998	0.016	0.012	0.996	0.018
3	0.009	0.999	0.011	0.009	0.998	0.014
4	0.008	0.999	0.009	0.008	0.999	0.011
5	0.006	1.000	0.007	0.007	0.999	0.009
6	0.005	1.000	0.006	0.005	0.999	0.007
7	0.004	1.000	0.005	0.004	1.000	0.006
8	0.003	1.000	0.004	0.003	1.000	0.005
	40-item					
	Mathematics			Reading		
Number of Dimensions	RMSEA	CFI	SRMR	RMSEA	CFI	SRMR
1	**0.026**	**0.984**	**0.039**	**0.015**	**0.981**	**0.030**
2	0.014	0.995	0.021	0.012	0.988	0.024
3	0.011	0.997	0.018	0.010	0.992	0.020
4	0.010	0.998	0.015	0.009	0.994	0.017
5	0.008	0.999	0.013	0.007	0.996	0.014
6	0.007	0.999	0.011	0.007	0.997	0.013
7	0.006	0.999	0.010	0.006	0.998	0.012
8	0.006	0.999	0.009	0.005	0.998	0.010

Note. The values reported in the cells are based on the Mplus WLSM estimator. The fit indices from the Mplus WLSMV estimator were all identical to the reported cell values at the second decimal. RMSEA values are the lower bounds from the associated 95% confidence interval.

Table 21. *Fit Indices from Mplus Full Information Maximum Likelihood Estimation for the Population Datasets*

	20-item					
	Mathematics			Reading		
Number of Dimensions	AIC	BIC	AICc	AIC	BIC	AICc
1	1248953	1249318	1248953	1117749	1118113	1117749
2	1244981	1245518	1244981	1115982	1116519	1115982
3	1244321	1245022	1244321	1115645	1116347	1115646
4	1244117	1244973	1244117	1115726	1116583	1115726
5	1243943	1244945	1243943	1115615	1116617	1115615
6	**1243397**	**1244536**	**1243398**	1115089	1116227	1115089
7	-	-	-	**1114943**	**1116209**	**1114943**
	40-item					
	Mathematics			Reading		
Number of Dimensions	AIC	BIC	AICc	AIC	BIC	AICc
1	2100520	2101248	2100520	1650934	1651662	1650934
2	2089852	2090936	2089852	1647287	1648370	1647287
3	2086424	2087854	2086424	1645765	1647194	1645765
4	2085692	2087459	2085692	1644673	1646438	1644673
5	2083979	2086073	2083979	1644211	1646304	1644211
6	2083017	2085429	2083017	1643874	**1646286**	1643874
7	**2082400**	**2085122**	**2082400**	**1643686**	1646407	**1643686**

Note. Dashes indicate no converged solution. AIC and AIC_c values are identical due to the large sample size.

Sampling Analysis

Recall that 500 samples with the sample sizes of 500 and 1000 were drawn from each population dataset. Then, each sample dataset was analyzed using parallel analysis, revised parallel analysis, DETECT, NOHARM, and WLSM, WLSMV, and MLR estimators in Mplus, and the number of dimensions suggested by different criteria was recorded. For some criteria, a decision for the number of dimensions was not reached for every replication in some conditions. There were two different occasions for no-decision replications. The first type of no-decision replications occurred as a result of the study design. As stated before, when a sample dataset is analyzed by NOHARM or the WLSM and WLSMV estimators in Mplus, models up to eight dimensions were fitted to each sample data. When a sample dataset is analyzed by the MLR estimator in Mplus, models up to six dimensions were fitted. In some replications, for instance, Mplus-based mean-

adjusted chi-square statistics were still significant for the eight-dimensional model, or MLR-based AIC value was the minimum for the six-dimensional model. Therefore, a decision was not available for some replications even after fitting the eight-dimensional or six-dimensional models. The second type of no-decision replications occurred due to convergence problems. For instance, one-, two-, and three-dimensional models were fit to a sample dataset using the MLR estimator in Mplus and solutions converged, but a solution was not available for higher dimensional models due to non-convergence; a decision was not reached by using the information from the converged solutions. The proportions of datasets in which the dimensionality decision could not be reached for different criteria are presented in Table 22 and Table 23.

The first type of no-decision replications due to the upper limit of the fitted models in the analysis occurred for NOHARM-based Achi, chi-square difference test with unadjusted log-likelihood statistics, and AIC based on the Mplus FIML estimation with the MLR estimator. These instances were generally observed in conditions such that the sample size was 1000 and the number of items was 40. The second type of no-decision replications due to convergence problems occurred for mean-adjusted and mean-and-variance adjusted chi-square statistics from Mplus WLS estimation, and criterion related to the FIML estimation with the MLR estimator in Mplus. The most severe results were observed for the chi-square difference test with unadjusted log-likelihood statistics obtained from the MLR estimator for the 40-item mathematics and reading tests. The proportion of datasets in which the decision was not reached was about 30% to 80% in those conditions.

The datasets in which the decision could not be reached due to convergence issues were eliminated from subsequent analysis. However, the datasets in which the decision could not be reached due to the limited number of fitted models were included in the subsequent analysis. For those datasets, the number of suggested dimensions was assumed to be eight for NOHARM Achi and six for Mplus MLR related criterion, which were the highest dimensional models fitted.

Table 22. *Proportion of Datasets in Which the Dimensionality Decision Cannot be Reached After Fitting Models up to a Certain Number of Latent Dimensions*

Subject	Mathematics				Reading			
Number of Items	20		40		20		40	
Sample Size	500	1000	500	1000	500	1000	500	1000
Method								
Parallel Analysis	-	-	-	-	-	-	-	-
Revised Parallel Ana.	-	-	-	-	-	-	-	-
Detect	-	-	-	-	-	-	-	-
Noharm – Achi	-	-	-	-	-	-	1%	6%
Noharm – ALR	-	-	-	-	-	-	-	-
Noharm – T_{MChi}	-	-	-	-	-	-	-	-
Noharm – T_{MVChi}	-	-	-	-	-	-	-	-
Mplus – WLS_{MChi}	-	-	-	-	-	-	-	-
Mplus – WLS_{MVChi}	-	-	-	-	-	-	-	-
Mplus – MLR_{Chi1}	-	1%	21%	21%	-	-	3%	6%
Mplus – MLR_{Chi2}	-	-	-	-	-	-	-	-
Mplus – WLS_{MRmsea}	-	-	-	-	-	-	-	-
Mplus – $WLS_{MVRmsea}$	-	-	-	-	-	-	-	-
Mplus – WLS_{SRMR}	-	-	-	-	-	-	-	-
Mplus – WLS_{MCfi}	-	-	-	-	-	-	-	-
Mplus – WLS_{MVCfi}	-	-	-	-	-	-	-	-
Mplus – MLR_{AIC}	-	1%	2%	7%	-	-	-	-
Mplus – MLR_{BIC}	-	-	-	-	-	-	-	-
Mplus – MLR_{AICc}	-	-	-	1%	-	-	-	-

Note 1. The numbers in the cells are based on 500 replications. Dashes indicate 0%. For Noharm, Mplus-WLSM and Mplus-WLSMV based indices, models up to eight dimensions were fitted. For Mplus-MLR based indices, models up to six dimensions were fitted.

Note2. Noharm-Achi: Noharm based approximate chi-square statistics, Noharm-ALR: Noharm based approximate likelihood ratio chi-square statistics, Noharm – T_{MChi}: Noharm based mean-adjusted chi-square statistics, Noharm – T_{MVChi}: Noharm based mean-and-variance adjusted chi-square statistics, Mplus – WLS_{MChi}: Mplus weighted least squares mean-adjusted chi-square statistics, Mplus – WLS_{MVChi}: Mplus weighted least squares mean-and-variance adjusted chi-square statistics, Mplus – MLR_{Chi1}: Mplus marginal maximum likelihood unadjusted chi-square difference test, Mplus – MLR_{Chi2}: Mplus marginal maximum likelihood adjusted chi-square difference test, Mplus-WLS_{MRmsea}: Mplus weighted least squares root mean square error approximation with mean-adjusted chi-square, Mplus-$WLS_{MVRmsea}$: Mplus weighted least squares root mean square error approximation with mean-and-variance adjusted chi-square, Mplus-WLS_{SRMR}: Mplus weighted least squares standardized root mean square residual, Mplus-WLS_{MCfi}: Mplus weighted least squares comparative fit index with mean-adjusted chi-square statistic, Mplus-WLS_{MVCfi}: Mplus weighted least squares comparative fit index with mean-and-variance adjusted chi-square statistic, Mplus – MLR_{AIC}: Mplus Akaike information criterion, Mplus – MLR_{BIC}: Mplus Bayesian information criterion, Mplus – MLR_{AICc}: Mplus corrected Akaike information criterion.

Table 23. *Proportion of Datasets in Which the Dimensionality Decision Cannot be Reached Due to Convergence Problems*

Subject	Mathematics				Reading			
Number of Items	20		40		20		40	
Sample Size	500	1000	500	1000	500	1000	500	1000
Method								
Parallel Analysis	-	-	-	-	-	-	-	-
Revised Parallel Ana.	-	-	-	-	-	-	-	-
Detect	-	-	-	-	-	-	-	-
Noharm – Achi	-	-	-	-	-	-	-	-
Noharm – ALR	-	-	-	-	-	-	-	-
Noharm – T_{MChi}	-	-	-	-	-	-	-	-
Noharm – T_{MVChi}	-	-	-	-	-	-	-	-
Mplus – WLS_{MChi}	1%	2%	1%	1%	1%	2%	1%	1%
Mplus – WLS_{MVChi}	-	1%	-	1%	1%	2%	-	1%
Mplus – MLR_{Chi1}	3%	9%	30%	51%	1%	3%	36%	20%
Mplus – MLR_{Chi2}	-	1%	4%	9%	-	1%	25%	20%
Mplus – WLS_{MRmsea}	-	-	-	-	-	-	-	-
Mplus – $WLS_{MVRmsea}$	-	-	-	-	-	-	-	-
Mplus – WLS_{SRMR}	-	-	-	-	1%	-	3%	-
Mplus – WLS_{MCfi}	-	-	-	-	-	-	-	-
Mplus – WLS_{MVCfi}	-	-	-	-	-	-	-	-
Mplus – MLR_{AIC}	2%	5%	7%	26%	-	1%	24%	17%
Mplus – MLR_{BIC}	-	-	3%	3%	-	-	23%	14%
Mplus – MLR_{AICc}	1%	2%	4%	5%	-	1%	23%	15%

Note 1. The numbers in the cells are based on 500 replications. Dashes indicate 0%. For Noharm, Mplus-WLSM and Mplus-WLSMV-based indices, models up to eight dimensions were fitted. For Mplus-MLR-based indices, models up to six dimensions were fitted.

Note2. Noharm-Achi: Noharm based approximate chi-square statistics, Noharm-ALR: Noharm based approximate likelihood ratio chi-square statistics, Noharm – T_{MChi}: Noharm based mean adjusted chi-square statistics, Noharm – T_{MVChi}: Noharm based mean-and-variance adjusted chi-square statistics, Mplus – WLS_{MChi}: Mplus weighted least squares mean adjusted chi-square statistics, Mplus – WLS_{MVChi}: Mplus weighted least squares mean-and-variance adjusted chi-square statistics, Mplus – MLR_{Chi1}: Mplus marginal maximum likelihood unadjusted chi-square difference test, Mplus – MLR_{Chi2}: Mplus marginal maximum likelihood adjusted chi-square difference test, Mplus-WLS_{MRmsea}: Mplus weighted least squares root mean square error approximation with mean adjusted chi-square, Mplus-$WLS_{MVRmsea}$: Mplus weighted least squares root mean square error approximation with mean-and-variance adjusted chi-square, Mplus-WLS_{SRMR}: Mplus weighted least squares standardized root mean square residual, Mplus-WLS_{MCfi}: Mplus weighted least squares comparative fit index with mean adjusted chi-square statistic, Mplus-WLS_{MVCfi}: Mplus weighted least squares comparative fit index with mean-and-variance adjusted chi-square statistic, Mplus – MLR_{AIC}: Mplus Akaike information criterion, Mplus – MLR_{BIC}: Mplus Bayesian information criterion, Mplus – MLR_{AICc}: Mplus corrected Akaike information criterion.

The average estimates for the suggested number of dimensions by each criterion across 500 replications within each of the eight conditions are reported in Table 24. Parallel analysis selected the one-dimensional model every time for seven conditions out of eight. It sometimes favored the two-dimensional solution for the 40-item mathematics test when the sample size was 1000. Although revised parallel analysis sometimes selected more complex models, one- or two-dimensional models were more frequently selected. DETECT analysis never selected the one-dimensional model for any replication and indicated models with a wide range from two to seven dimensions. On average, DETECT tended to select four dimensions in many occasions.

The model selection behavior of mean-adjusted chi-square statistics from NOHARM and Mplus were very similar to each other. Similarly, the mean-and-variance adjusted chi-square statistics from NOHARM and Mplus analyses closely matched. On average, mean-adjusted chi-square statistics tended to select more complex models than the mean-and-variance adjusted chi-square statistics from the NOHARM and Mplus analyses. The number of dimensions identified by the mean-adjusted chi-square statistics and mean-and-variance adjusted chi-square statistics given by NOHARM and Mplus analyses increased as the sample size and number of items increased.

NOHARM ALR and Achi statistics tended to select less complex models compared to the mean-adjusted and mean-and-variance adjusted chi-square statistics. For instance, NOHARM ALR tended to select the one-dimensional model almost every time for the 20-item and 40-item reading tests regardless of sample size, and selected two-dimensional models in some occasions for the 20-item and 40-item mathematics test, especially when the sample size was 1000. Similarly, NOHARM Achi tended to select one- or two-dimensional models, except the 40-item reading test in which NOHARM Achi tended to select more complex models.

The average estimate for the number of dimensions suggested by the unadjusted chi-square statistics from FIML estimation were similar to the mean-adjusted chi-square statistics from Mplus and NOHARM analyses, and tended to select more complex models.

Table 24. *Average Estimates for the Dimensionality Decisions across 500 Replications*

Subject	Mathematics				Reading			
Number of Items	20		40		20		40	
Sample Size	500	1000	500	1000	500	1000	500	1000
PA	1.00	1.00	1.00	1.24	1.00	1.00	1.00	1.00
Revised PA	1.26	1.77	1.45	2.13	1.11	1.40	1.00	1.04
Detect	3.61	3.91	3.95	4.07	3.68	3.75	3.86	3.88
Noharm – Achi	1.05	1.56	1.89	2.16	1.25	1.72	4.00	5.60
Noharm – ALR	1.02	1.31	1.28	1.98	1.04	1.06	1.02	1.10
Noharm – T_{MChi}	2.00	2.59	3.12	4.17	1.62	2.27	2.48	3.56
Noharm – T_{MVChi}	1.73	2.43	2.25	3.47	1.36	2.07	1.34	2.63
Mplus – WLS_{MChi}	2.00	2.53	3.32	4.18	1.63	2.23	3.19	3.82
Mplus – WLS_{MVChi}	1.82	2.47	2.27	3.42	1.37	2.07	1.58	2.75
Mplus – MLR_{Chi1}	2.31	2.61	4.25	4.61	2.07	2.26	1.84	1.67
Mplus – MLR_{Chi2}	1.80	2.20	2.13	2.56	1.28	1.62	1.57	1.56
Mplus – WLS_{MRmsea}	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Mplus – $WLS_{MVRmsea}$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Mplus – WLS_{SRMR}	1.03	1.00	2.89	1.02	1.78	1.00	5.53	1.53
Mplus – WLS_{MCfi}	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Mplus – WLS_{MVCfi}	1.00	1.00	1.00	1.00	1.01	1.00	1.01	1.00
Mplus – MLR_{AIC}	2.04	2.49	2.61	3.49	1.59	2.04	1.40	1.76
Mplus – MLR_{BIC}	1.00	1.00	1.00	1.02	1.00	1.00	1.00	1.00
Mplus – MLR_{AICc}	1.61	2.27	1.48	2.46	1.19	1.82	1.00	1.24

Note. PA: Parallel Analysis, RPA: Revised Parallel Analysis, Noharm-Achi: Noharm based approximate chi-square statistics, Noharm-ALR: Noharm based approximate likelihood ratio chi-square statistics, Noharm – T_{MChi}: Noharm based mean-adjusted chi-square statistics, Noharm – T_{MVChi}: Noharm based mean-and-variance adjusted chi-square statistics, Mplus – WLS_{MChi}: Mplus weighted least squares mean-adjusted chi-square statistics, Mplus – WLS_{MVChi}: Mplus weighted least squares mean-and-variance adjusted chi-square statistics, Mplus – MLR_{Chi1}: Mplus marginal maximum likelihood unadjusted chi-square difference test, Mplus – MLR_{Chi2}: Mplus marginal maximum likelihood adjusted chi-square difference test, Mplus-WLS_{MRmsea}: Mplus weighted least squares root mean square error approximation with mean-adjusted chi-square, Mplus-$WLS_{MVRmsea}$: Mplus weighted least squares root mean square error approximation with mean-and-variance adjusted chi-square, Mplus-WLS_{SRMR}: Mplus weighted least squares standardized root mean square residual, Mplus-WLS_{MCfi}: Mplus weighted least squares comparative fit index with mean adjusted chi-square statistic, Mplus-WLS_{MVCfi}: Mplus weighted least squares comparative fit index with mean-and-variance adjusted chi-square statistic, Mplus – MLR_{AIC}: Mplus Akaike information criterion, Mplus – MLR_{BIC}: Mplus Bayesian information criterion, Mplus – MLR_{AICc}: Mplus corrected Akaike information criterion.

The average estimate for the number of dimensions suggested by the adjusted chi-square statistics from FIML estimation were similar to the mean-and-variance adjusted chi-square statistics from Mplus and NOHARM analyses, and tended to select less complex models. The number of dimensions identified by the unadjusted chi-square statistics based on the MLR estimator in Mplus increased as the sample size and number of items increased. In contrast, the number of dimensions identified by the adjusted chi-square statistics seemed to be not influenced by the number of items or sample size.

The model fit indices such as RMSEA, CFI, and BIC selected one-dimensional models almost every time in all occasions. The behavior of the SRMR index in determining the number of dimensions was different than other fit indices. For instance, the number of dimensions selected by the SRMR index decreased as the sample size increased. SRMR tended to select one-dimensional models for most occasions, but selected very complex models for the 40-item mathematics and reading tests when the sample size was 500. AIC and AIC_c both tended to select more complex models compared to the other fit indices, and AIC_c selected less complex models than the AIC.

The standard deviation and range for the suggested number of dimensions by each criterion across 500 replications within each of the eight conditions are reported in Table 25 and Table 26 as a measure of model selection uncertainty or consistency. The fit indices such as RMSEA with a cut-off value of 0.05, CFI with a cut-off value of 0.95, and BIC were the most consistent criteria as they selected the same model (one-dimensional model) across all replications in all occasions. Similarly, parallel analysis selected the one-dimensional model for every replication in all occasions, except the 40-item mathematics test when the sample size was 1000. Among chi-square statistics, NOHARM ALR and Achi were more consistent than the others. In general, NOHARM and Mplus-based mean-and-variance adjusted chi-square statistics had less variability than the corresponding mean-adjusted chi-square statistics in their decisions regarding the number of dimensions.

Table 25. *Standard Deviation of Estimates for the Dimensionality Decisions Across 500 Replications*

Subject	Mathematics				Reading			
Number of Items	20		40		20		40	
Sample Size	500	1000	500	1000	500	1000	500	1000
PA	-	-	-	0.43	-	-	-	-
Revised PA	0.48	0.65	0.51	0.39	0.32	0.56	-	0.21
Detect	0.63	0.67	0.82	0.74	0.67	0.71	0.85	0.80
Noharm – Achi	0.22	0.50	0.36	0.37	0.47	0.63	1.26	1.45
Noharm – ALR	0.15	0.46	0.45	0.14	0.20	0.23	0.15	0.32
Noharm – T_{MChi}	0.66	0.74	0.97	0.99	0.73	0.79	1.04	1.09
Noharm – T_{MVChi}	0.62	0.64	0.58	0.79	0.57	0.74	0.54	0.80
Mplus – WLS_{MChi}	0.63	0.71	1.07	1.08	0.76	0.76	1.21	1.13
Mplus – WLS_{MVChi}	0.59	0.64	0.56	0.90	0.59	0.76	0.69	0.89
Mplus – MLR_{Chi1}	0.84	0.93	1.39	1.37	0.89	0.94	1.43	1.47
Mplus – MLR_{Chi2}	0.60	0.63	0.58	0.68	0.48	0.65	0.69	0.79
Mplus – WLS_{MRmsea}	-	-	-	-	-	-	-	-
Mplus – $WLS_{MVRmsea}$	-	-	-	-	-	-	-	-
Mplus – WLS_{SRMR}	0.18	-	0.71	0.13	0.59	-	1.18	0.62
Mplus – WLS_{MCfi}	-	-	-	-	-	-	0.04	-
Mplus – WLS_{MVCfi}	-	-	-	-	0.08	-	0.09	-
Mplus – MLR_{AIC}	0.69	0.81	0.89	1.14	0.72	0.83	0.73	1.07
Mplus – MLR_{BIC}	-	0.06	-	0.13	-	-	-	-
Mplus – MLR_{AICc}	0.54	0.63	0.50	0.72	0.40	0.70	-	0.50

Note 1. The numbers in the cells are based on the replications in which the dimensionality decision can be reached and the replications in which the dimensionality decision cannot be reached after fitting models up to a certain number of latent traits. For the replications in which the dimensionality decision cannot be reached after fitting models up to a certain number of latent traits, the maximum number of latent traits considered in the study (six for Mplus-MLR and eight for Noharm, Mplus-WLSM, and Mplus-WLSMV based indices) was used in computation. Dashes indicate zero.

Note2. PA: Parallel Analysis, RPA: Revised Parallel Analysis, Noharm-Achi: Noharm based approximate chi-square statistics, Noharm-ALR: Noharm based approximate likelihood ratio chi-square statistics, Noharm – T_{MChi}: Noharm based mean adjusted chi-square statistics, Noharm – T_{MVChi}: Noharm based mean-and-variance adjusted chi-square statistics, Mplus – WLS_{MChi}: Mplus weighted least squares mean adjusted chi-square statistics, Mplus – WLS_{MVChi}: Mplus weighted least squares mean-and-variance adjusted chi-square statistics, Mplus – MLR_{Chi1}: Mplus marginal maximum likelihood unadjusted chi-square difference test, Mplus – MLR_{Chi2}: Mplus marginal maximum likelihood adjusted chi-square difference test, Mplus-WLS_{MRmsea}: Mplus weighted least squares root mean square error approximation with mean adjusted chi-square, Mplus-$WLS_{MVRmsea}$: Mplus weighted least squares root mean square error approximation with mean-and-variance adjusted chi-square, Mplus-WLS_{SRMR}: Mplus weighted least squares standardized root mean square residual, Mplus-WLS_{MCfi}: Mplus weighted least squares comparative fit index with mean adjusted chi-square statistic, Mplus-WLS_{MVCfi}: Mplus weighted least squares comparative fit index with mean-and-variance adjusted chi-square statistic, Mplus – MLR_{AIC}: Mplus Akaike information criterion, Mplus – MLR_{BIC}: Mplus Bayesian information criterion, Mplus – MLR_{AICc}: Mplus corrected Akaike information criterion.

Table 26. *Minimum – Maximum Estimates for the Dimensionality Decision Across 500 Replications*

Subject	Mathematics				Reading			
Number of Items	20		40		20		40	
Sample Size	500	1000	500	1000	500	1000	500	1000
PA	1-1	1-1	1-1	1-2	1-1	1-1	1-1	1-1
Revised PA	1-4	1-5	1-3	1-4	1-3	1-3	1-1	1-3
Detect	2-6	2-6	2-6	3-7	2-6	2-6	2-7	2-7
Noharm – Achi	1-2	1-3	1-3	2-4	1-4	1-4	1-8	3-8
Noharm – ALR	1-2	1-2	1-2	1-2	1-2	1-2	1-3	1-3
Noharm – T_{MChi}	1-5	2-6	2-8	2-8	1-6	1-5	1-7	1-7
Noharm – T_{MVChi}	1-4	1-6	1-6	2-6	1-4	1-5	1-4	1-6
Mplus – WLS_{MChi}	1-4	1-5	1-8	1-8	1-5	1-5	1-8	1-8
Mplus – WLS_{MVChi}	1-4	1-4	1-6	1-6	1-4	1-5	1-4	1-6
Mplus – MLR_{Chi1}	1-6	1-6	1-6	1-6	1-6	1-6	1-6	1-6
Mplus – MLR_{Chi2}	1-4	1-4	1-4	1-5	1-3	1-4	1-4	1-4
Mplus – WLS_{MRmsea}	1-1	1-1	1-1	1-1	1-1	1-1	1-1	1-1
Mplus – $WLS_{MVRmsea}$	1-1	1-1	1-1	1-1	1-1	1-1	1-1	1-1
Mplus – WLS_{SRMR}	1-2	1-1	2-5	1-2	1-3	1-1	2-8	1-4
Mplus – WLS_{MCfi}	1-1	1-1	1-1	1-1	1-1	1-1	1-2	1-1
Mplus – WLS_{MVCfi}	1-1	1-1	1-1	1-1	1-2	1-1	1-2	1-1
Mplus – MLR_{AIC}	1-5	1-6	1-6	1-6	1-4	1-5	1-6	1-5
Mplus – MLR_{BIC}	1-1	1-2	1-1	1-2	1-1	1-1	1-1	1-1
Mplus – MLR_{AICc}	1-3	1-6	1-2	1-6	1-3	1-4	1-1	1-3

Note. PA: Parallel Analysis, RPA: Revised Parallel Analysis, Noharm-Achi: Noharm based approximate chi-square statistics, Noharm-ALR: Noharm based approximate likelihood ratio chi-square statistics, Noharm – T_{MChi}: Noharm based mean adjusted chi-square statistics, Noharm – T_{MVChi}: Noharm based mean-and-variance adjusted chi-square statistics, Mplus – WLS_{MChi}: Mplus weighted least squares mean adjusted chi-square statistics, Mplus – WLS_{MVChi}: Mplus weighted least squares mean-and-variance adjusted chi-square statistics, Mplus – MLR_{Chi1}: Mplus marginal maximum likelihood unadjusted chi-square difference test, Mplus – MLR_{Chi2}: Mplus marginal maximum likelihood adjusted chi-square difference test, Mplus-WLS_{MRmsea}: Mplus weighted least squares root mean square error approximation with mean adjusted chi-square, Mplus-$WLS_{MVRmsea}$: Mplus weighted least squares root mean square error approximation with mean-and-variance adjusted chi-square, Mplus-WLS_{SRMR}: Mplus weighted least squares standardized root mean square residual, Mplus-WLS_{MCfi}: Mplus weighted least squares comparative fit index with mean adjusted chi-square statistic, Mplus-WLS_{MVCfi}: Mplus weighted least squares comparative fit index with mean-and-variance adjusted chi-square statistic, Mplus – MLR_{AIC}: Mplus Akaike information criterion, Mplus – MLR_{BIC}: Mplus Bayesian information criterion, Mplus – MLR_{AICc}: Mplus corrected Akaike information criterion.

Study 2

In Study 2, 500 datasets were generated for 640 conditions by manipulating the number of major factors, amount of variance accounted for by major factors, amount of variance accounted for by minor factors, inter-factor correlations, number of items, and sample size. For each replication, the generated factor loading matrices for major and minor factors were saved to check whether the generated values matched with the corresponding conditions.

Figure 4 represents the minimum, maximum, and average values observed for the variance accounted for by the first minor factor, which was always the largest minor factor, in the generated minor factor loading matrices across all replications. A similar depiction is used throughout this chapter when reporting the results for Study 2. Each figure has four panels, and each panel represents a condition as an interaction between the levels of inter-factor correlations and variance accounted for by minor factors. For each panel, the x-axis represents 40 different structures of major dimensions generated in the study in ascending order of the number of major dimensions. For instance, "30.10" indicates a factor structure with two major dimensions accounting for 30% and 10% of the variance, and "20.20.5.5" indicates a factor structure with four major dimensions accounting for 20%, 20%, 5%, and 5% of the variance, respectively. The y-axis represents the relevant statistics.

In Figure 4, the results are aggregated across different conditions of sample size and number of items within each panel. On average, the first minor factor in the generated factor loading matrices accounted for about 1% of the total variance for the conditions in which the total variance accounted for by all minor factors was 10%. In those conditions, the amount of variance accounted for by the first minor factor was not more than 1.8% or less than 0.3% for any replication. For the conditions in which the total variance accounted for by all minor factors was 20%, the first minor factor accounted for about 1.9% of the variance on average, and the amount of variance accounted for by the first minor factor was not more than 3.5% or less than 0.8% for any replication. Recall that there were 50 minor factors generated for each replication, and each successive minor factor accounted for 0.9 times the variance accounted for by the

preceding minor factor. For instance, if the first minor factor accounted for 2% of the variance, the successive minor factors accounted for 1.8%, 1.62%, 1.46%, 1.31%, and so on.

Figure 5 represents the average value observed for the variance accounted for by the major dimensions in the generated major factor loading matrices in all conditions. The results are aggregated across different conditions of sample size and number of items within each panel. Although some minor deviations occurred, Figure 4 and Figure 5 indicate that the generated factor loading matrices for the major and minor dimensions matched the factor structures proposed to be simulated for Study 2.

All simulated datasets within each of the 640 conditions were analyzed using parallel analysis, revised parallel analysis, DETECT, NOHARM, and WLSM, WLSMV, and MLR estimators in Mplus, and the number of dimensions suggested by the different criteria was identified. The reference for all outcome variables was the number of major dimensions (quasi-true number of dimensions) in the true generating model. The proportion of datasets in which the quasi-true number of dimensions was correctly identified, the bias with respect to the quasi-true number of dimensions, and the root mean squared deviation from the quasi-true number of dimensions were reported. Due to the large number of simulation conditions and large number of dimensionality assessment methods considered in Study 2, the results are presented separately for each method.

Parallel Analysis. The results of parallel analysis are shown for all 640 conditions in Figure 6, Figure 7, and Figure 8. The parallel analysis selected a one-dimensional model almost all the time regardless of the underlying factor structure as depicted in Figure 6, Figure 7, and Figure 8. The only exception was the condition in which there were two major dimensions accounting for 40% and 20% of the variance, respectively. In this condition, parallel analysis correctly identified two major dimensions for 20% of the replications. The patterns in Figure 7 and Figure 8 are due to identifying one dimension for almost every replication in all conditions. There was no bias for one-dimensional structures, and the bias was equal to -1, -2, and -3 for structures with two, three, and four major dimensions, respectively.

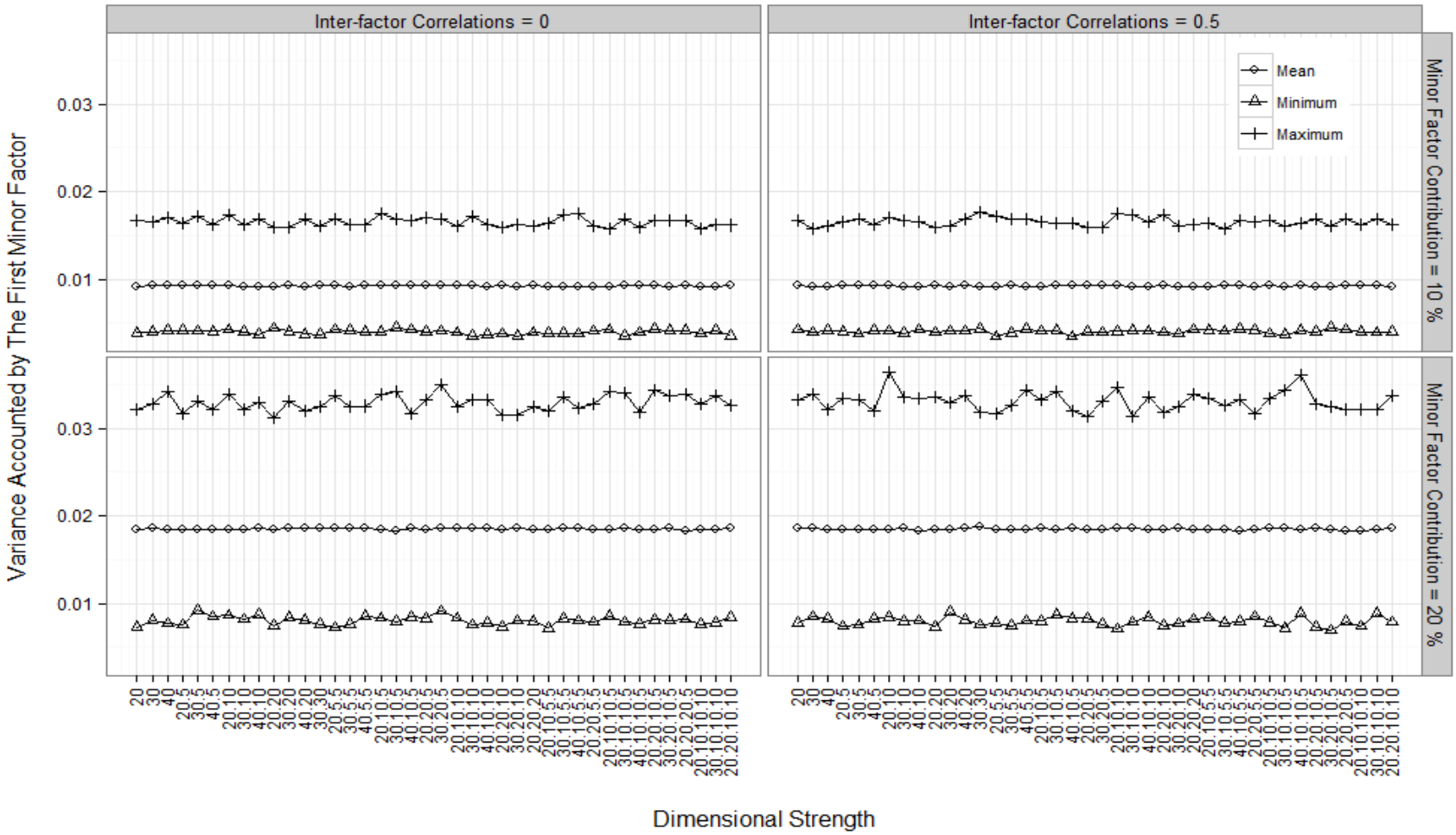


Figure 4. The Average, Minimum, and Maximum Values for the Variance Accounted for by the First Minor Factor (Largest Minor Factor) in Generated Minor Factor Loading Matrices across Simulation Conditions

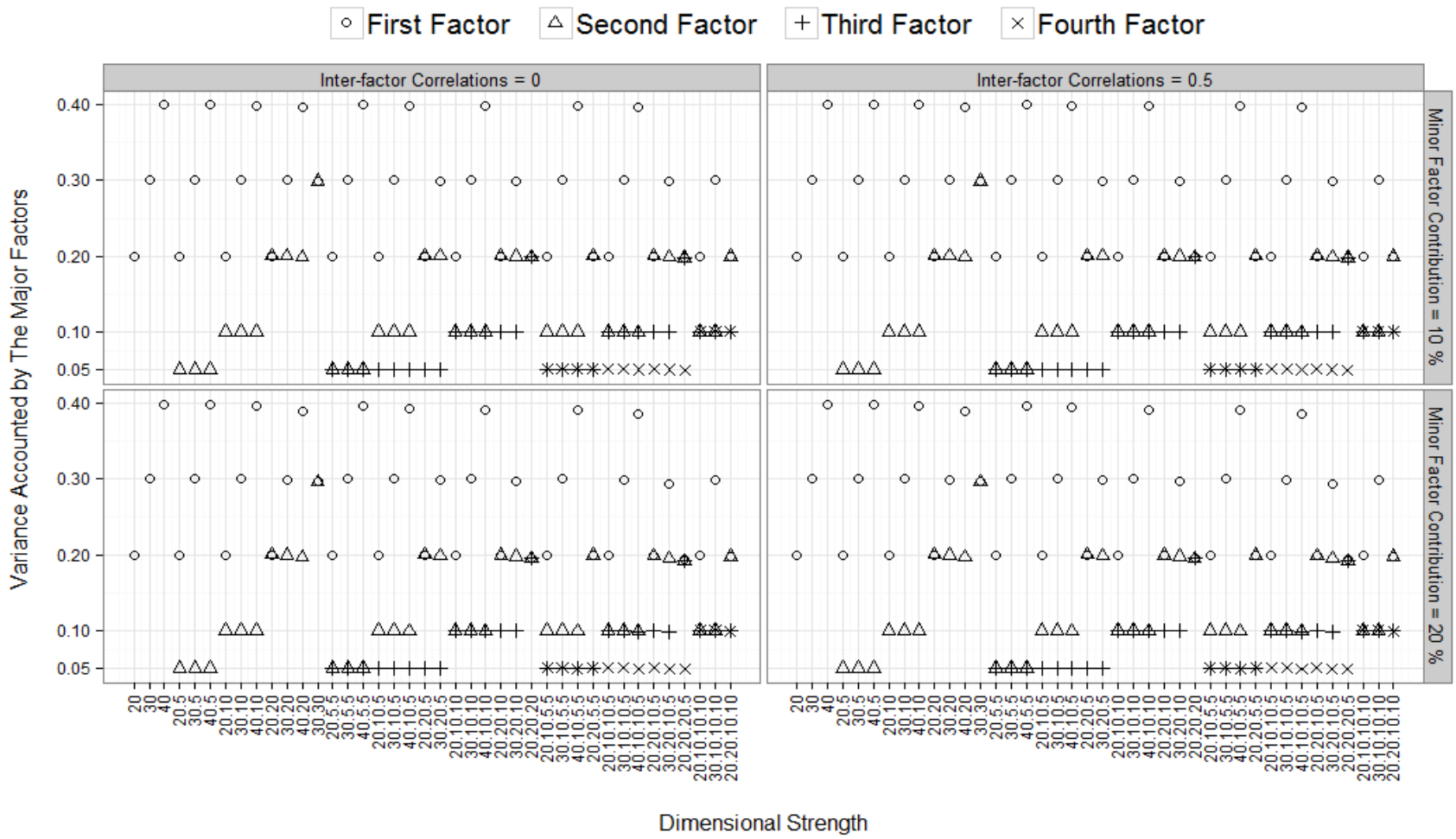


Figure 5. The Average Values for the Variance Accounted for by the Major Dimensions in Generated Major Factor Loading Matrices across Simulation Conditions

The results for parallel analysis were as expected and not surprising. As illustrated in Table 2, the interpretation of the first eigenvalue depends on the underlying factor structure. When the underlying structure is complex, the first eigenvalue seems to be an indicator of the total variance accounted for by the whole factor structure. Figure 9 depicts this fact for all conditions simulated in Study 2. As an exception among all other figures presented in this chapter, the x-axis in Figure 9 was rearranged in ascending order of the total variance accounted for by the major factors in order to make the pattern more visible. As seen in Figure 9, the average first eigenvalue across all replications is reflecting the magnitude of the total variance accounted for by the whole structure, and increases as a function of the total variance accounted for by the whole structure. For instance, the average first eigenvalue was approximately equal to 11 and 22 for the 20- and 40- item tests for the conditions in which there were two major dimensions accounting 40% and 10% of the total variance, there was no correlation among the dimensions, and minor factors were accounting for 10% of the variance. The first eigenvalues indicated about 55% percent of the variance, which was roughly reflecting the sum of variances accounted for by the major and minor factors in the generating structure. As a result of this fact, parallel analysis tends to select one-dimensional models regardless of the generating factor structure when the structure is complex.

Revised Parallel Analysis. The results of revised parallel analysis are shown for all 640 conditions in Figure 10, Figure 11, and Figure 12. The results were slightly different than parallel analysis. Revised parallel analysis correctly identified one dimension almost every time and incorrectly identified two dimensions for a small amount of replications when the quasi-true number of dimensions was one. When there were two major dimensions and each dimension accounted for at least 10 % of the variance, the proportion of correctly identifying two major dimensions increased, especially when the sample size was 1000, number of items was 40, and inter-factor correlation was zero. However, when there was one strong and one relatively weaker major dimension, revised parallel analysis could not identify the weaker dimension. Also, when the inter-factor correlation was 0.5, revised parallel analysis tended to identify one dimension when the quasi-true number of dimensions was two.

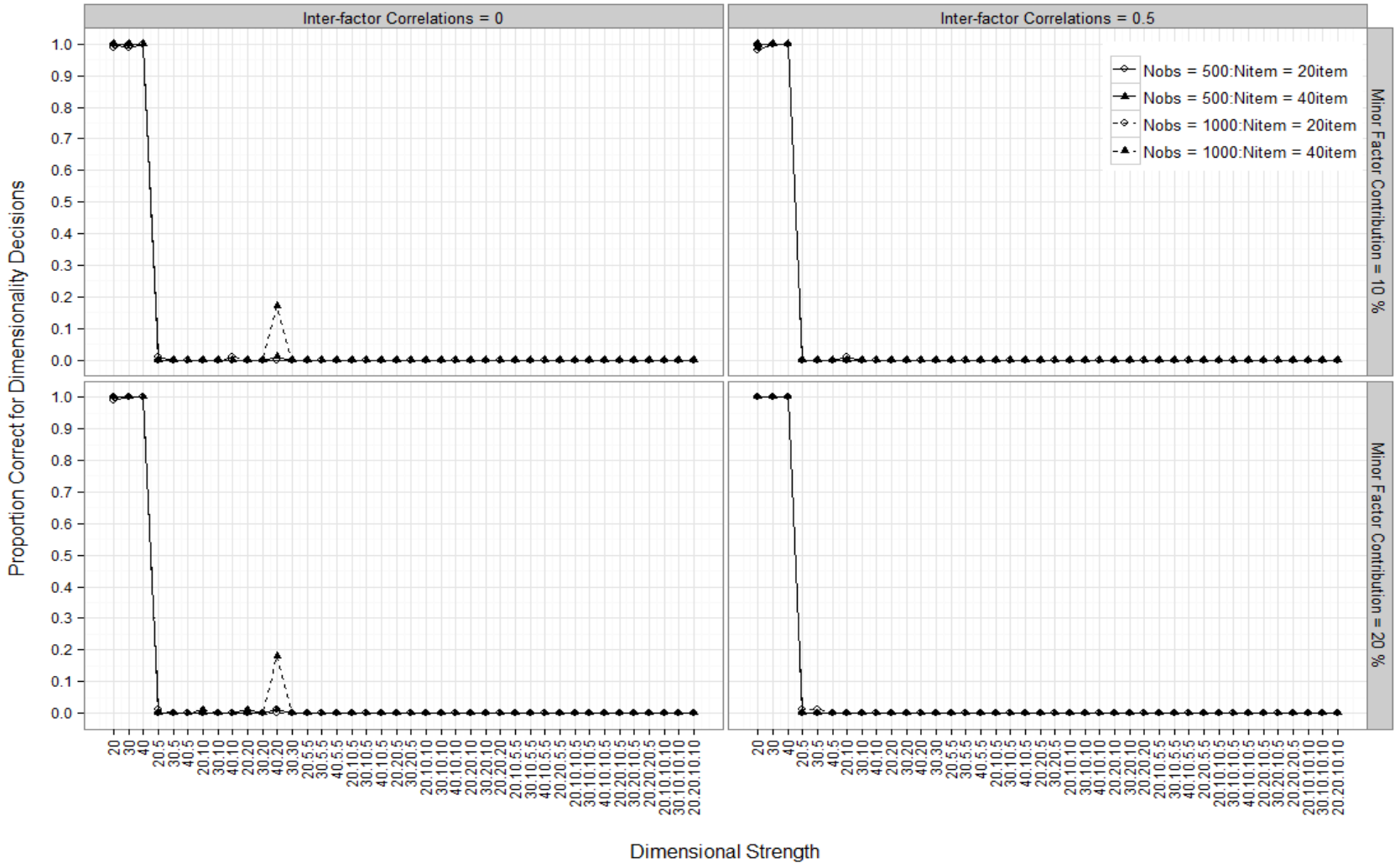


Figure 6. Proportion of Datasets in Which the Quasi-true Number of Dimensions is Correctly Identified by **Parallel Analysis**

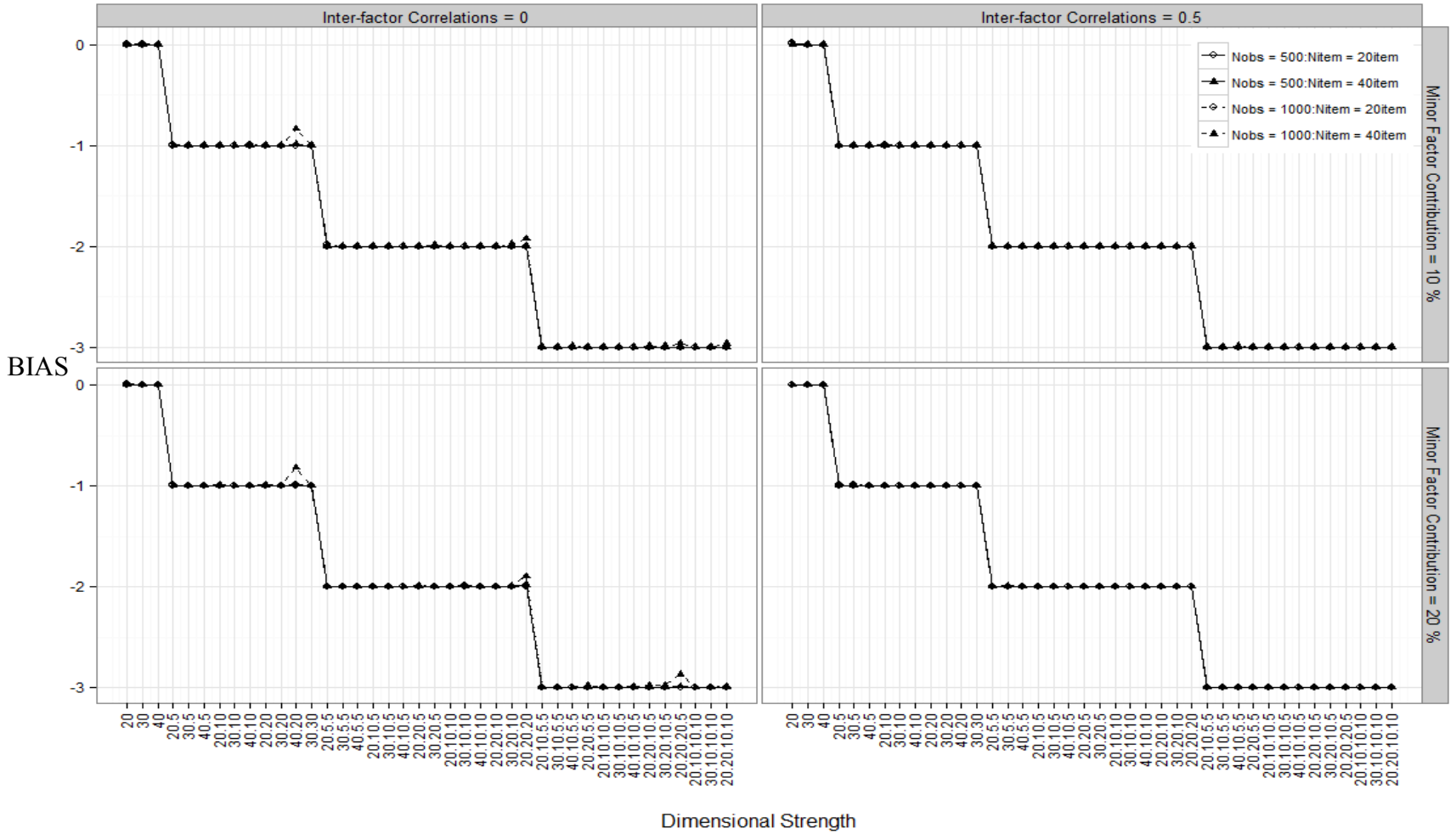


Figure 7. Bias with respect to the Quasi-true Number of Dimensions for **Parallel Analysis**

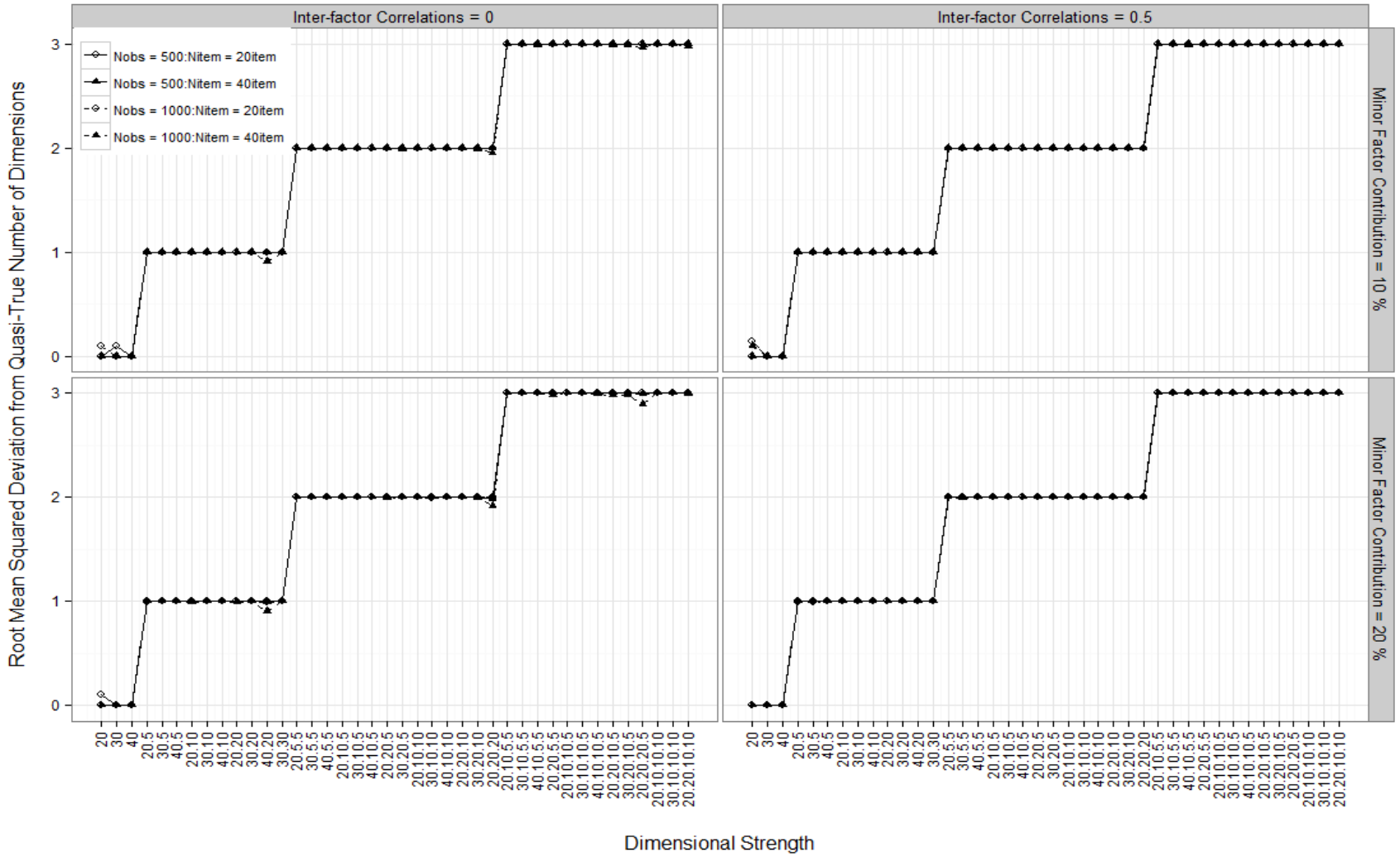


Figure 8. Root Mean Squared Deviation from the Quasi-true Number of Dimensions for **Parallel Analysis**

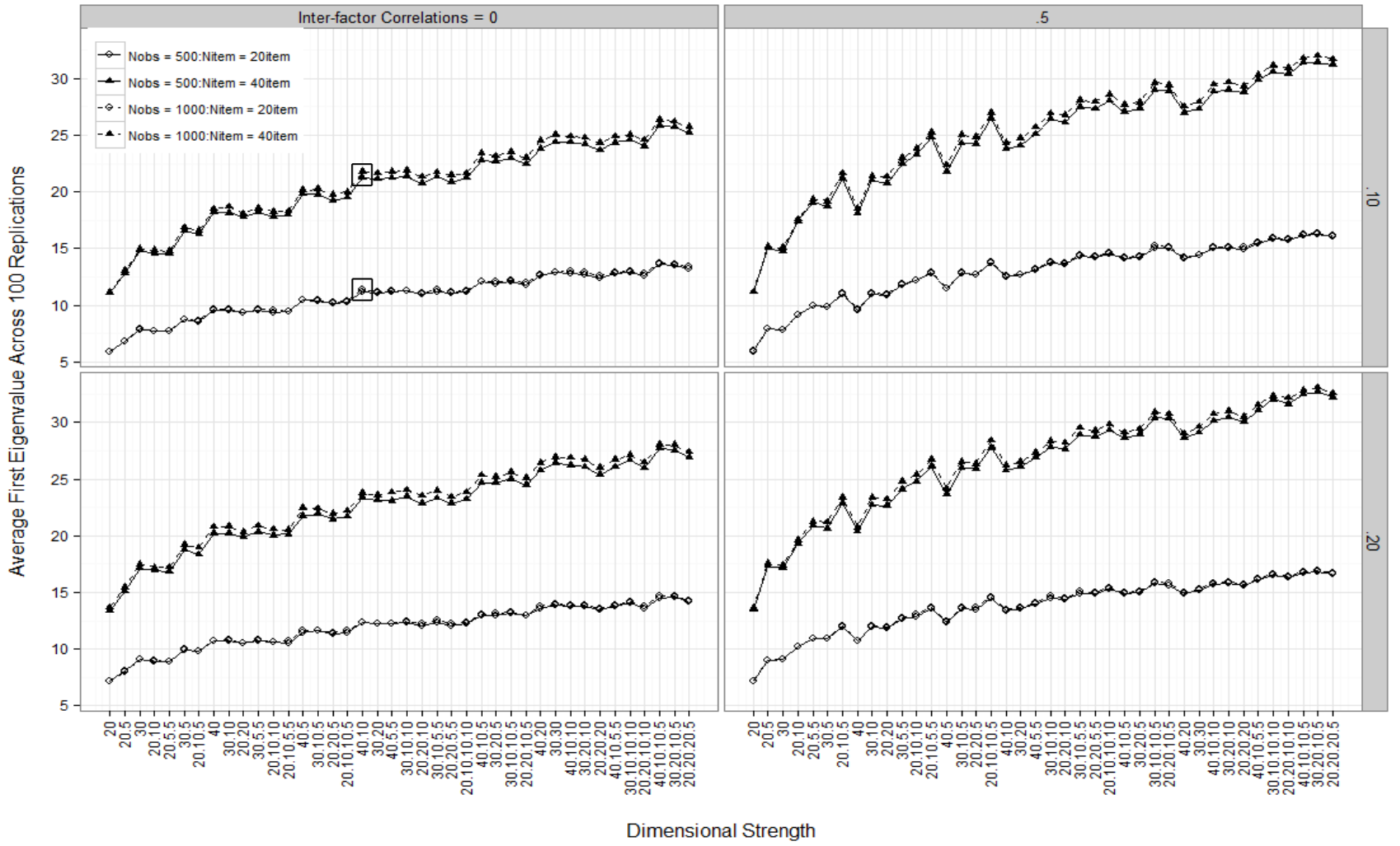


Figure 9. Average Value of the First Eigenvalue Across 100 Replications

When the generating model included three or four major dimensions, revised parallel analysis tended to identify two dimensions when the sample size was 1000, number of items was 40, and inter-factor correlation was 0, and tended to identify only one dimension for other occasions.

DETECT. The results from the DETECT analysis are given in Figure 13, Figure 14, and Figure 15. Figure 13 shows that DETECT was not very successful in identifying the major dimensions when the quasi-true number of dimensions was one or two. In most of these conditions, the proportions of correctly identifying the number of major dimensions were close to zero. Relatively higher success rates, between 30% and 70%, occurred when the quasi-true number of dimensions was three or four, especially for the conditions in which the sample size was 500 and the number of items was 20. There seemed to be an interaction among the number of items, inter-factor correlation, and the quasi-true number of dimensions in terms of the proportion correct for dimensionality decisions. When the inter-factor correlation was zero, the proportion correct was higher in conditions with three major dimensions than the conditions with four major dimensions. When the inter-factor correlation was 0.5, the proportion correct was again highest in conditions with three major dimensions for the 20-item tests, but it was highest in conditions with four major dimensions for the 40-item tests.

Figure 14 shows a systematic bias in dimensionality decisions given by the DETECT analysis. DETECT tended to identify three or four dimensions in most conditions. For instance, the bias was equal to about 3 when the quasi-true number of dimensions was one, indicating that DETECT identified four dimensions on average in those conditions. Similarly, the bias was between 1.5 and 2 for the conditions in which the quasi-true number of dimensions was two, indicating that DETECT identified three or four dimensions on average in those conditions. The bias was between 0 and 1 for the conditions with three major dimensions, and between 0 and -1 for the conditions with four major dimensions. DETECT tended to identify three or four dimensions in most conditions regardless of the underlying factor structure.

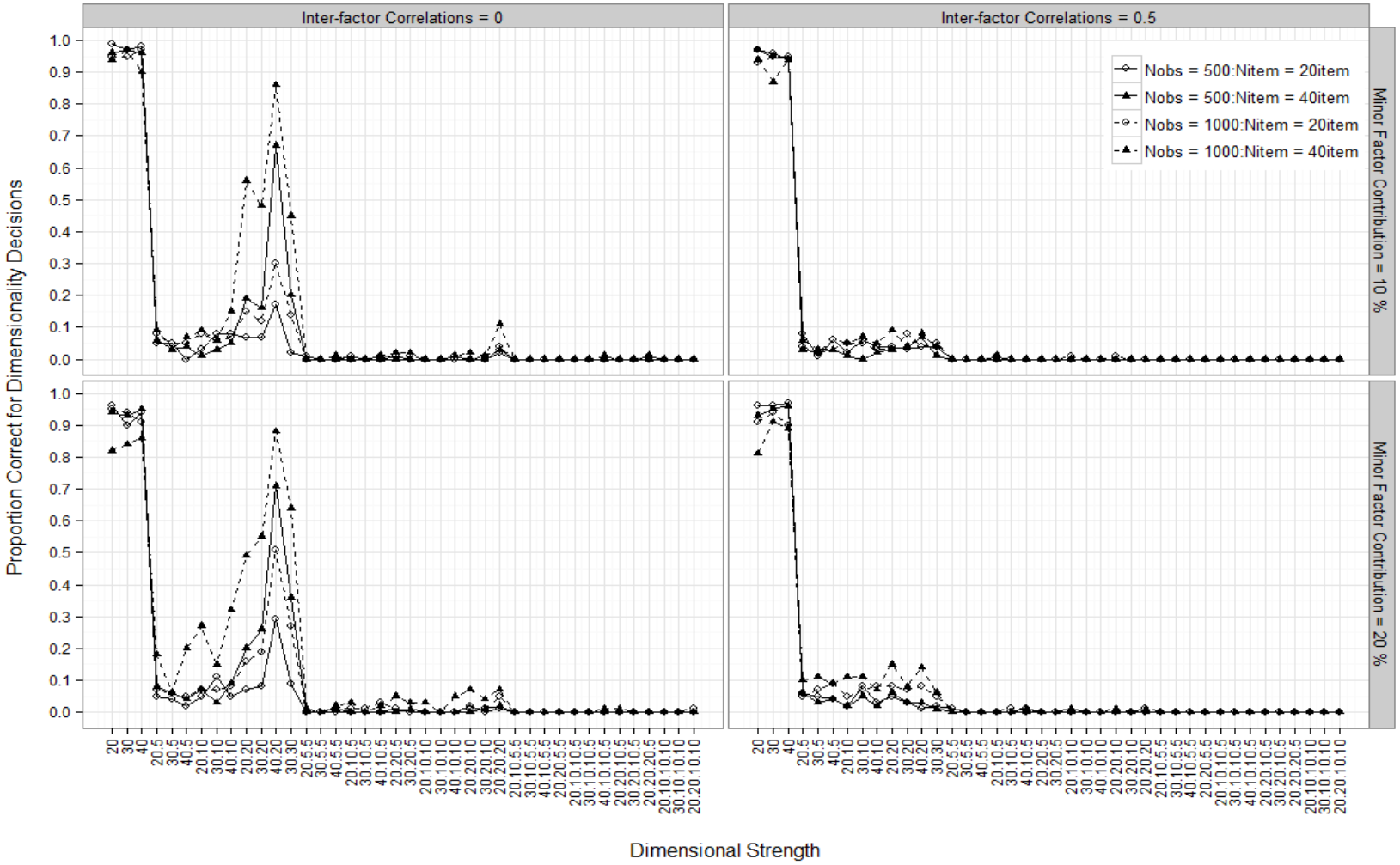


Figure 10. Proportion of Datasets in Which the Quasi-true Number of Dimensions is Correctly Identified by **Revised Parallel Analysis**

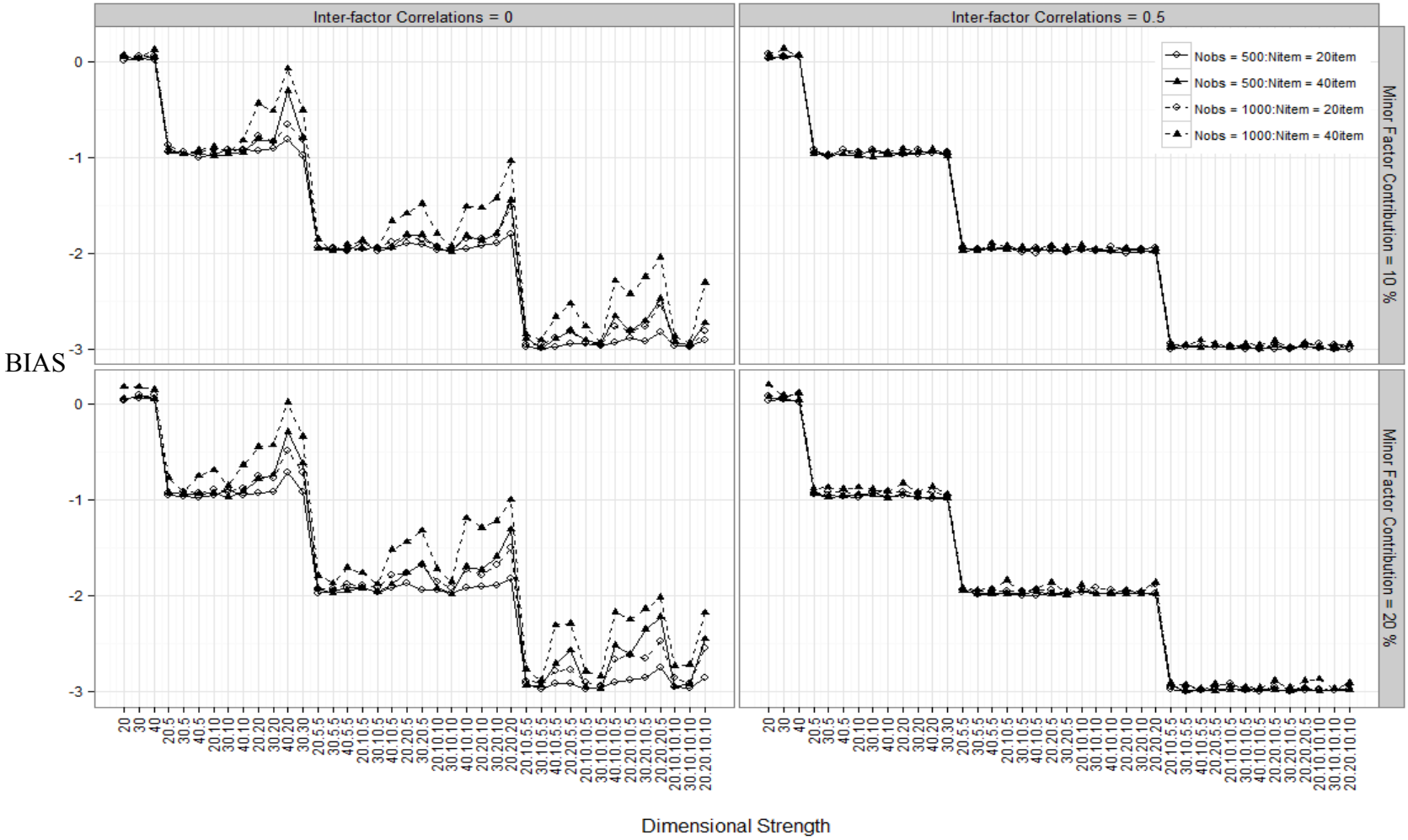


Figure 11. Bias with respect to the Quasi-true Number of Dimensions for **Revised Parallel Analysis**

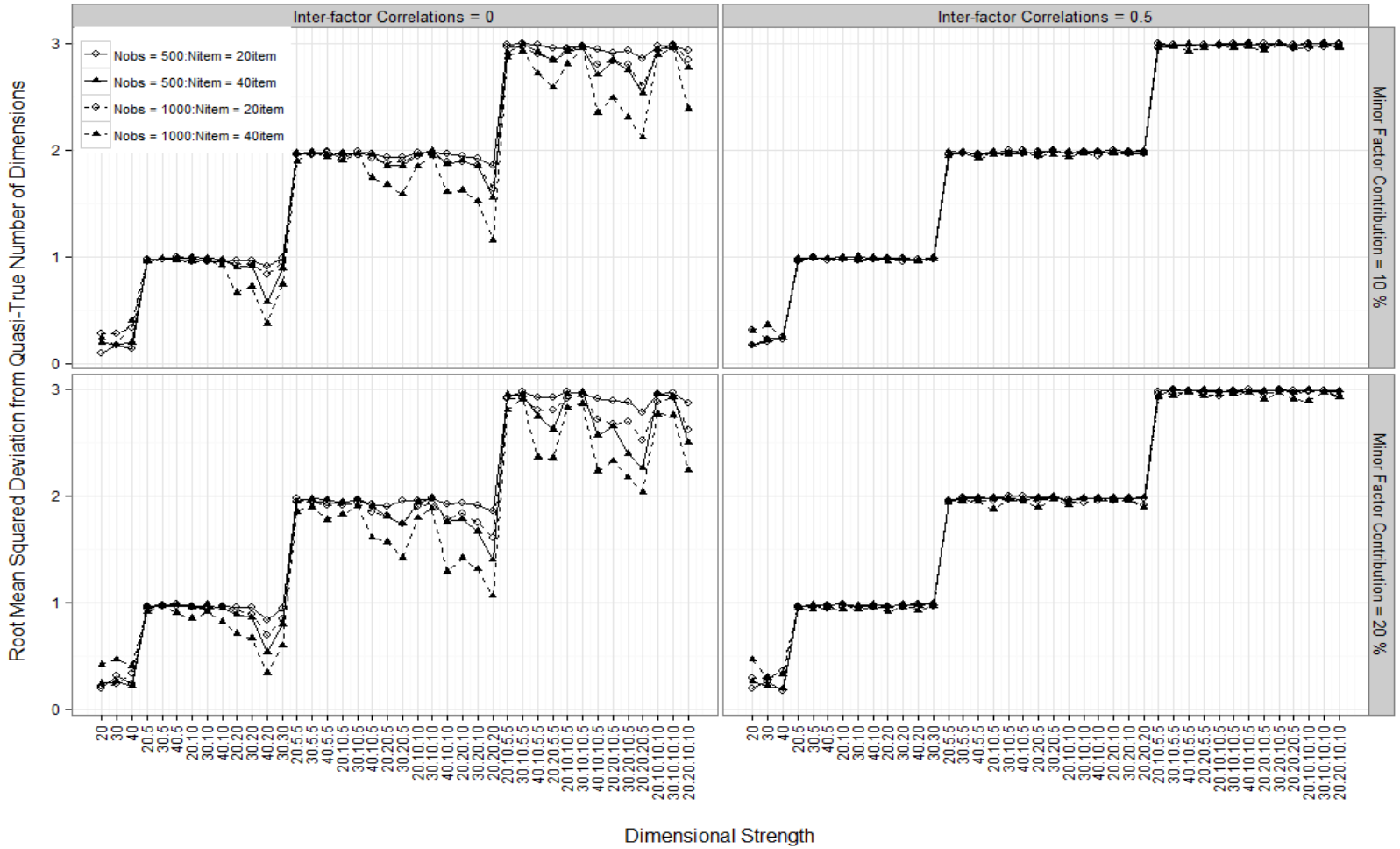


Figure 12. Root Mean Squared Deviation from the Quasi-true Number of Dimensions for **Revised Parallel Analysis**

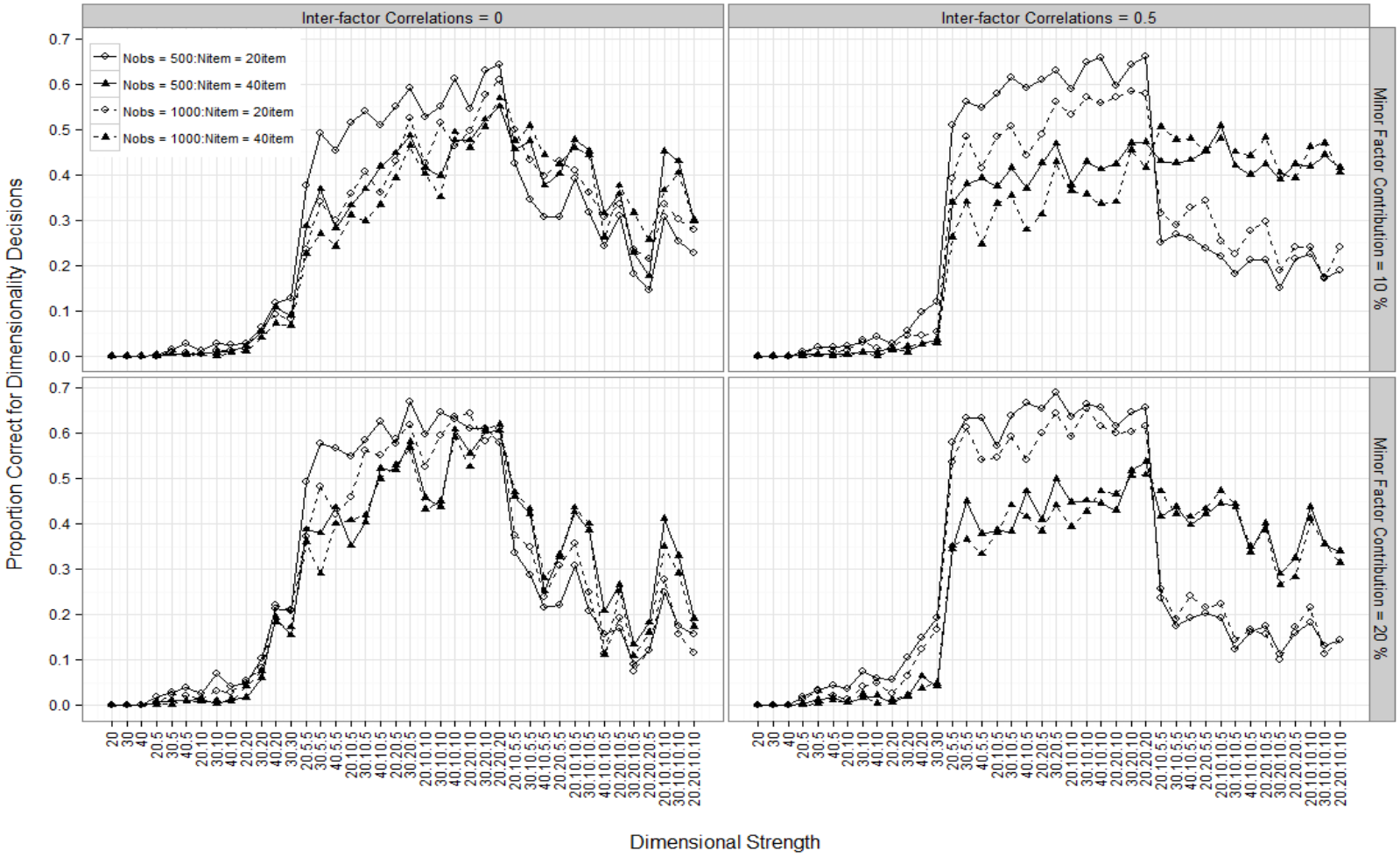


Figure 13. Proportion of Datasets in Which the Quasi-true Number of Dimensions is Correctly Identified by **DETECT**

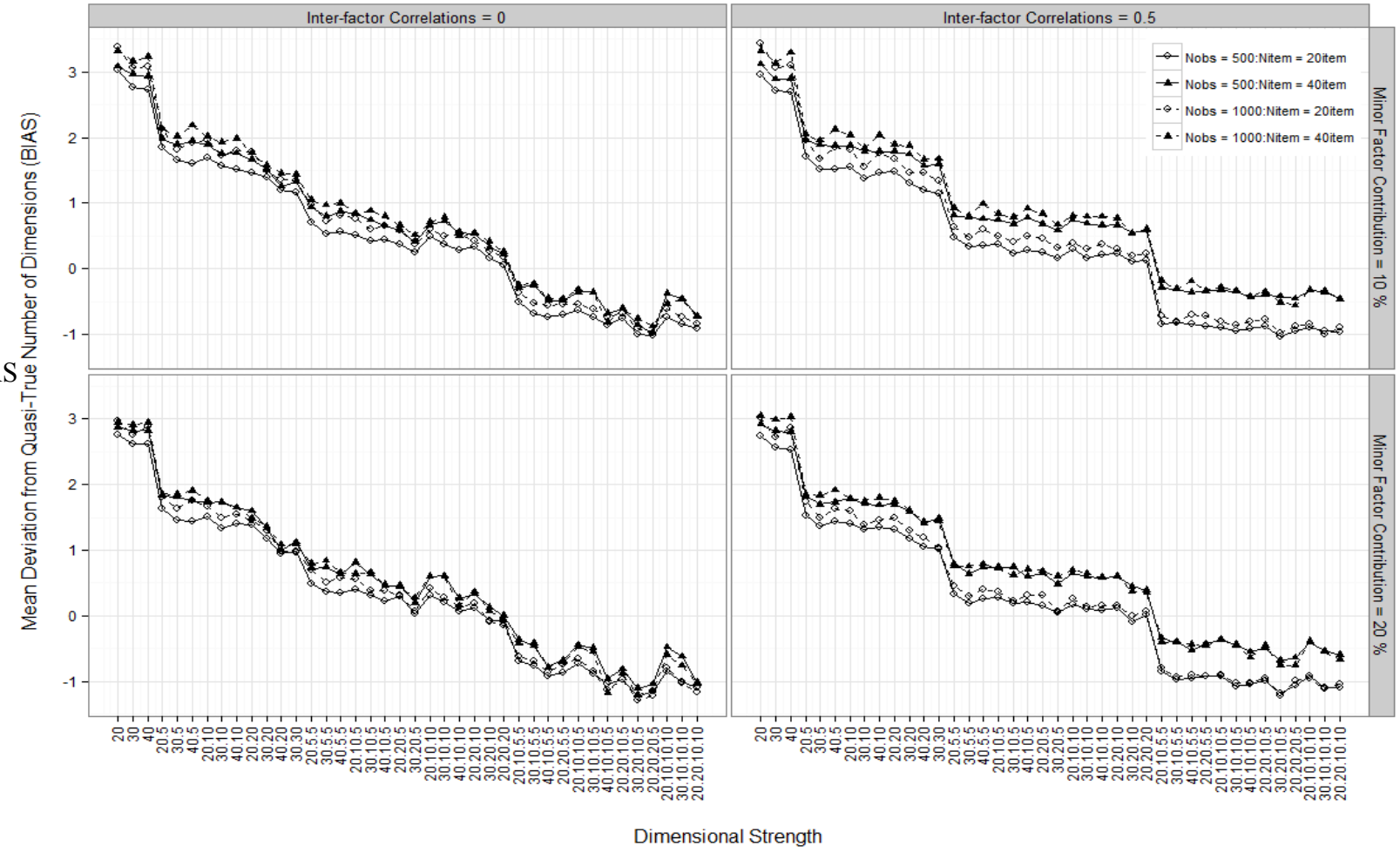


Figure 14. Bias with respect to the Quasi-true Number of Dimensions for **DETECT**

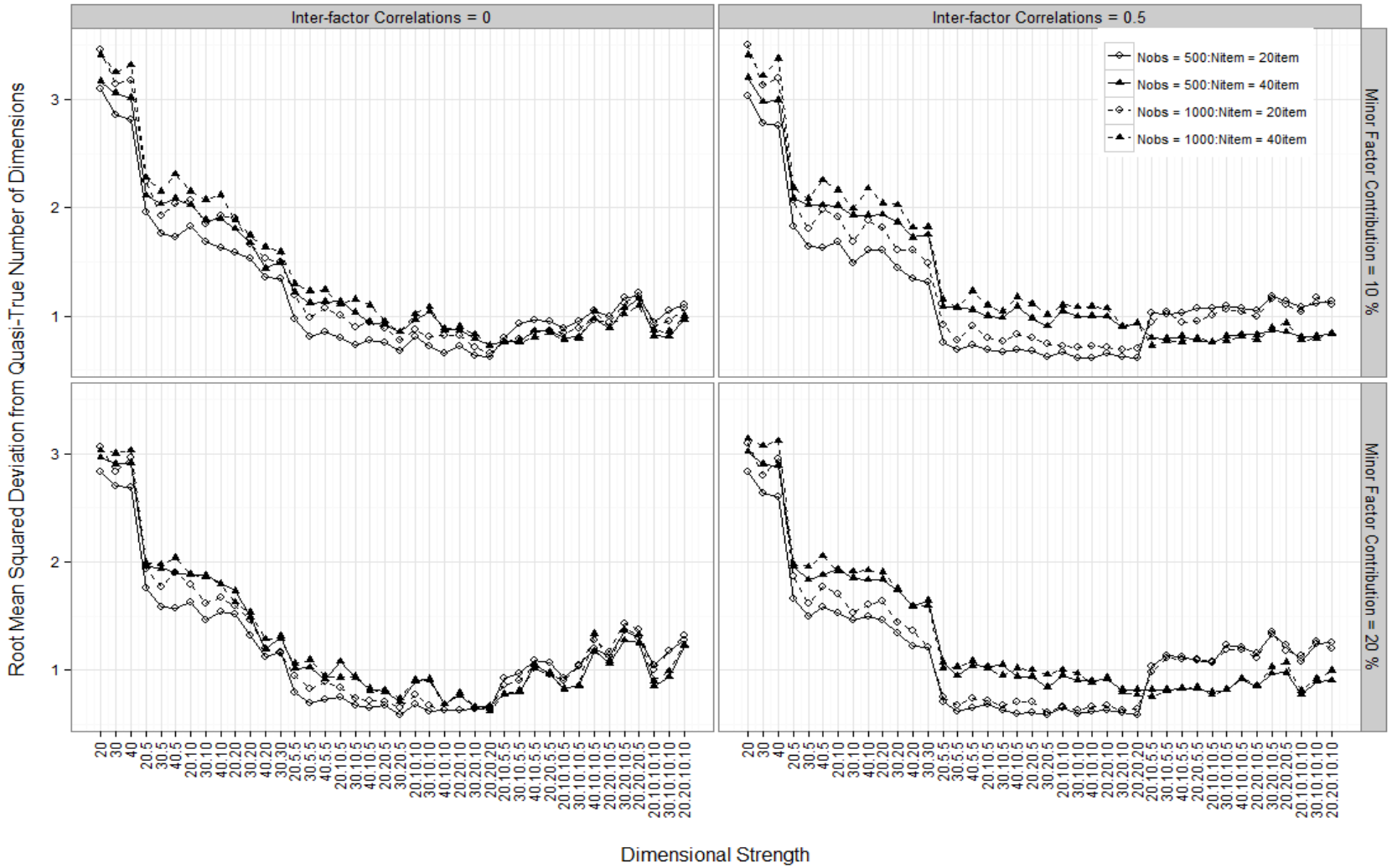


Figure 15. Root Mean Squared Deviation from the Quasi-true Number of Dimensions for **DETECT**

NOHARM-based Chi-Square Statistics. The results for the NOHARM-based chi-square statistics are shown in Figure 16 through Figure 27. When a simulated dataset was analyzed using NOHARM, the largest model fitted was the eight-dimensional model. As described in Study 1, a chi-square statistic was still significant after fitting the eight-dimensional model for some replications, and a decision was not reached. The number of dimensions was assumed to be eight, which is the largest model fitted, for those replications when computing the relevant statistics and creating the figures. The convergence problem never occurred in the NOHARM analysis for any replication, so there was no condition in which the dimensionality decision was not reached due to convergence problems.

Table 27 shows the proportion of simulated datasets with no-decision after computing the NOHARM Achi statistic for the eight-dimensional model. These replications only occurred when the sample size was 1000 and the number of items was 20. Figure 16 shows that the NOHARM Achi statistic was not successful in correctly selecting the model with the quasi-true number of dimensions. Figure 17 indicates the bias in decisions regarding the number of major dimensions. The NOHARM Achi tended to select one-dimensional models almost every time when the sample size was 500 regardless of the underlying factor structure, and selected more complex models in rare occasions when the sample size was 1000.

Table 28 shows the proportion of simulated datasets with no-decision after computing the NOHARM ALR statistic for the eight-dimensional model. NOHARM ALR selected more complex models compared to NOHARM Achi. NOHARM ALR indicated at least eight dimensions for a large amount of replications when the sample size was 1000 and number of items was 20. Particularly, when the total amount of variance accounted for by the major dimensions exceeded 60%, the proportions of datasets in which NOHARM ALR indicated at least eight dimensions were more than 90%. Figure 19 shows that NOHARM ALR was relatively more successful than NOHARM Achi in correctly selecting the model with the quasi-true number of dimensions, but the success rate was not more than 30% in most conditions.

Table 27. *Proportion of Datasets in Which the Dimensionality Decision Cannot be Reached by the NOHARM Approximate Chi-Square Statistic*

Minor Factors	10 %				20 %			
Number of Items	20		40		20		40	
Sample Size	500	1000	500	1000	500	1000	500	1000
20	-	-	-	-	-	-	-	-
30	-	-	-	-	-	-	-	-
40	-	-	-	-	-	-	-	-
20.5	-	-	-	-	-	-	-	-
30.5	-	-	-	-	-	-	-	-
40.5	-	-	-	-	-	-	-	-
20.10	-	-	-	-	-	-	-	-
30.10	-	-	-	-	-	-	-	-
40.10	-	0.01	-	-	-	-	-	-
20.20	-	-	-	-	-	-	-	-
30.20	-	-	-	-	-	-	-	-
40.20	-	0.02	-	-	-	0.09	-	-
30.30	-	-	-	-	-	0.03	-	-
20.5.5	-	-	-	-	-	-	-	-
30.5.5	-	-	-	-	-	-	-	-
40.5.5	-	-	-	-	-	0.01	-	-
20.10.5	-	-	-	-	-	-	-	-
30.10.5	-	-	-	-	-	-	-	-
40.10.5	-	0.01	-	-	-	0.04	-	-
20.20.5	-	-	-	-	-	-	-	-
30.30.5	-	-	-	-	-	-	-	-
20.10.10	-	-	-	-	-	-	-	-
30.10.10	-	-	-	-	-	-	-	-
40.10.10	-	0.01	-	-	-	0.07	-	-
20.20.10	-	-	-	-	-	-	-	-
30.20.10	-	-	-	-	-	0.02	-	-
20.20.20	-	0.01	-	-	-	0.05	-	-
20.10.5.5	-	-	-	-	-	-	-	-
30.10.5.5	-	-	-	-	-	-	-	-
40.10.5.5	-	0.03	-	-	-	0.07	-	-
20.20.5.5	-	-	-	-	-	-	-	-
20.10.10.5	-	-	-	-	-	-	-	-
30.10.10.5	-	-	-	-	-	-	-	-
40.10.10.5	-	0.08	-	-	-	0.16	-	-
20.20.10.5	-	-	-	-	-	-	-	-
30.20.10.5	-	0.02	-	-	-	0.08	-	-
20.20.20.5	-	0.04	-	-	-	0.12	-	-
20.10.10.10	-	-	-	-	-	-	-	-
30.10.10.10	-	-	-	-	-	0.01	-	-
20.20.10.10	-	-	-	-	-	0.02	-	-

Note. Dashes indicate zero.

Table 28. *Proportion of Datasets in Which the Dimensionality Decision Cannot be Reached by the NOHARM Approximate Likelihood Ratio Chi-Square Statistic*

Minor Factors	10 %				20 %			
Number of Items	20		40		20		40	
Sample Size	500	1000	500	1000	500	1000	500	1000
20	-	0.02	-	-	-	0.03	-	-
30	-	0.01	-	-	-	0.01	-	-
40	-	0.03	-	-	-	0.23	-	0.02
20.5	-	0.02	-	-	-	0.01	-	-
30.5	-	0.01	-	-	-	0.02	-	-
40.5	-	0.14	-	0.01	0.02	0.50	-	0.06
20.10	-	0.01	-	-	-	0.03	-	-
30.10	-	0.01	-	-	-	0.06	-	-
40.10	0.01	0.37	-	0.05	0.05	0.75	-	0.18
20.20	-	0.03	-	-	-	0.14	-	0.01
30.20	-	0.27	-	0.02	0.02	0.63	-	0.05
40.20	0.11	0.83	0.01	0.28	0.22	0.97	0.01	0.61
30.30	0.07	0.77	-	0.15	0.18	0.95	0.01	0.51
20.5.5	-	0.01	-	-	-	0.01	-	-
30.5.5	-	0.01	-	-	-	0.11	-	-
40.5.5	0.01	0.41	-	0.04	0.08	0.76	-	0.26
20.10.5	-	0.01	-	-	-	0.04	-	-
30.10.5	-	0.08	-	-	0.01	0.36	-	0.01
40.10.5	0.06	0.65	-	0.18	0.16	0.87	0.01	0.50
20.20.5	-	0.14	-	0.01	0.01	0.44	-	0.02
30.30.5	0.04	0.57	-	0.08	0.11	0.84	-	0.31
20.10.10	-	0.02	-	-	-	0.17	-	-
30.10.10	0.01	0.26	-	-	0.05	0.60	-	0.05
40.10.10	0.15	0.81	0.01	0.43	0.32	0.95	0.05	0.70
20.20.10	0.01	0.34	-	0.01	0.06	0.67	-	0.10
30.20.10	0.12	0.79	0.01	0.30	0.22	0.96	0.02	0.61
20.20.20	0.13	0.85	0.01	0.40	0.30	0.96	0.04	0.67
20.10.5.5	-	0.03	-	-	-	0.20	-	-
30.10.5.5	0.01	0.29	-	0.01	0.07	0.59	-	0.08
40.10.5.5	0.20	0.82	0.03	0.50	0.38	0.96	0.10	0.71
20.20.5.5	0.02	0.39	-	0.02	0.08	0.69	-	0.16
20.10.10.5	-	0.17	-	-	0.02	0.38	-	0.02
30.10.10.5	0.06	0.44	-	0.08	0.15	0.83	0.01	0.36
40.10.10.5	0.36	0.93	0.14	0.64	0.52	0.99	0.29	0.88
20.20.10.5	0.09	0.58	-	0.14	0.17	0.87	0.01	0.43
30.20.10.5	0.31	0.92	0.10	0.57	0.52	0.99	0.27	0.81
20.20.20.5	0.35	0.95	0.14	0.63	0.52	0.99	0.39	0.83
20.10.10.10	0.02	0.29	-	0.01	0.04	0.59	-	0.12
30.10.10.10	0.15	0.72	0.01	0.35	0.30	0.96	0.05	0.58
20.20.10.10	0.17	0.78	0.02	0.42	0.33	0.96	0.13	0.63

Note. Dashes indicate zero.

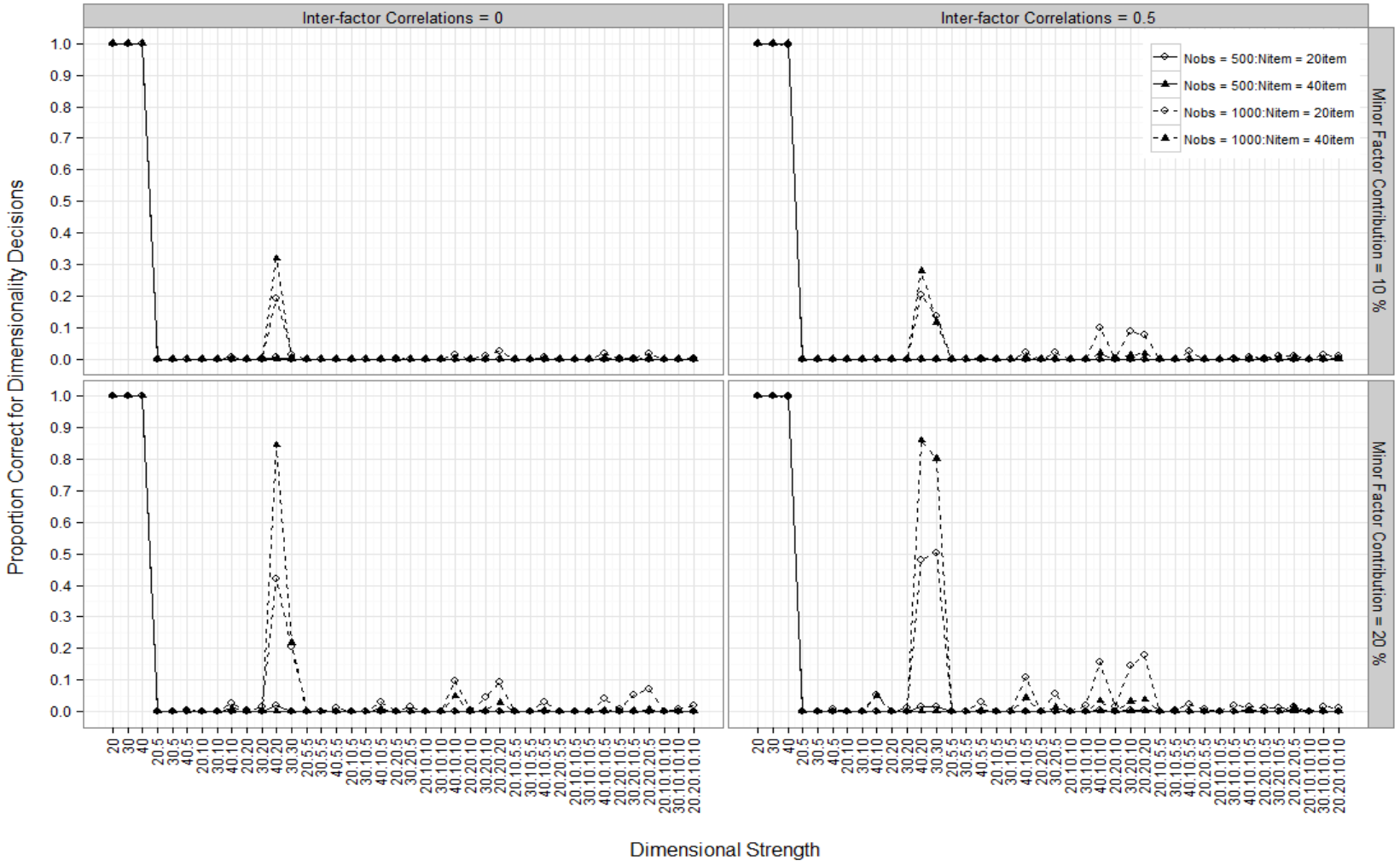


Figure 16. Proportion of Datasets in Which the Quasi-true Number of Dimensions is Correctly Identified by the **NOHARM Approximate Chi-Square Statistic**

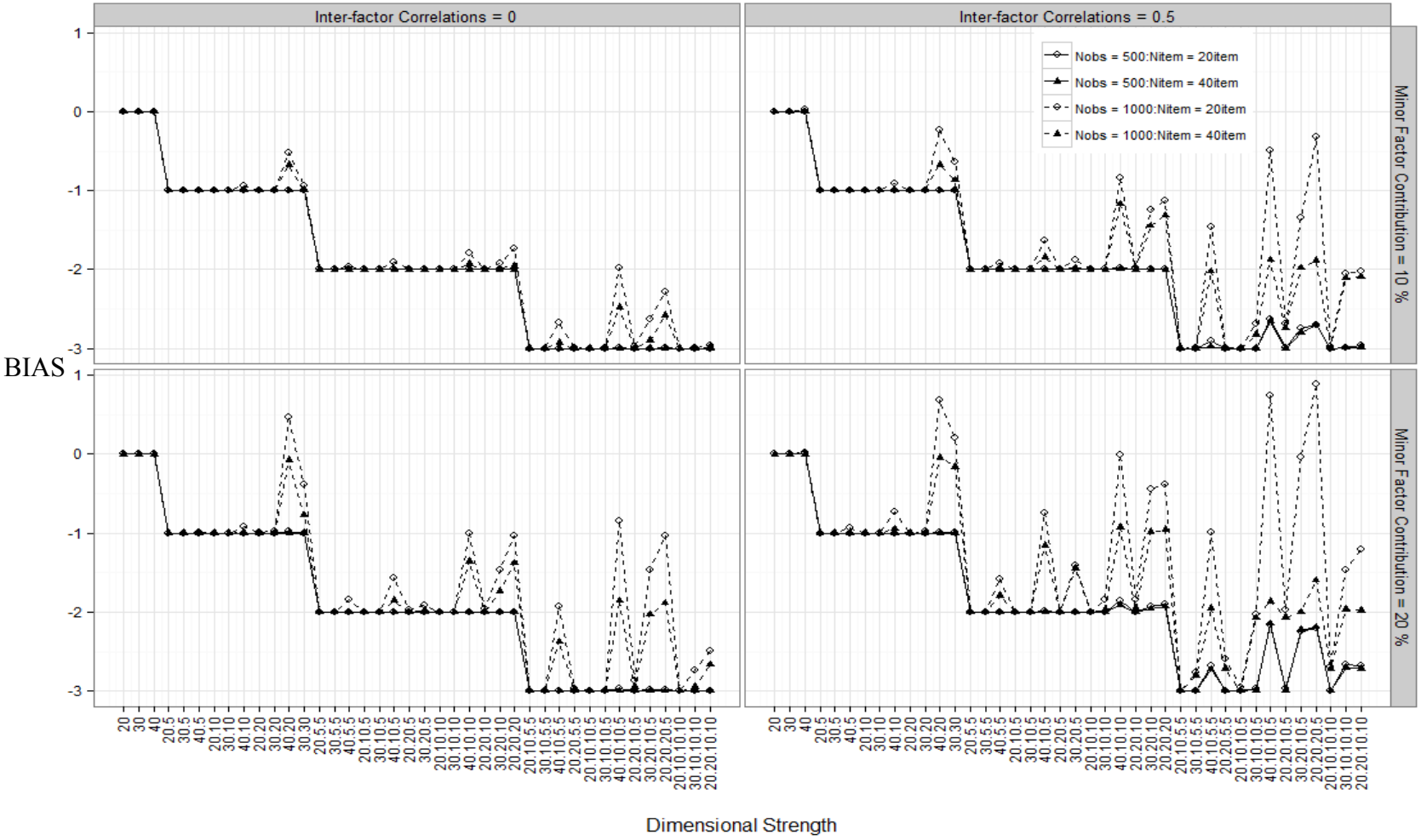


Figure 17. Bias with respect to the Quasi-true Number of Dimensions for the **NOHARM Approximate Chi-Square Statistic**

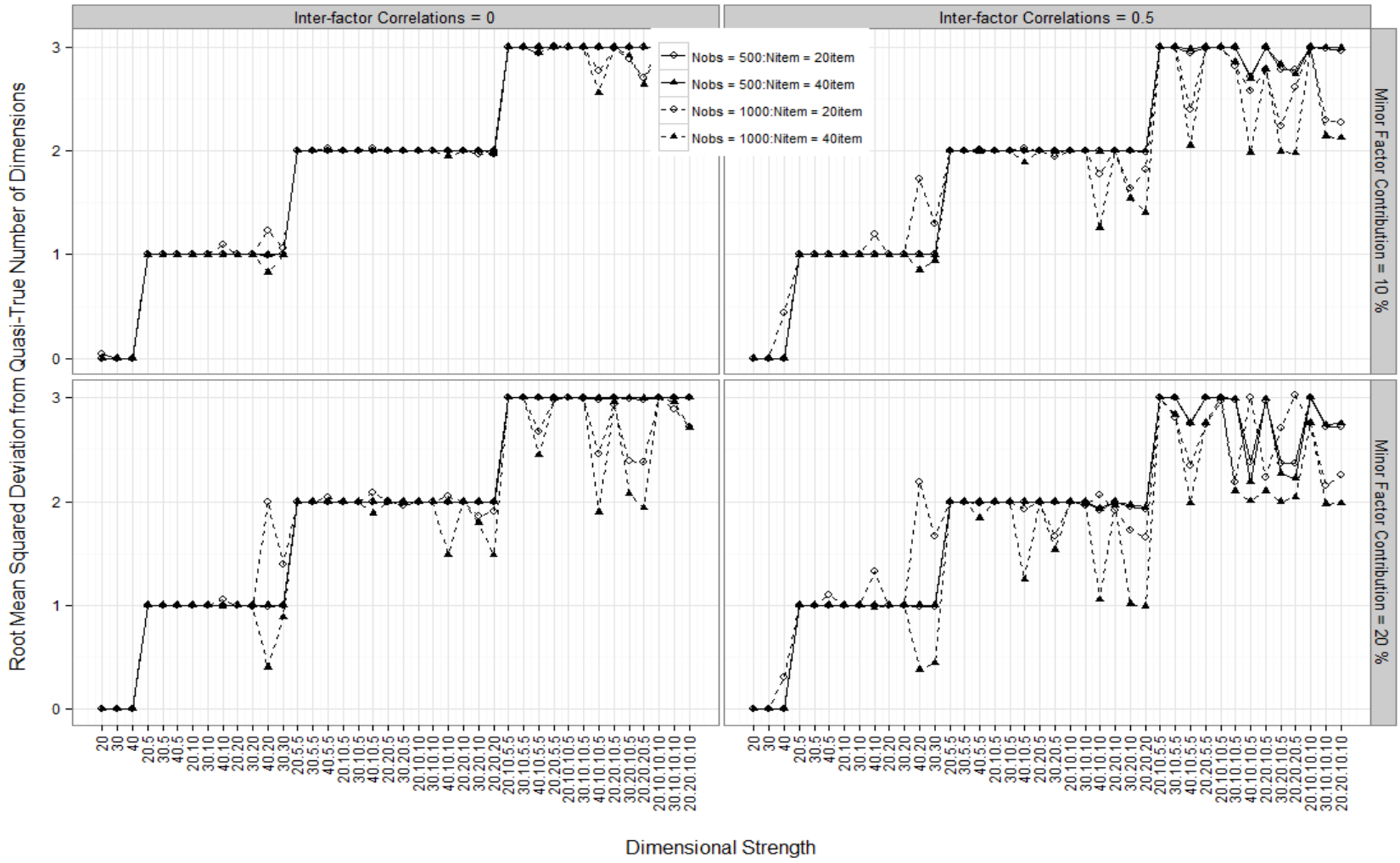


Figure 18. Root Mean Squared Deviation from the Quasi-true Number of Dimensions for the **NOHARM Approximate Chi-Square Statistic**

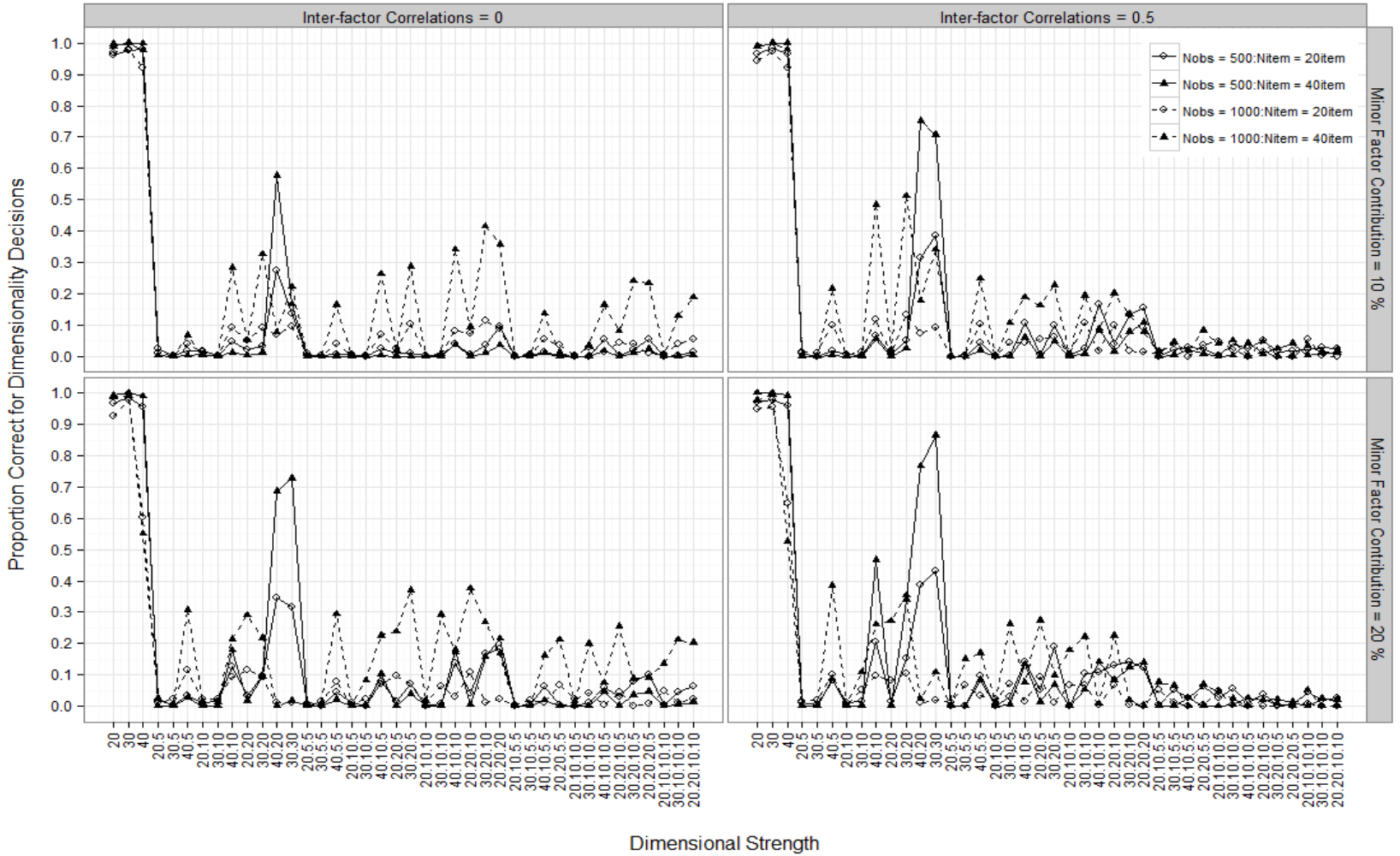


Figure 19. Proportion of Datasets in Which the Quasi-true Number of Dimensions is Correctly Identified by the **NOHARM Approximate Likelihood Ratio Chi-Square Statistic**

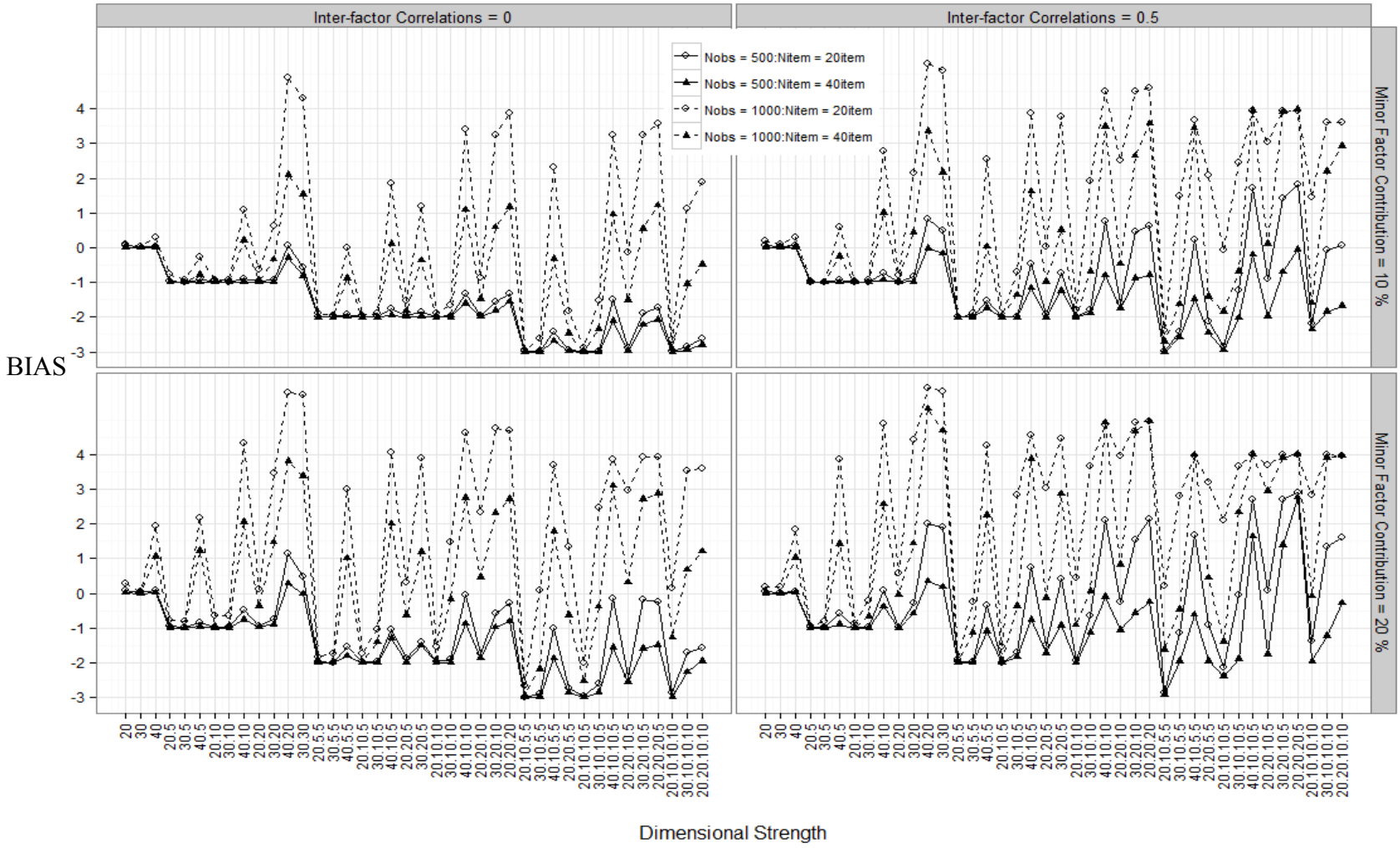


Figure 20. Bias with respect to the Quasi-true Number of Dimensions for the **NOHARM Approximate Likelihood Ratio Chi-Square Statistic**

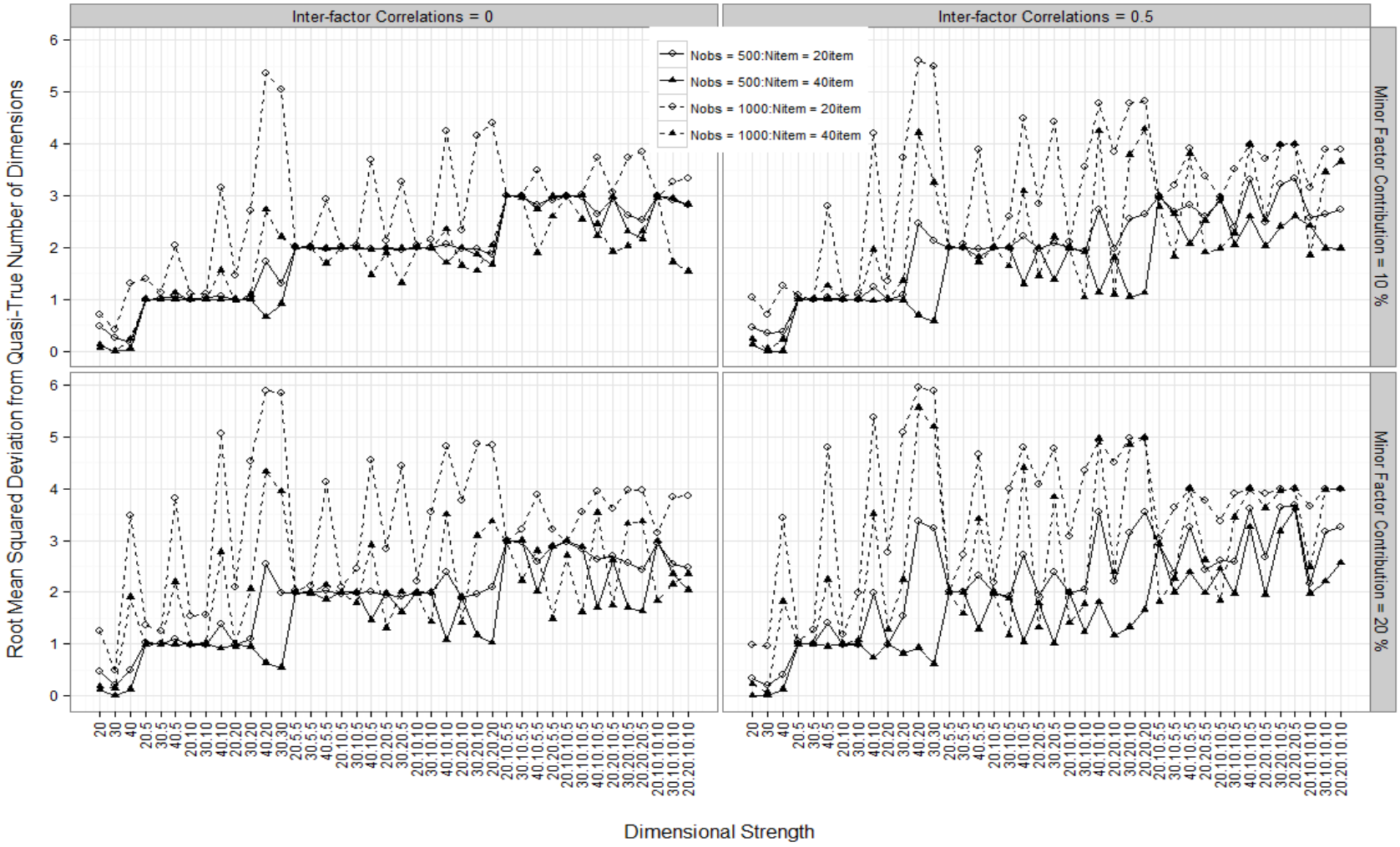


Figure 21. Root Mean Squared Deviation from the Quasi-true Number of Dimensions for the **NOHARM Approximate Likelihood Ratio Chi-Square Statistic**

Figure 20 shows the bias in dimensionality decisions with respect to the quasi-true number of dimensions for the NOHARM ALR chi-square statistic. When the sample size was 1000, the bias was positive and large in most conditions. This was largely due to the replications in which the decision was not reached after fitting the eight-dimensional model, and the suggested number of dimensions was assumed to be eight for those replications. When the sample size was 500, the bias was negative, indicating that the number of dimensions suggested by NOHARM ALR was less than the quasi-true number of dimensions in the generating model.

Table 29 and Table 30 show the proportions of simulated datasets with no-decision after computing the NOHARM mean-adjusted and mean-and-variance adjusted chi-square statistics for the eight-dimensional model. When the total amount of variance accounted for by the major dimensions exceeded 40%, the proportions of datasets in which at least eight dimensions were identified approached 100%, particularly for the sample size of 1000. The proportions of no-decision replications were relatively lower for the mean-and-variance adjusted chi-square statistics than the mean-adjusted chi-square statistics. Figure 22 and Figure 25 show the proportions of datasets in which the quasi-true number of dimensions was correctly identified by the NOHARM mean-adjusted and mean-and-variance adjusted chi-square statistics. The proportions of correct decisions were similar for both statistics across conditions. On average, mean-and-variance adjusted chi-square statistics provided slightly higher rates than the mean-adjusted chi-square statistics in correctly identifying the model with the quasi-true number of dimensions. But, the success rates for both statistics were low in general except for the conditions with one major dimension. Figure 23 and Figure 24 show the bias in decisions with respect to the quasi-true number of dimensions for the NOHARM mean-adjusted and mean-and-variance adjusted chi-square statistics. In most conditions, the bias was positive and large, indicating that both statistics tend to select the models with more dimensions than the quasi-true number of dimensions. In rare conditions in which there were one or two relatively weaker dimensions accounting for 5% of the variance, the bias was negative, and both statistics tended to select models with less dimensions than the quasi-true number of dimensions.

Table 29. *Proportion of Datasets in Which the Dimensionality Decision Cannot be Reached by the NOHARM Mean-Adjusted Chi-square Statistic*

Minor Factors	10 %				20 %			
Number of Items	20		40		20		40	
Sample Size	500	1000	500	1000	500	1000	500	1000
20	0.02	0.09	-	0.01	0.08	0.29	-	0.09
30	0.05	0.14	-	0.02	0.14	0.46	0.02	0.22
40	0.21	0.46	0.03	0.24	0.56	0.94	0.31	0.78
20.5	0.04	0.13	-	0.01	0.11	0.40	0.01	0.17
30.5	0.07	0.22	0.01	0.07	0.24	0.70	0.07	0.47
40.5	0.42	0.80	0.21	0.59	0.87	0.99	0.65	0.87
20.10	0.06	0.20	-	0.05	0.17	0.57	0.04	0.31
30.10	0.15	0.37	0.01	0.20	0.44	0.86	0.26	0.71
40.10	0.74	0.96	0.51	0.82	0.97	1.00	0.86	0.92
20.20	0.23	0.59	0.04	0.28	0.57	0.95	0.31	0.74
30.20	0.67	0.95	0.38	0.77	0.95	1.00	0.81	0.91
40.20	1.00	1.00	0.90	0.94	1.00	1.00	0.94	0.97
30.30	0.98	1.00	0.89	0.96	1.00	1.00	0.96	0.98
20.5.5	0.08	0.19	0.01	0.06	0.18	0.58	0.06	0.36
30.5.5	0.17	0.45	0.03	0.26	0.50	0.88	0.29	0.76
40.5.5	0.76	0.97	0.57	0.81	0.97	1.00	0.87	0.92
20.10.5	0.11	0.30	0.02	0.14	0.33	0.78	0.14	0.57
30.10.5	0.35	0.67	0.14	0.47	0.75	0.98	0.57	0.90
40.10.5	0.91	1.00	0.79	0.93	1.00	1.00	0.91	0.95
20.20.5	0.48	0.86	0.21	0.58	0.83	1.00	0.63	0.88
30.30.5	0.87	1.00	0.71	0.92	1.00	1.00	0.92	0.95
20.10.10	0.22	0.56	0.07	0.31	0.59	0.92	0.33	0.77
30.10.10	0.61	0.88	0.39	0.78	0.92	1.00	0.81	0.94
40.10.10	0.99	1.00	0.89	0.96	1.00	1.00	0.93	0.96
20.20.10	0.71	0.96	0.49	0.78	0.96	1.00	0.81	0.94
30.20.10	0.98	1.00	0.87	0.96	1.00	1.00	0.96	0.98
20.20.20	0.99	1.00	0.89	0.96	1.00	1.00	0.95	0.97
20.10.5.5	0.27	0.59	0.10	0.38	0.58	0.91	0.39	0.81
30.10.5.5	0.64	0.88	0.43	0.75	0.92	1.00	0.83	0.65
40.10.5.5	0.99	1.00	0.91	0.95	1.00	1.00	0.94	0.97
20.2/0.5.5	0.74	0.97	0.50	0.81	0.95	1.00	0.81	0.94
20.10.10.5	0.48	0.77	0.25	0.57	0.78	0.98	0.65	0.89
30.10.10.5	0.81	0.99	0.67	0.91	0.99	1.00	0.93	0.97
40.10.10.5	1.00	1.00	0.95	0.97	1.00	1.00	0.98	0.98
20.20.10.5	0.90	0.99	0.71	0.92	1.00	1.00	0.92	0.96
30.20.10.5	1.00	1.00	0.94	0.98	1.00	1.00	0.99	0.97
20.20.20.5	1.00	1.00	0.94	0.97	1.00	1.00	0.97	0.98
20.10.10.10	0.65	0.89	0.48	0.78	0.93	1.00	0.82	0.96
30.10.10.10	0.95	1.00	0.87	0.96	1.00	1.00	0.97	0.98
20.20.10.10	0.98	1.00	0.88	0.96	1.00	1.00	0.95	0.97

Note. Dashes indicate zero.

Table 30. *Proportion of Datasets in Which the Dimensionality Decision Cannot be Reached by the NOHARM Mean-and-Variance Adjusted Chi-Square Statistic*

Minor Factors	10 %				20 %			
Number of Items	20		40		20		40	
Sample Size	500	1000	500	1000	500	1000	500	1000
20	0.01	0.05	-	-	0.03	0.20	-	0.03
30	0.01	0.08	-	0.01	0.07	0.36	-	0.09
40	0.09	0.35	0.01	0.11	0.38	0.88	0.12	0.67
20.5	0.01	0.07	-	-	0.05	0.29	-	0.07
30.5	0.02	0.14	-	0.02	0.12	0.59	0.01	0.30
40.5	0.27	0.70	0.07	0.42	0.74	0.98	0.44	0.82
20.10	0.03	0.13	-	0.01	0.08	0.45	-	0.16
30.10	0.06	0.28	-	0.08	0.27	0.79	0.05	0.56
40.10	0.59	0.92	0.28	0.73	0.93	1.00	0.74	0.91
20.20	0.12	0.49	-	0.16	0.39	0.90	0.09	0.62
30.20	0.51	0.91	0.16	0.66	0.88	1.00	0.63	0.89
40.20	0.98	1.00	0.81	0.93	1.00	1.00	0.93	0.96
30.30	0.95	1.00	0.78	0.95	1.00	1.00	0.95	0.98
20.5.5	0.03	0.12	-	0.02	0.09	0.49	0.01	0.21
30.5.5	0.08	0.35	-	0.11	0.34	0.80	0.09	0.61
40.5.5	0.63	0.95	0.35	0.73	0.94	1.00	0.75	0.90
20.10.5	0.05	0.24	-	0.05	0.21	0.67	0.03	0.42
30.10.5	0.21	0.59	0.02	0.31	0.59	0.96	0.30	0.84
40.10.5	0.84	0.99	0.63	0.91	0.99	1.00	0.88	0.94
20.20.5	0.32	0.78	0.07	0.44	0.71	1.00	0.38	0.84
30.30.5	0.79	0.99	0.52	0.88	0.99	1.00	0.88	0.95
20.10.10	0.12	0.44	0.01	0.16	0.42	0.88	0.13	0.65
30.10.10	0.45	0.83	0.17	0.63	0.85	1.00	0.63	0.91
40.10.10	0.96	1.00	0.82	0.95	1.00	1.00	0.91	0.96
20.20.10	0.57	0.94	0.26	0.70	0.90	1.00	0.68	0.93
30.20.10	0.94	1.00	0.76	0.95	1.00	1.00	0.95	0.98
20.20.20	0.97	1.00	0.80	0.95	1.00	1.00	0.94	0.97
20.10.5.5	0.17	0.50	0.02	0.23	0.42	0.87	0.16	0.70
30.10.5.5	0.51	0.82	0.23	0.64	0.85	0.99	0.67	0.55
40.10.5.5	0.96	1.00	0.83	0.93	1.00	1.00	0.93	0.97
20.2/0.5.5	0.62	0.94	0.31	0.70	0.89	1.00	0.66	0.92
20.10.10.5	0.34	0.70	0.07	0.41	0.65	0.96	0.42	0.85
30.10.10.5	0.70	0.97	0.46	0.86	0.97	1.00	0.87	0.96
40.10.10.5	1.00	1.00	0.94	0.97	1.00	1.00	0.98	0.98
20.20.10.5	0.81	0.99	0.55	0.88	0.98	1.00	0.87	0.96
30.20.10.5	0.99	1.00	0.91	0.97	1.00	1.00	0.99	0.97
20.20.20.5	1.00	1.00	0.92	0.96	1.00	1.00	0.97	0.98
20.10.10.10	0.54	0.84	0.26	0.68	0.86	1.00	0.66	0.94
30.10.10.10	0.90	1.00	0.74	0.95	1.00	1.00	0.97	0.98
20.20.10.10	0.94	1.00	0.77	0.95	1.00	1.00	0.94	0.97

Note. Dashes indicate zero.

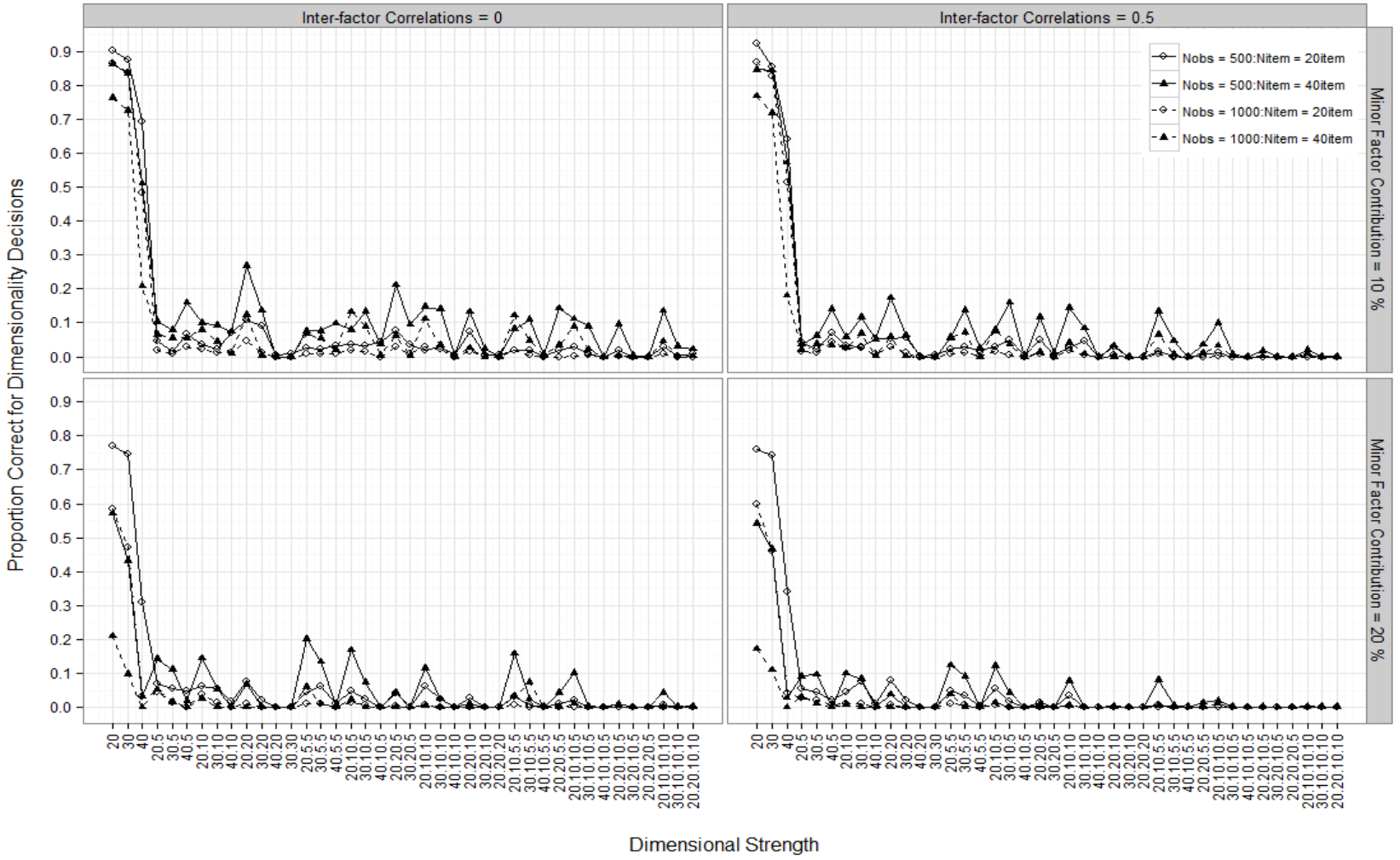


Figure 22. Proportion of Datasets in Which the Quasi-true Number of Dimensions is Correctly Identified by the **NOHARM Mean-Adjusted Chi-Square Statistic**

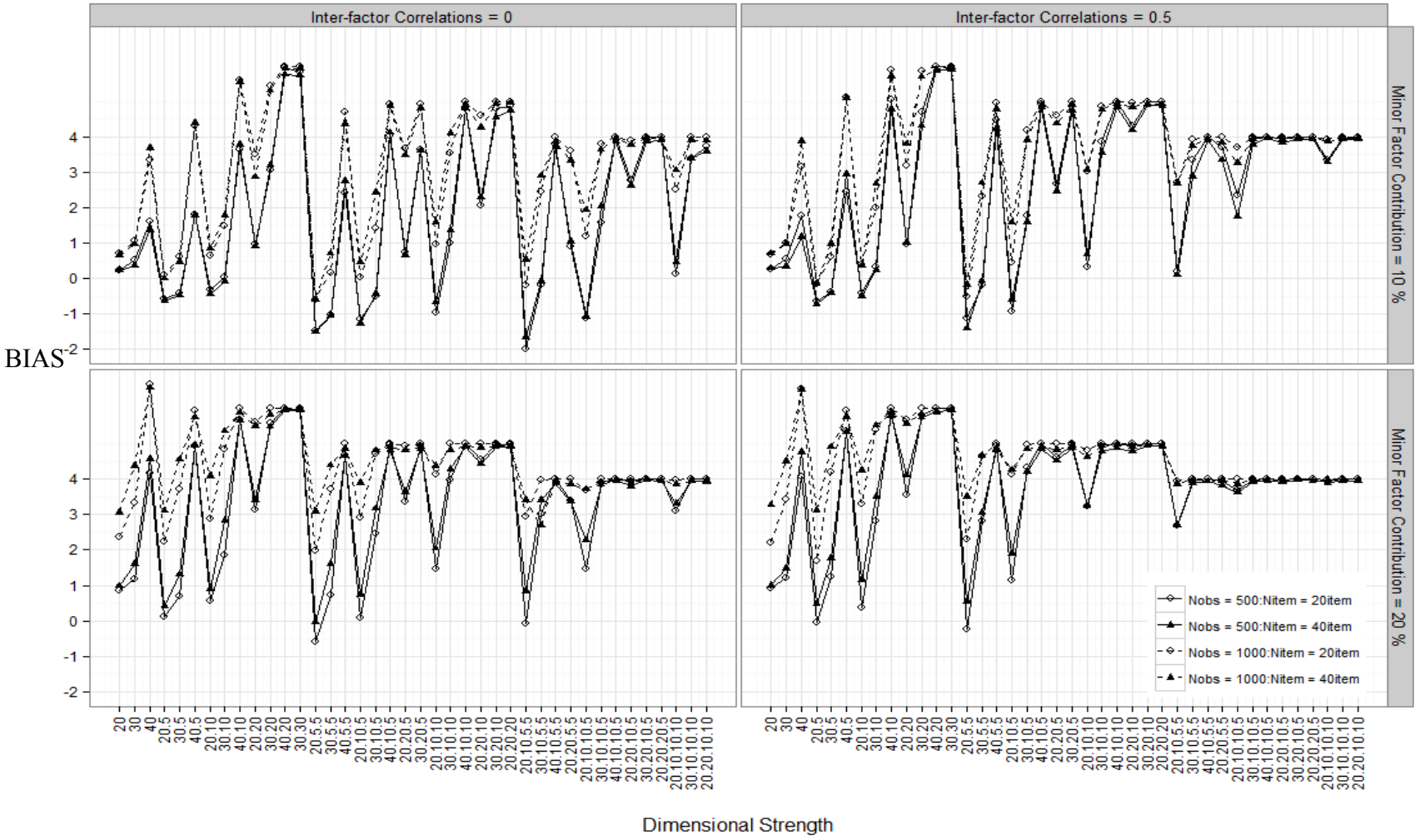


Figure 23. Bias with respect to the Quasi-true Number of Dimensions for the **NOHARM Mean-Adjusted Chi-Square Statistic**

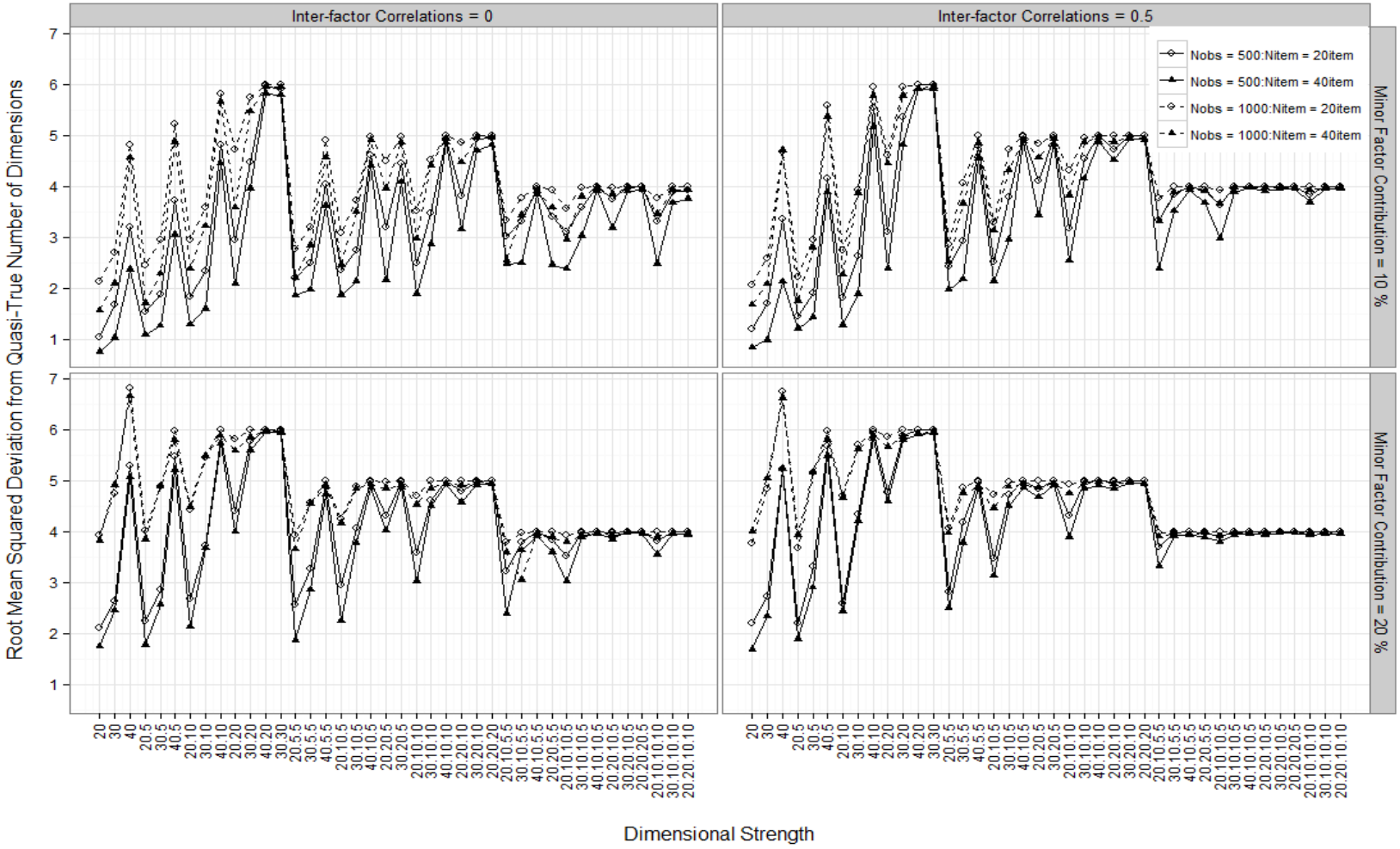


Figure 24. Root Mean Squared Deviation from the Quasi-true Number of Dimensions for the **NOHARM Mean-Adjusted Chi-Square Statistic**

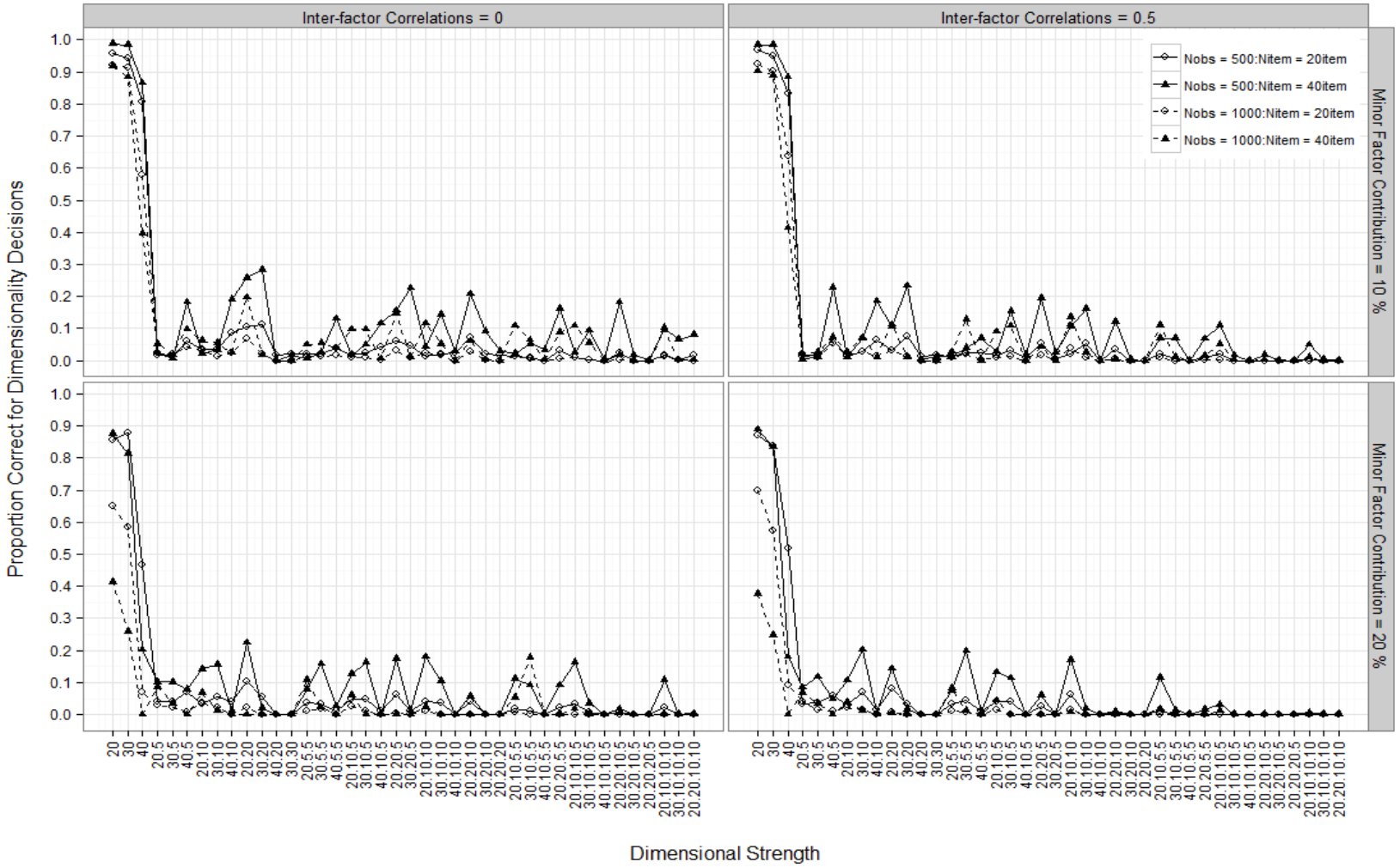


Figure 25. Proportion of Datasets in Which the Quasi-true Number of Dimensions is Correctly Identified by the **NOHARM Mean-and-Variance Adjusted Chi-Square Statistic**

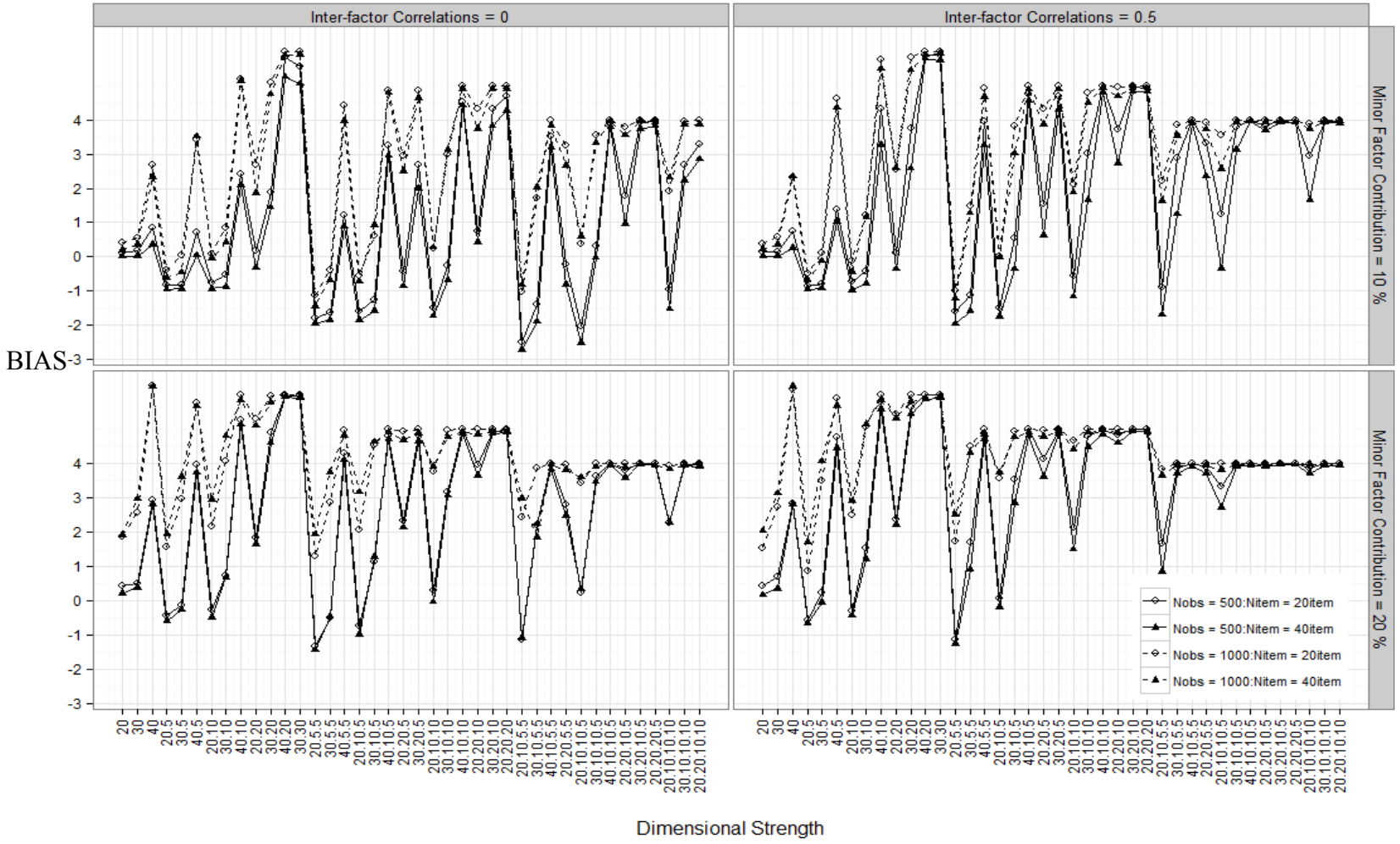


Figure 26. Bias with respect to the Quasi-true Number of Dimensions for the **NOHARM Mean-and-Variance Adjusted Chi-Square Statistic**

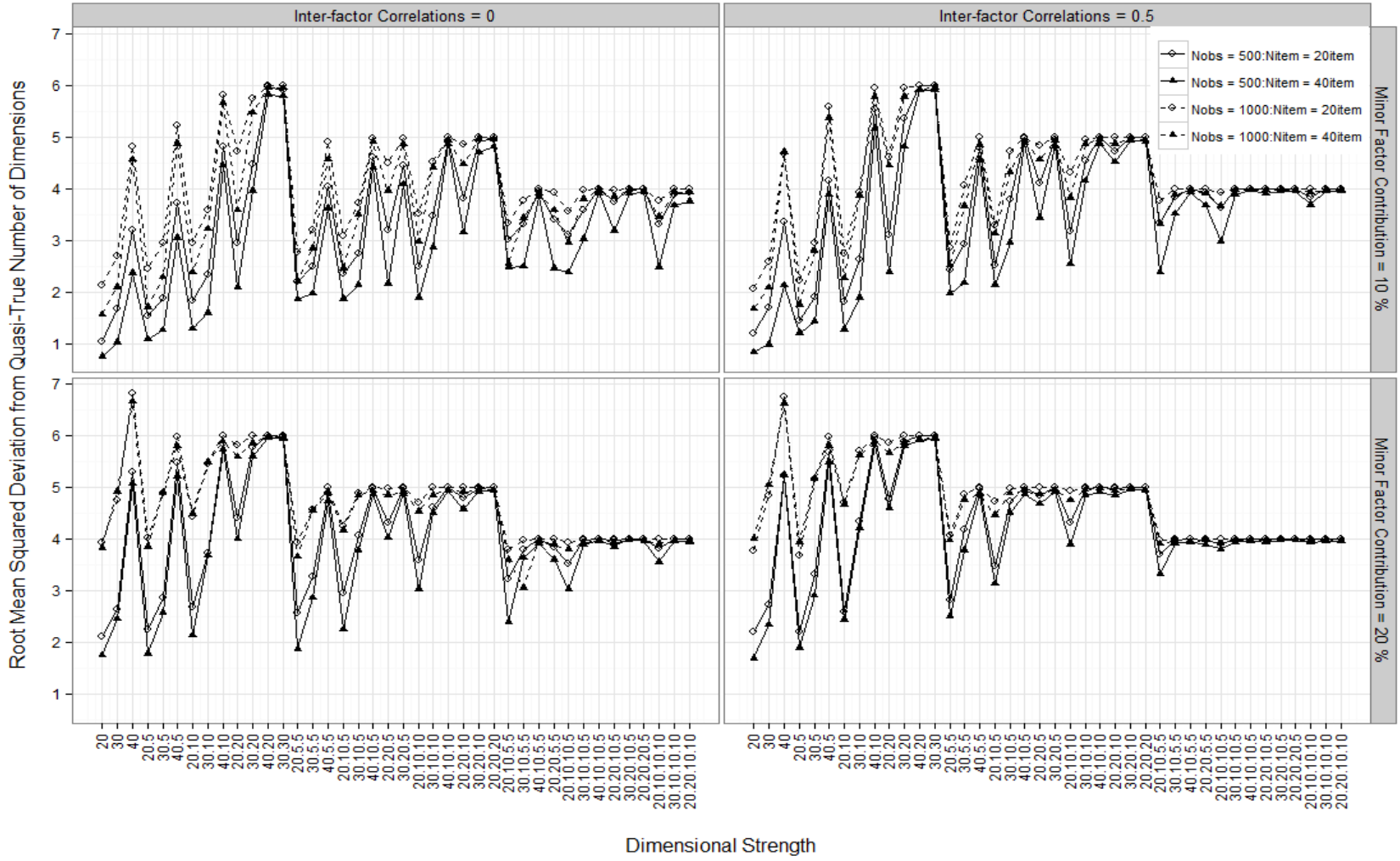


Figure 27. Root Mean Squared Deviation from the Quasi-true Number of Dimensions for the **NOHARM Mean-and-Variance Adjusted Chi-Square Statistic**

Mplus WLS-based Chi-Square Statistics. Table 31 through Table 34 and Figure 28 through Figure 33 show results for mean-adjusted and mean-and-variance adjusted chi-square statistics based on the Mplus WLS estimation. In Table 31 and Table 33, the proportion of replications in which a decision was not reached after fitting the eight-dimensional model was 0% when the number of items was 20, and below 25% when the number of items was 40 for both statistics. The proportion of replications in which both statistics identified at least eight dimensions exceeded 10%, particularly in conditions where the total amount of variance accounted for by major dimensions was more than 60%. These numbers were much lower than the NOHARM-based mean-adjusted and mean-and-variance adjusted chi-square statistics. On the other hand, as reported in Table 32 and Table 34, mean-adjusted and mean-and-variance adjusted chi-square statistics based on the Mplus WLS estimation suffered from convergence issues. The proportion of replications in which a decision was not reached due to convergence problems hit 90% in some conditions, and higher rates were observed particularly when the sample size was 1000, the number of items was 40, and the total variance accounted for by minor factors was 20%. The replications with no-decision due to convergence problems were eliminated from further analysis. For the other replications in which a decision was not reached after fitting the eight-dimensional model, the number of suggested dimensions was assumed to be eight for further computations.

Figure 28 and Figure 31 show the proportions of replications in which the quasi-true number of dimensions was correctly identified by the mean-adjusted and mean-and-variance adjusted chi-square statistics based on the Mplus WLS estimation. Compared to the NOHARM-based chi-square statistics, Mplus WLS-based chi-square statistics were more successful in identifying the quasi-true number of dimensions. In general, the proportions of correct decisions were higher for the 20-item test and varied between 10% and 70% in most conditions. For the conditions with 40 items, the proportions of correct decisions were lower and varied between 0% and 50% in most conditions. On average across all conditions, the mean-and-variance adjusted chi-square statistic performed slightly better than the mean-adjusted chi-square statistic in recovering the quasi-true number of dimensions.

Figure 29 and Figure 32 show the bias in dimensionality decisions with respect to the quasi-true number of dimensions for the Mplus WLS-based mean-adjusted and mean-and-variance adjusted chi-square statistics. When there were 40 items, both statistics showed a positive bias in most conditions, indicating that they tended to identify more dimensions than the quasi-true number of dimensions in the generating model. In some 40-item conditions, when there were one or two weaker dimensions accounting for only 5% of the variance and the total amount of variance accounted for by minor dimensions was 10%, both statistics showed a negative bias in small amounts. When there were 20 items, both statistics showed a negative bias in most conditions and identified fewer than the quasi-true number of dimensions. However, in some 20-item conditions, when there were two or more than two strong dimensions and the total amount of variance accounted for by minor dimensions was 20%, both statistics showed a positive bias.

Mplus FIML-based Chi-Square Difference Test Statistics. The results for the adjusted and unadjusted log-likelihood ratio chi-square difference tests based on the Mplus FIML estimation are presented in Table 35 through Table 38 and in Figure 34 through Figure 39. The models with up to six dimensions were fitted using the Mplus MLR estimator, and chi-square difference tests were computed using adjusted and unadjusted log-likelihoods at each step. In some replications, the chi-square difference test was still significant after fitting the six-dimensional model and a decision was not reached. In some other replications, the decision was not reached due to convergence issues at some point when fitting models with one to six dimensions. Table 35 and Table 37 show the proportions of no-decision replications after successfully fitting models with up to six dimensions. The proportions were 0% for most conditions and at very low rates in other conditions varying between 0% and 8% for the chi-square difference test with adjusted log-likelihoods. However, the proportions were relatively at higher rates, particularly in 40-item conditions, varying between 0% and 36% for the chi-square difference test with unadjusted log-likelihoods. Table 36 and Table 38 show the proportions of no-decision replications due to convergence problems. Similarly, the proportions were relatively at higher rates particularly in the 40-item conditions, varying between 0 % and 67 % for the chi-square difference test with unadjusted log-likelihoods.

Table 31. *Proportion of Datasets in Which the Dimensionality Decision Cannot be Reached by the Mplus WLS Mean-Adjusted Chi-Square Statistic*

Minor Factors	10 %				20 %			
Number of Items	20		40		20		40	
Sample Size	500	1000	500	1000	500	1000	500	1000
20	-	-	-	-	-	-	-	-
30	-	-	-	-	-	-	-	-
40	-	-	-	-	-	-	0.01	-
20.5	-	-	-	-	-	-	-	-
30.5	-	-	-	-	-	-	-	-
40.5	-	-	-	-	-	-	0.02	0.02
20.10	-	-	-	-	-	-	-	-
30.10	-	-	-	-	-	-	-	-
40.10	-	-	0.01	-	-	-	0.07	0.05
20.20	-	-	-	-	-	-	-	-
30.20	-	-	-	-	-	-	0.04	0.03
40.20	-	-	0.06	0.01	-	-	0.18	0.11
30.30	-	-	0.05	0.02	-	-	0.19	0.13
20.5.5	-	-	-	-	-	-	-	-
30.5.5	-	-	-	-	-	-	-	-
40.5.5	-	-	0.01	-	-	-	0.07	0.05
20.10.5	-	-	-	-	-	-	-	-
30.10.5	-	-	-	-	-	-	0.01	0.01
40.10.5	-	-	0.02	0.01	-	-	0.13	0.09
20.20.5	-	-	-	-	-	-	0.01	0.01
30.30.5	-	-	0.02	0.01	-	-	0.12	0.08
20.10.10	-	-	-	-	-	-	-	-
30.10.10	-	-	-	-	-	-	0.03	0.02
40.10.10	-	-	0.06	0.02	-	-	0.17	0.10
20.20.10	-	-	0.01	-	-	-	0.04	0.04
30.20.10	-	-	0.04	0.02	-	-	0.19	0.14
20.20.20	-	-	0.06	0.01	-	-	0.21	0.15
20.10.5.5	-	-	-	-	-	-	-	-
30.10.5.5	-	-	-	-	-	-	0.03	0.03
40.10.5.5	-	-	0.06	0.02	-	-	0.16	0.12
20.20.5.5	-	-	-	-	-	-	0.03	0.03
20.10.10.5	-	-	-	-	-	-	0.01	0.01
30.10.10.5	-	-	0.01	-	-	-	0.11	0.07
40.10.10.5	-	-	0.10	0.03	-	-	0.22	0.13
20.20.10.5	-	-	0.01	0.01	-	-	0.09	0.11
30.20.10.5	-	-	0.10	0.03	-	-	0.22	0.12
20.20.20.5	-	-	0.12	0.05	-	-	0.23	0.12
20.10.10.10	-	-	-	-	-	-	0.04	0.03
30.10.10.10	-	-	0.04	0.02	-	-	0.18	0.13
20.20.10.10	-	-	0.03	0.02	-	-	0.18	0.14

Note. Dashes indicate zero.

Table 32. *Proportion of Datasets in Which the Dimensionality Decision Cannot be Reached Due to Convergence Problems by the Mplus WLS Mean-Adjusted Chi-Square Statistic*

Minor Factors	10 %				20 %			
Number of Items	20		40		20		40	
Sample Size	500	1000	500	1000	500	1000	500	1000
20	0.02	0.01	0.01	-	0.03	0.04	-	0.01
30	0.03	0.01	-	0.01	0.04	0.05	0.01	0.03
40	0.05	0.07	0.04	0.04	0.09	0.13	0.15	0.24
20.5	0.02	0.02	-	-	0.03	0.04	0.01	0.01
30.5	0.03	0.04	0.01	0.02	0.04	0.07	0.03	0.05
40.5	0.08	0.08	0.10	0.16	0.15	0.20	0.35	0.45
20.10	0.03	0.03	-	0.01	0.04	0.06	0.01	0.02
30.10	0.03	0.04	0.02	0.02	0.05	0.10	0.06	0.09
40.10	0.11	0.14	0.23	0.35	0.19	0.33	0.49	0.66
20.20	0.04	0.05	0.02	0.03	0.06	0.10	0.08	0.11
30.20	0.09	0.10	0.13	0.17	0.18	0.22	0.33	0.42
40.20	0.24	0.30	0.57	0.72	0.46	0.55	0.73	0.83
30.30	0.27	0.29	0.55	0.66	0.42	0.54	0.71	0.82
20.5.5	0.03	0.03	0.01	0.02	0.03	0.05	0.02	0.03
30.5.5	0.04	0.04	0.02	0.02	0.07	0.09	0.06	0.09
40.5.5	0.10	0.14	0.30	0.37	0.25	0.31	0.51	0.65
20.10.5	0.04	0.04	0.01	0.02	0.04	0.06	0.03	0.05
30.10.5	0.04	0.07	0.03	0.06	0.11	0.15	0.14	0.21
40.10.5	0.15	0.23	0.41	0.56	0.29	0.46	0.66	0.82
20.20.5	0.04	0.06	0.04	0.05	0.11	0.13	0.16	0.19
30.30.5	0.14	0.16	0.28	0.37	0.28	0.34	0.55	0.64
20.10.10	0.03	0.06	0.02	0.02	0.05	0.09	0.06	0.09
30.10.10	0.08	0.09	0.11	0.12	0.18	0.23	0.34	0.39
40.10.10	0.29	0.40	0.65	0.79	0.49	0.57	0.76	0.87
20.20.10	0.08	0.09	0.11	0.15	0.18	0.21	0.33	0.38
30.20.10	0.24	0.31	0.51	0.63	0.45	0.53	0.69	0.78
20.20.20	0.21	0.27	0.51	0.62	0.42	0.52	0.67	0.78
20.10.5.5	0.04	0.04	0.02	0.03	0.07	0.09	0.07	0.09
30.10.5.5	0.06	0.09	0.09	0.12	0.18	0.20	0.34	0.42
40.10.5.5	0.26	0.37	0.61	0.77	0.48	0.58	0.75	0.86
20.2/0.5.5	0.08	0.08	0.11	0.15	0.16	0.22	0.31	0.36
20.10.10.5	0.05	0.08	0.04	0.04	0.09	0.11	0.13	0.18
30.10.10.5	0.10	0.16	0.25	0.34	0.28	0.36	0.55	0.67
40.10.10.5	0.42	0.49	0.74	0.91	0.61	0.74	0.77	0.87
20.20.10.5	0.11	0.16	0.26	0.32	0.27	0.34	0.52	0.61
30.20.10.5	0.36	0.45	0.70	0.84	0.62	0.74	0.75	0.87
20.20.20.5	0.38	0.44	0.69	0.85	0.59	0.70	0.76	0.88
20.10.10.10	0.07	0.09	0.08	0.12	0.15	0.19	0.26	0.37
30.10.10.10	0.22	0.26	0.49	0.61	0.44	0.53	0.69	0.80
20.20.10.10	0.22	0.26	0.48	0.61	0.42	0.50	0.69	0.77

Note. Dashes indicate zero.

Table 33. *Proportion of Datasets in Which the Dimensionality Decision Cannot be Reached by the Mplus WLS Mean-and-Variance Adjusted Chi-Square Statistic*

Minor Factors	10 %				20 %			
Number of Items	20		40		20		40	
Sample Size	500	1000	500	1000	500	1000	500	1000
20	-	-	-	-	-	-	-	-
30	-	-	-	-	-	-	-	-
40	-	-	-	-	-	-	-	-
20.5	-	-	-	-	-	-	-	-
30.5	-	-	-	-	-	-	-	-
40.5	-	-	-	-	-	-	0.01	0.01
20.10	-	-	-	-	-	-	-	-
30.10	-	-	-	-	-	-	-	-
40.10	-	-	-	-	-	-	0.02	0.02
20.20	-	-	-	-	-	-	-	-
30.20	-	-	-	-	-	-	0.01	0.02
40.20	-	-	0.02	0.01	-	-	0.11	0.09
30.30	-	-	0.01	0.02	-	-	0.13	0.10
20.5.5	-	-	-	-	-	-	-	-
30.5.5	-	-	-	-	-	-	-	-
40.5.5	-	-	-	-	-	-	0.03	0.03
20.10.5	-	-	-	-	-	-	-	-
30.10.5	-	-	-	-	-	-	-	-
40.10.5	-	-	0.01	-	-	-	0.06	0.06
20.20.5	-	-	-	-	-	-	-	-
30.30.5	-	-	-	-	-	-	0.05	0.05
20.10.10	-	-	-	-	-	-	-	-
30.10.10	-	-	-	-	-	-	0.01	0.01
40.10.10	-	-	0.03	0.02	-	-	0.11	0.08
20.20.10	-	-	-	-	-	-	0.01	0.02
30.20.10	-	-	0.02	0.01	-	-	0.11	0.10
20.20.20	-	-	0.02	0.01	-	-	0.14	0.11
20.10.5.5	-	-	-	-	-	-	-	-
30.10.5.5	-	-	-	-	-	-	-	0.01
40.10.5.5	-	-	0.02	0.01	-	-	0.11	0.09
20.2/0.5.5	-	-	-	-	-	-	-	0.01
20.10.10.5	-	-	-	-	-	-	-	-
30.10.10.5	-	-	0.01	-	-	-	0.04	0.04
40.10.10.5	-	-	0.06	0.03	-	-	0.19	0.13
20.20.10.5	-	-	-	-	-	-	0.04	0.06
30.20.10.5	-	-	0.06	0.02	-	-	0.18	0.11
20.20.20.5	-	-	0.07	0.04	-	-	0.20	0.11
20.10.10.10	-	-	-	-	-	-	0.01	0.01
30.10.10.10	-	-	0.01	0.01	-	-	0.11	0.10
20.20.10.10	-	-	0.01	0.01	-	-	0.10	0.10

Note. Dashes indicate zero.

Table 34. *Proportion of Datasets in Which the Dimensionality Decision Cannot be Reached Due to Convergence Problems by the Mplus WLS Mean-and-Variance Adjusted Chi-Square Statistics*

Minor Factors	10 %				20 %			
Number of Items	20		40		20		40	
Sample Size	500	1000	500	1000	500	1000	500	1000
20	0.01	-	0.01	-	0.02	0.03	-	0.01
30	0.02	0.01	-	-	0.03	0.04	0.01	0.02
40	0.03	0.06	0.02	0.03	0.07	0.12	0.08	0.18
20.5	0.01	0.02	-	-	0.02	0.03	-	0.01
30.5	0.02	0.03	-	0.01	0.03	0.06	0.02	0.03
40.5	0.06	0.07	0.04	0.11	0.12	0.18	0.24	0.38
20.10	0.02	0.02	-	0.01	0.03	0.05	0.01	0.01
30.10	0.03	0.03	0.01	0.01	0.03	0.09	0.02	0.06
40.10	0.08	0.12	0.15	0.27	0.16	0.30	0.38	0.60
20.20	0.03	0.05	0.01	0.03	0.05	0.07	0.03	0.08
30.20	0.07	0.08	0.07	0.13	0.14	0.18	0.22	0.35
40.20	0.20	0.27	0.45	0.66	0.39	0.50	0.67	0.82
30.30	0.20	0.26	0.41	0.60	0.36	0.49	0.64	0.80
20.5.5	0.02	0.02	-	0.01	0.02	0.04	0.01	0.02
30.5.5	0.03	0.03	0.01	0.01	0.05	0.08	0.03	0.06
40.5.5	0.08	0.12	0.18	0.29	0.19	0.27	0.40	0.60
20.10.5	0.03	0.03	-	0.01	0.03	0.05	0.01	0.03
30.10.5	0.03	0.05	0.02	0.03	0.08	0.13	0.08	0.14
40.10.5	0.12	0.20	0.28	0.48	0.24	0.41	0.56	0.79
20.20.5	0.03	0.05	0.02	0.04	0.08	0.12	0.09	0.14
30.30.5	0.09	0.14	0.17	0.31	0.22	0.29	0.44	0.59
20.10.10	0.02	0.05	0.01	0.01	0.04	0.07	0.03	0.06
30.10.10	0.05	0.08	0.05	0.08	0.14	0.20	0.22	0.33
40.10.10	0.23	0.36	0.51	0.72	0.40	0.53	0.71	0.86
20.20.10	0.06	0.07	0.05	0.11	0.13	0.19	0.20	0.31
30.20.10	0.18	0.28	0.37	0.54	0.37	0.49	0.62	0.76
20.20.20	0.16	0.25	0.39	0.56	0.34	0.48	0.62	0.76
20.10.5.5	0.03	0.04	0.01	0.01	0.04	0.08	0.03	0.06
30.10.5.5	0.05	0.08	0.05	0.07	0.14	0.17	0.21	0.36
40.10.5.5	0.19	0.32	0.47	0.70	0.41	0.54	0.70	0.84
20.2/0.5.5	0.06	0.07	0.06	0.10	0.12	0.18	0.19	0.29
20.10.10.5	0.03	0.07	0.01	0.02	0.07	0.10	0.05	0.12
30.10.10.5	0.07	0.14	0.13	0.24	0.23	0.32	0.45	0.61
40.10.10.5	0.32	0.44	0.65	0.88	0.53	0.70	0.75	0.87
20.20.10.5	0.08	0.13	0.15	0.24	0.21	0.30	0.40	0.57
30.20.10.5	0.28	0.40	0.58	0.80	0.53	0.68	0.73	0.87
20.20.20.5	0.30	0.40	0.57	0.81	0.50	0.66	0.75	0.88
20.10.10.10	0.05	0.07	0.03	0.07	0.10	0.16	0.16	0.31
30.10.10.10	0.17	0.23	0.34	0.53	0.37	0.48	0.62	0.77
20.20.10.10	0.17	0.22	0.34	0.52	0.34	0.46	0.60	0.76

Note. Dashes indicate zero.

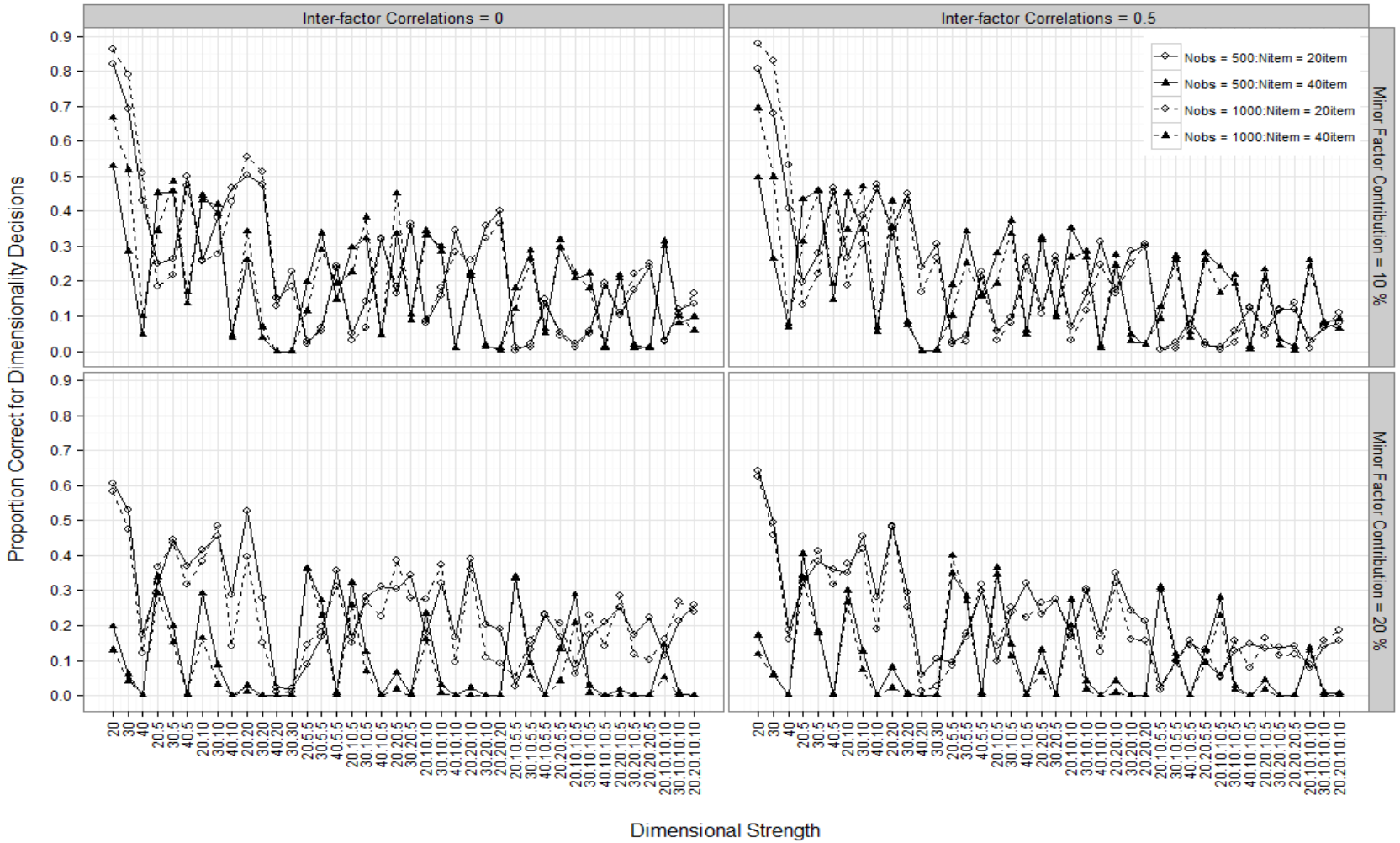


Figure 28. Proportion of Datasets in Which the Quasi-true Number of Dimensions is Correctly Identified by the **Mplus WLS Mean-Adjusted Chi-Square Statistic**

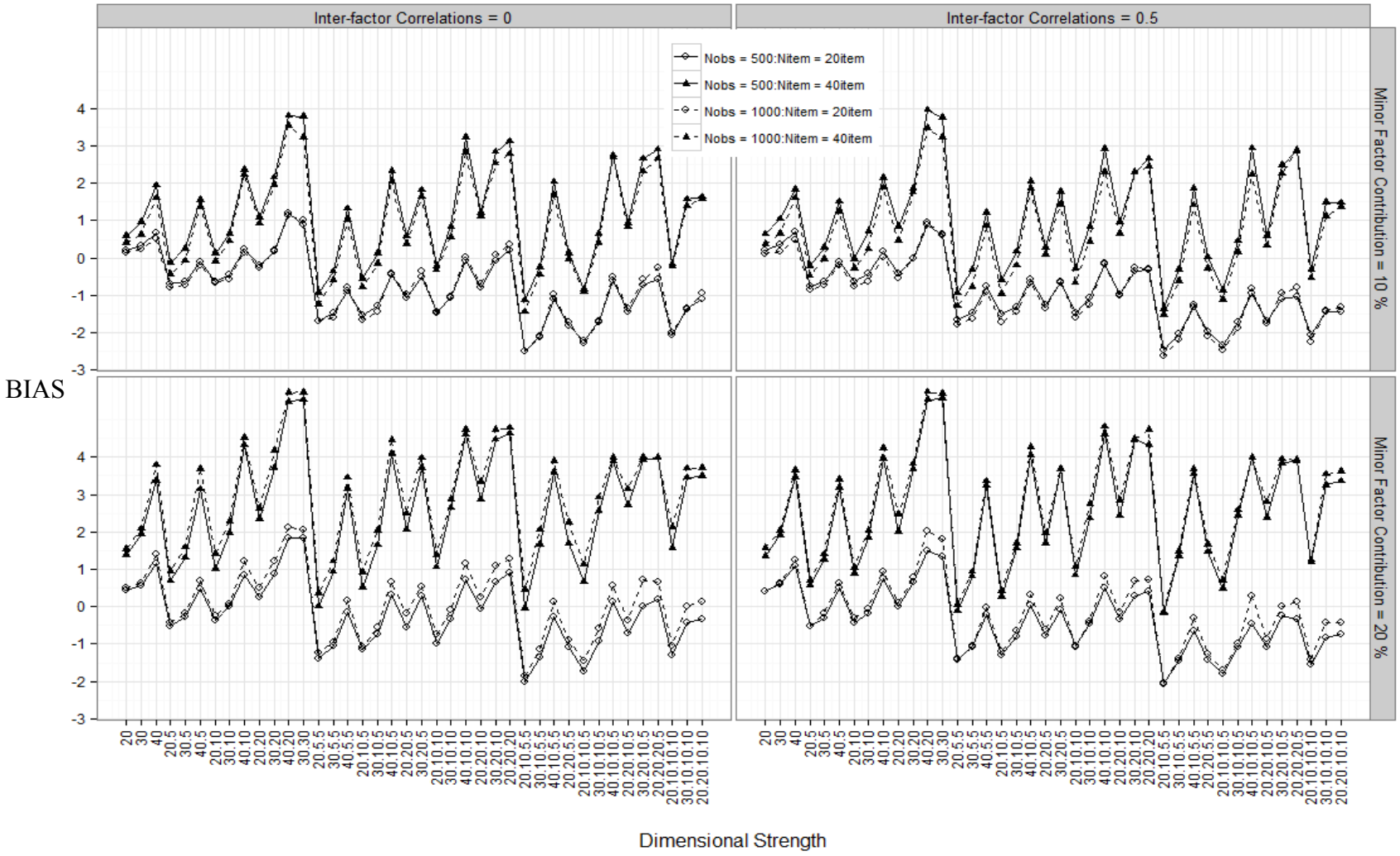


Figure 29. Bias with respect to the Quasi-true Number of Dimensions for the **Mplus WLS Mean-Adjusted Chi-Square Statistic**

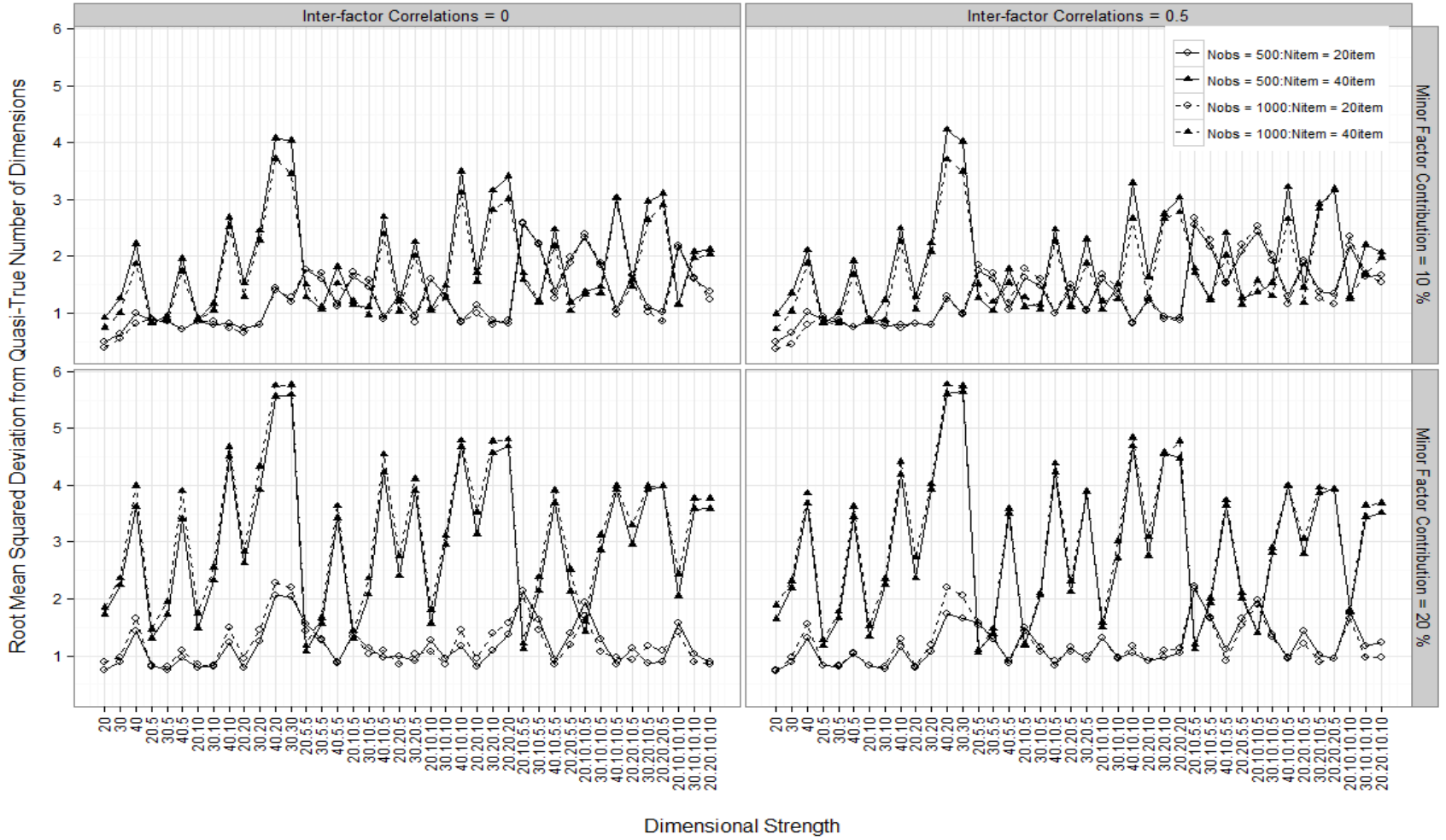


Figure 30. Root Mean Squared Deviation from the Quasi-true Number of Dimensions for the **Mplus WLS Mean-Adjusted Chi-Square Statistic**

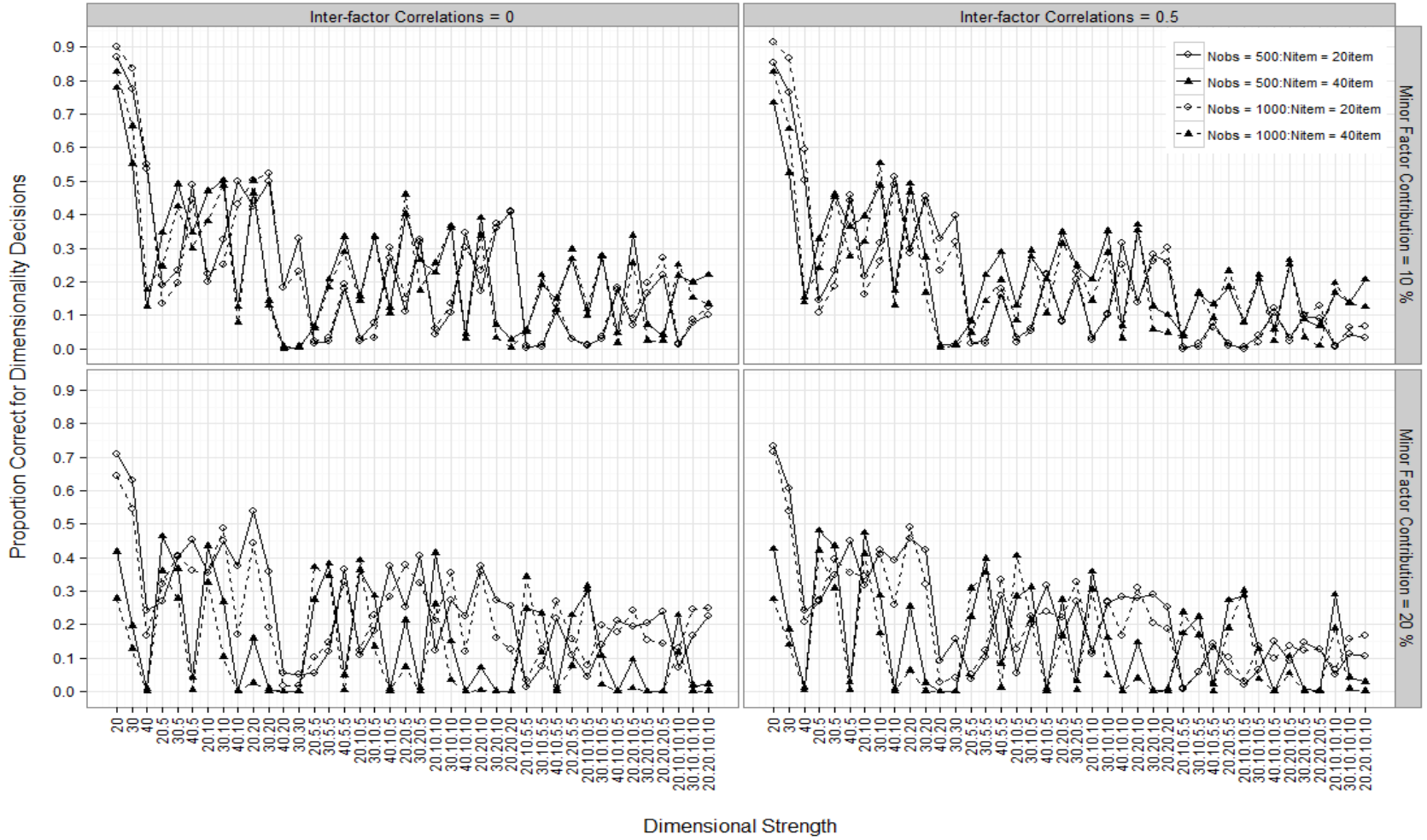


Figure 31. Proportion of Datasets in Which the Quasi-true Number of Dimensions is Correctly Identified by the **Mplus WLS Mean-and-Variance Adjusted Chi-Square Statistic**

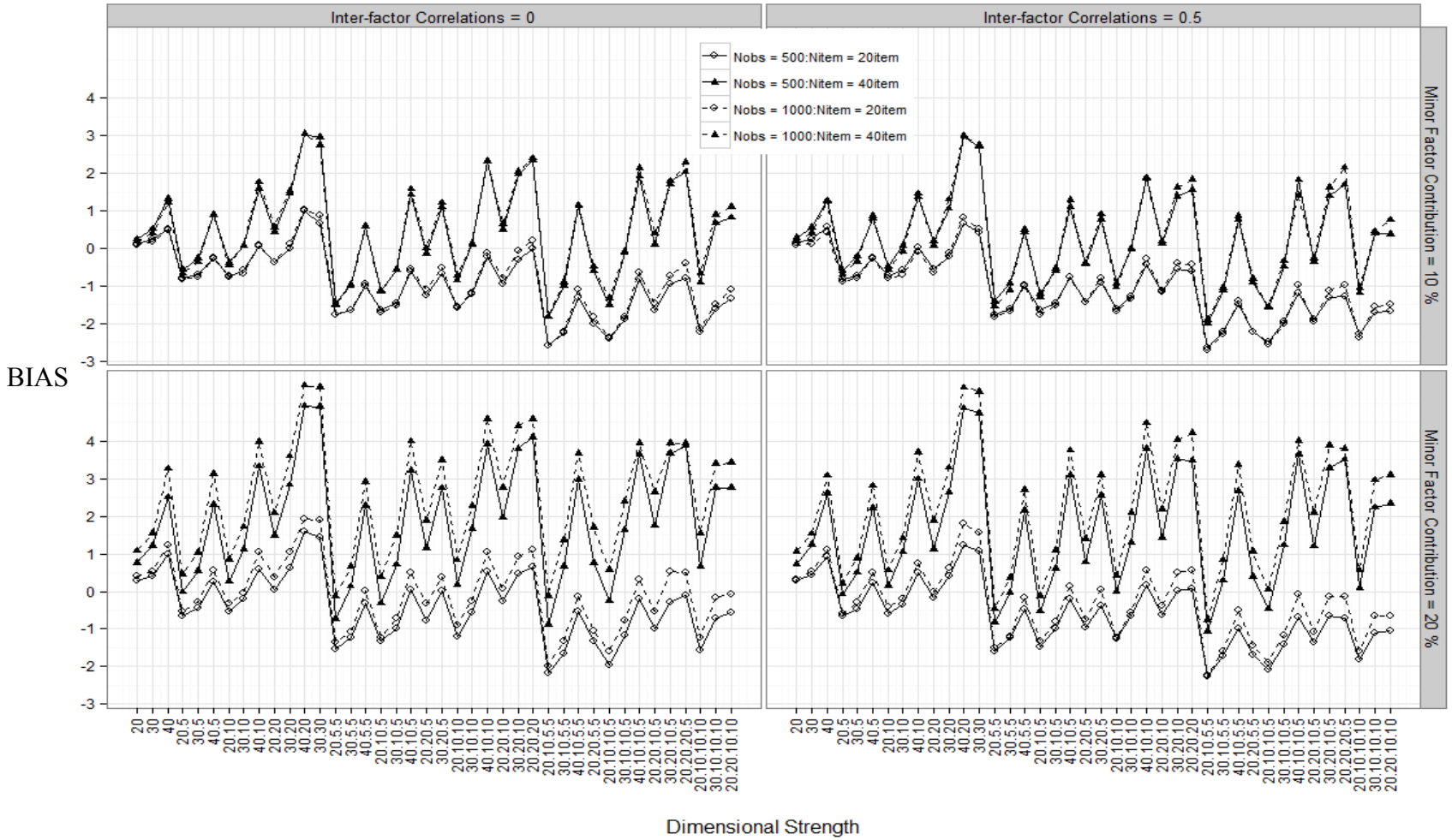


Figure 32. Bias with respect to the Quasi-true Number of Dimensions for the **Mplus WLS Mean-and-Variance Adjusted Chi-Square Statistic**

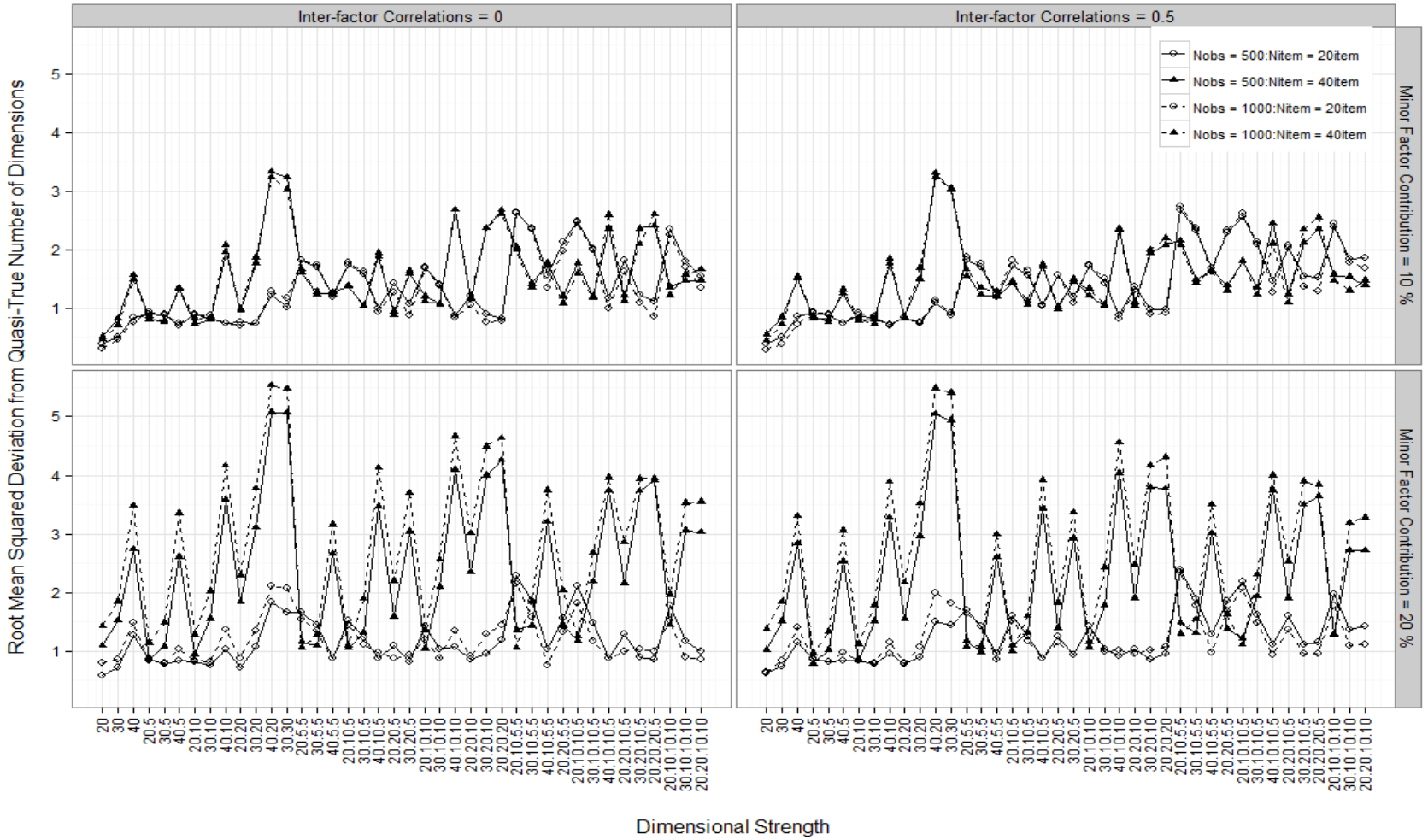


Figure 33. Root Mean Squared Deviation from the Quasi-true Number of Dimensions for the **Mplus WLS Mean-and-Variance Adjusted Chi-Square Statistic**

For the chi-square difference test with the adjusted log-likelihoods, the proportions varied between 0% and 60% and were lower than the unadjusted chi-square difference test on average. The replications with no-decision due to convergence problems were eliminated from further analysis. For the other replications in which a decision was not reached after fitting the six-dimensional model, the number of suggested dimensions was assumed to be six for further computations.

Figure 34 and Figure 37 show the proportions of replications in which the quasi-true number of dimensions were correctly identified for the chi-square difference tests with the adjusted and unadjusted log-likelihood values. When there was only one major dimension in the generating model, chi-square difference tests with adjusted log-likelihoods were more successful with proportion corrects around 90%, while chi-square difference test with the unadjusted log-likelihoods identified one major dimension for 30% to 50% of the replications. When there were two major dimensions in the generating model, both statistics were successful in identifying the quasi-true number of dimensions for the condition in which the number of items was 40. In most of the conditions with two major dimensions, the proportions of correct decisions varied between 40% and 90% when the number of items was 40 and between 20% and 40% when the number of items was 20. When there were more than two major dimensions, the proportions of correct decisions were very low and less than 10% in most conditions while rarely exceeding 30% for both adjusted and unadjusted chi-square statistics.

Figure 35 and Figure 38 show the bias with respect to the quasi-true number of dimensions for the chi-square difference tests with both adjusted and unadjusted log-likelihood values. The chi-square difference test with the unadjusted log-likelihood values showed positive bias for all conditions when there was only one major dimension, positive bias for the 40-item conditions and zero or slightly negative bias for the 20-item conditions when there were two major dimensions, positive bias in most 40-item conditions and negative bias in all 20-item conditions when there were three major dimensions, and negative bias in almost all conditions when there were four major dimensions.

Table 35. *Proportion of Datasets in Which the Dimensionality Decision Cannot be Reached by the Mplus MLR Unadjusted Likelihood Ratio Chi-Square Difference Test*

Minor Factors	10 %				20 %			
Number of Items	20		40		20		40	
Sample Size	500	1000	500	1000	500	1000	500	1000
20	-	-	0.09	0.19	0.01	0.01	0.12	0.36
30	-	0.03	0.28	0.17	0.02	0.02	0.23	0.07
40	0.01	0.01	0.04	0.01	0.01	0.03	0.04	0.01
20.5	0.01	0.03	0.04	0.08	0.01	0.02	0.07	0.06
30.5	0.01	0.02	0.09	0.02	0.03	0.01	0.10	0.03
40.5	0.01	-	0.02	-	0.01	0.01	0.02	0.01
20.10	0.01	0.01	0.08	0.13	0.02	-	0.06	0.06
30.10	0.01	-	-	-	-	0.02	-	-
40.10	-	-	-	-	-	-	-	-
20.20	0.02	0.01	0.10	0.02	0.01	0.03	0.05	0.03
30.20	0.01	0.01	0.01	-	0.04	-	-	-
40.20	-	-	-	0.01	-	0.01	-	0.01
30.30	-	-	0.01	-	-	0.01	-	0.01
20.5.5	-	0.01	0.17	0.02	-	0.02	0.16	0.01
30.5.5	-	0.01	0.04	-	0.01	-	0.02	-
40.5.5	-	-	0.01	0.01	-	-	0.01	-
20.10.5	0.02	0.03	0.11	0.06	0.02	0.04	0.07	0.03
30.10.5	0.01	0.01	0.01	-	-	0.01	-	-
40.10.5	-	-	-	-	0.01	-	0.01	-
20.20.5	0.01	0.01	0.01	0.01	0.02	0.01	0.01	0.01
30.30.5	-	0.01	0.13	0.08	0.01	0.02	0.15	0.10
20.10.10	-	0.01	0.09	0.04	0.01	0.03	0.08	0.05
30.10.10	-	0.01	0.07	0.04	0.01	0.02	0.07	0.09
40.10.10	0.01	-	0.09	0.08	-	0.02	0.15	0.19
20.20.10	-	-	0.02	0.04	0.01	0.02	0.03	0.16
30.20.10	-	0.01	0.10	0.11	0.01	0.02	0.19	0.18
20.20.20	0.01	0.01	0.15	0.10	-	0.03	0.19	0.21
20.10.5.5	0.02	0.03	0.16	0.15	0.03	0.07	0.23	0.19
30.10.5.5	0.02	0.01	0.13	0.14	0.03	0.02	0.14	0.14
40.10.5.5	0.02	0.01	0.07	0.08	-	0.01	0.12	0.16
20.2/0.5.5	0.01	-	0.05	0.11	0.01	0.01	0.11	0.21
20.10.10.5	0.01	0.02	0.13	0.16	0.02	0.05	0.24	0.13
30.10.10.5	0.01	0.01	0.09	0.07	0.01	0.02	0.19	0.17
40.10.10.5	-	0.01	0.10	0.07	0.02	0.03	0.13	0.16
20.20.10.5	0.02	0.02	0.13	0.09	0.03	0.01	0.17	0.17
30.20.10.5	0.01	-	0.07	0.10	0.03	0.02	0.21	0.14
20.20.20.5	-	0.01	0.15	0.10	0.02	0.03	0.12	0.19
20.10.10.10	0.01	0.02	0.20	0.12	0.02	0.04	0.16	0.17
30.10.10.10	0.01	0.01	0.11	0.08	-	0.01	0.16	0.18
20.20.10.10	0.01	0.02	0.12	0.11	0.02	0.03	0.16	0.16

Note. Dashes indicate zero.

Table 36. *Proportion of Datasets in Which the Dimensionality Decision Cannot be Reached Due to Convergence Problems by the Mplus MLR Unadjusted Likelihood Ratio Chi-Square Difference Test*

Minor Factors	10 %				20 %			
Number of Items	20		40		20		40	
Sample Size	500	1000	500	1000	500	1000	500	1000
20	0.18	0.05	0.23	0.31	0.13	0.10	0.43	0.30
30	-	0.02	0.05	0.09	0.03	0.06	0.12	0.08
40	0.04	0.02	0.06	0.07	0.05	0.07	0.07	0.08
20.5	-	0.01	0.43	0.30	0.01	0.10	0.40	0.26
30.5	0.01	0.03	0.09	0.05	0.05	0.07	0.09	0.07
40.5	0.02	0.05	0.06	0.11	0.09	0.10	0.10	0.09
20.10	0.11	0.04	0.18	0.22	0.19	0.12	0.28	0.23
30.10	0.15	0.18	0.28	0.33	0.28	0.29	0.37	0.53
40.10	0.22	0.19	0.23	0.17	0.24	0.17	0.31	0.11
20.20	0.03	0.05	0.08	0.11	0.07	0.12	0.07	0.12
30.20	0.08	0.12	0.37	0.37	0.11	0.32	0.37	0.47
40.20	0.29	0.24	0.30	0.25	0.32	0.36	0.38	0.24
30.30	0.24	0.24	0.40	0.29	0.34	0.28	0.47	0.38
20.5.5	0.17	0.04	0.05	0.27	0.12	0.08	0.09	0.39
30.5.5	0.16	0.19	0.07	0.08	0.30	0.29	0.10	0.19
40.5.5	0.03	0.04	0.11	0.13	0.09	0.07	0.13	0.13
20.10.5	0.01	0.04	0.07	0.05	0.04	0.07	0.12	0.09
30.10.5	0.03	0.22	0.41	0.62	0.15	0.24	0.38	0.61
40.10.5	0.24	0.23	0.31	0.28	0.30	0.29	0.38	0.32
20.20.5	0.18	0.18	0.33	0.17	0.18	0.21	0.27	0.21
30.30.5	0.23	0.22	0.46	0.41	0.29	0.35	0.47	0.50
20.10.10	0.23	0.17	0.54	0.66	0.22	0.34	0.60	0.67
30.10.10	0.20	0.21	0.46	0.51	0.29	0.27	0.42	0.45
40.10.10	0.25	0.07	0.10	0.14	0.28	0.16	0.10	0.09
20.20.10	0.20	0.12	0.31	0.29	0.24	0.22	0.40	0.25
30.20.10	0.06	0.14	0.10	0.16	0.09	0.16	0.15	0.19
20.20.20	0.08	0.08	0.09	0.09	0.11	0.11	0.12	0.16
20.10.5.5	0.05	0.03	0.17	0.29	0.05	0.10	0.16	0.28
30.10.5.5	0.05	0.06	0.16	0.23	0.06	0.10	0.17	0.18
40.10.5.5	0.08	0.06	0.09	0.10	0.09	0.12	0.12	0.12
20.2/0.5.5	0.15	0.19	0.33	0.17	0.27	0.24	0.36	0.22
20.10.10.5	0.05	0.05	0.15	0.31	0.06	0.12	0.17	0.32
30.10.10.5	0.06	0.06	0.12	0.18	0.09	0.11	0.10	0.15
40.10.10.5	0.09	0.06	0.14	0.11	0.14	0.12	0.13	0.11
20.20.10.5	0.05	0.09	0.14	0.23	0.06	0.16	0.16	0.22
30.20.10.5	0.07	0.09	0.14	0.16	0.15	0.13	0.13	0.18
20.20.20.5	0.08	0.13	0.09	0.18	0.11	0.17	0.11	0.15
20.10.10.10	0.05	0.08	0.19	0.25	0.09	0.13	0.12	0.22
30.10.10.10	0.06	0.07	0.10	0.16	0.11	0.12	0.13	0.16
20.20.10.10	0.08	0.09	0.08	0.23	0.12	0.15	0.13	0.20

Note. Dashes indicate zero.

Table 37. *Proportion of Datasets in Which the Dimensionality Decision Cannot be Reached by the Mplus MLR Adjusted Likelihood Ratio Chi-Square Difference Test*

Minor Factors	10 %				20 %			
Number of Items	20		40		20		40	
Sample Size	500	1000	500	1000	500	1000	500	1000
20	-	-	-	-	-	-	-	-
30	-	-	-	-	-	-	-	-
40	-	-	-	-	-	-	-	-
20.5	-	-	-	-	-	-	-	-
30.5	-	-	-	-	-	-	-	-
40.5	-	-	-	-	-	-	-	-
20.10	-	-	-	-	-	-	-	-
30.10	-	-	-	-	-	-	-	-
40.10	-	-	-	-	-	-	-	-
20.20	-	-	-	-	-	-	-	-
30.20	-	-	-	-	-	-	-	-
40.20	-	-	-	-	-	-	-	-
30.30	-	-	-	-	-	-	-	-
20.5.5	-	-	-	-	-	-	-	-
30.5.5	-	-	-	-	-	-	-	-
40.5.5	-	-	-	-	-	-	-	-
20.10.5	-	-	-	-	-	-	-	-
30.10.5	-	-	-	-	-	-	-	-
40.10.5	-	-	-	-	-	-	-	-
20.20.5	-	-	-	-	-	-	-	-
30.30.5	-	-	0.01	0.01	-	-	-	0.02
20.10.10	-	-	-	-	-	0.01	-	0.01
30.10.10	-	-	-	-	-	-	-	0.01
40.10.10	-	-	0.01	0.03	-	0.01	0.01	0.08
20.20.10	-	-	-	-	-	-	-	0.03
30.20.10	-	0.01	0.01	0.01	-	0.01	0.04	0.05
20.20.20	0.01	-	0.01	-	0.01	0.01	0.02	0.03
20.10.5.5	-	-	-	0.01	-	0.01	-	0.01
30.10.5.5	-	-	-	-	0.01	-	0.01	0.01
40.10.5.5	-	-	-	0.02	0.01	0.01	0.01	0.05
20.2/0.5.5	-	-	-	-	-	-	-	0.01
20.10.10.5	-	-	-	0.01	-	0.01	0.01	0.03
30.10.10.5	-	-	0.01	0.01	-	0.01	0.01	0.02
40.10.10.5	0.01	0.01	-	0.03	-	-	0.01	0.06
20.20.10.5	-	-	0.01	0.01	-	0.01	0.02	0.06
30.20.10.5	0.01	-	-	0.02	0.01	0.01	0.02	0.05
20.20.20.5	-	0.03	-	0.01	0.01	0.01	0.01	0.06
20.10.10.10	-	-	0.01	0.02	-	0.01	-	0.02
30.10.10.10	-	-	-	0.01	-	-	-	0.04
20.20.10.10	-	0.01	-	0.03	0.01	0.02	0.01	0.05

Note. Dashes indicate zero.

Table 38. *Proportion of Datasets in Which the Dimensionality Decision Cannot be Reached Due to Convergence Problems by the Mplus MLR Adjusted Likelihood Ratio Chi-Square Difference Test*

Minor Factors	10 %				20 %			
Number of Items	20		40		20		40	
Sample Size	500	1000	500	1000	500	1000	500	1000
20	0.06	0.01	0.02	-	0.05	0.04	0.01	0.05
30	-	-	0.01	0.02	0.01	-	0.03	0.03
40	0.01	-	0.03	0.04	0.01	0.02	0.04	0.06
20.5	-	-	0.02	0.03	0.01	0.04	0.06	0.08
30.5	0.01	0.01	0.02	0.04	0.01	0.01	0.02	0.07
40.5	0.01	0.01	0.04	0.10	0.05	0.03	0.07	0.09
20.10	0.05	0.01	0.02	0.10	0.08	0.04	0.04	0.17
30.10	0.03	0.05	0.14	0.31	0.07	0.12	0.28	0.51
40.10	0.10	0.10	0.13	0.15	0.13	0.10	0.19	0.10
20.20	-	0.02	0.04	0.08	0.01	0.01	0.03	0.12
30.20	0.03	0.07	0.26	0.34	0.05	0.18	0.30	0.45
40.20	0.13	0.18	0.21	0.19	0.22	0.25	0.26	0.21
30.30	0.13	0.18	0.24	0.21	0.16	0.17	0.30	0.31
20.5.5	0.06	0.01	0.01	0.16	0.04	0.01	0.02	0.34
30.5.5	0.06	0.08	0.03	0.08	0.07	0.17	0.07	0.19
40.5.5	0.01	0.01	0.08	0.10	0.03	0.04	0.07	0.12
20.10.5	-	0.02	0.02	0.03	0.01	0.01	0.04	0.07
30.10.5	0.01	0.12	0.33	0.60	0.08	0.11	0.34	0.60
40.10.5	0.14	0.12	0.16	0.22	0.12	0.18	0.24	0.25
20.20.5	0.08	0.09	0.27	0.15	0.07	0.09	0.20	0.22
30.30.5	0.09	0.11	0.11	0.34	0.13	0.16	0.26	0.43
20.10.10	0.10	0.05	0.25	0.35	0.06	0.14	0.35	0.46
30.10.10	0.05	0.12	0.25	0.38	0.13	0.16	0.29	0.38
40.10.10	0.11	0.05	0.06	0.08	0.13	0.09	0.06	0.06
20.20.10	0.10	0.04	0.16	0.15	0.10	0.13	0.21	0.22
30.20.10	0.04	0.08	0.05	0.13	0.05	0.12	0.11	0.16
20.20.20	0.03	0.01	0.06	0.08	0.05	0.08	0.07	0.15
20.10.5.5	0.01	0.01	0.07	0.17	0.02	0.04	0.12	0.21
30.10.5.5	0.02	0.03	0.08	0.16	-	0.03	0.15	0.15
40.10.5.5	0.02	0.03	0.03	0.06	0.05	0.08	0.09	0.11
20.2/0.5.5	0.06	0.06	0.14	0.15	0.14	0.12	0.18	0.21
20.10.10.5	0.02	0.02	0.09	0.19	0.01	0.05	0.12	0.24
30.10.10.5	0.04	0.03	0.09	0.13	0.03	0.03	0.07	0.12
40.10.10.5	0.07	0.05	0.05	0.09	0.06	0.08	0.07	0.10
20.20.10.5	0.01	0.03	0.10	0.22	0.02	0.09	0.10	0.18
30.20.10.5	0.03	0.06	0.07	0.12	0.09	0.06	0.12	0.14
20.20.20.5	0.04	0.07	0.04	0.13	0.06	0.10	0.09	0.14
20.10.10.10	0.02	0.02	0.10	0.17	0.03	0.07	0.07	0.17
30.10.10.10	0.02	0.03	0.06	0.13	0.03	0.06	0.07	0.14
20.20.10.10	0.06	0.03	0.06	0.20	0.09	0.09	0.10	0.17

Note. Dashes indicate zero.

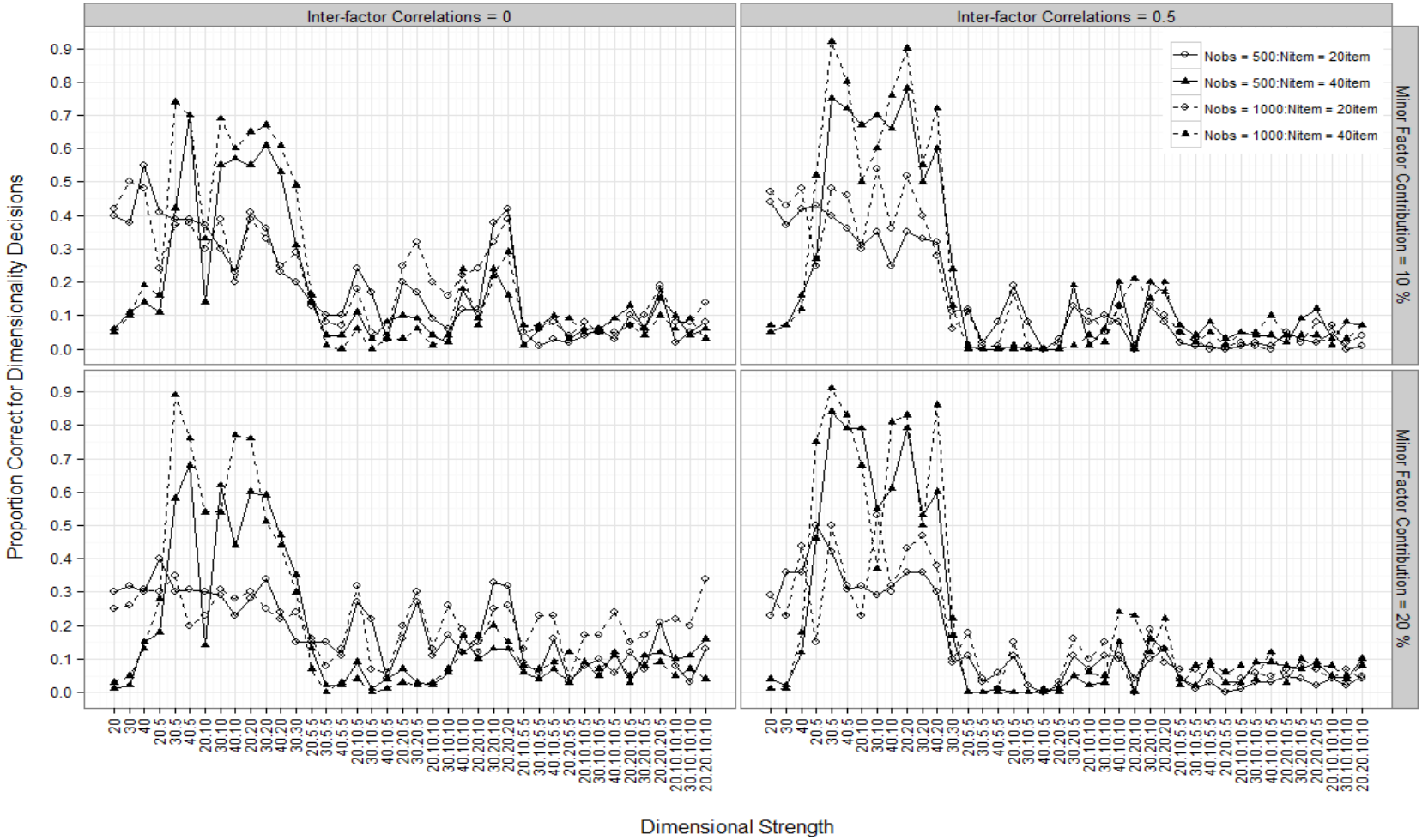


Figure 34. Proportion of Datasets in Which the Quasi-true Number of Dimensions is Correctly Identified by the **Mplus MLR Unadjusted Likelihood Ratio Chi-Square Difference Test**

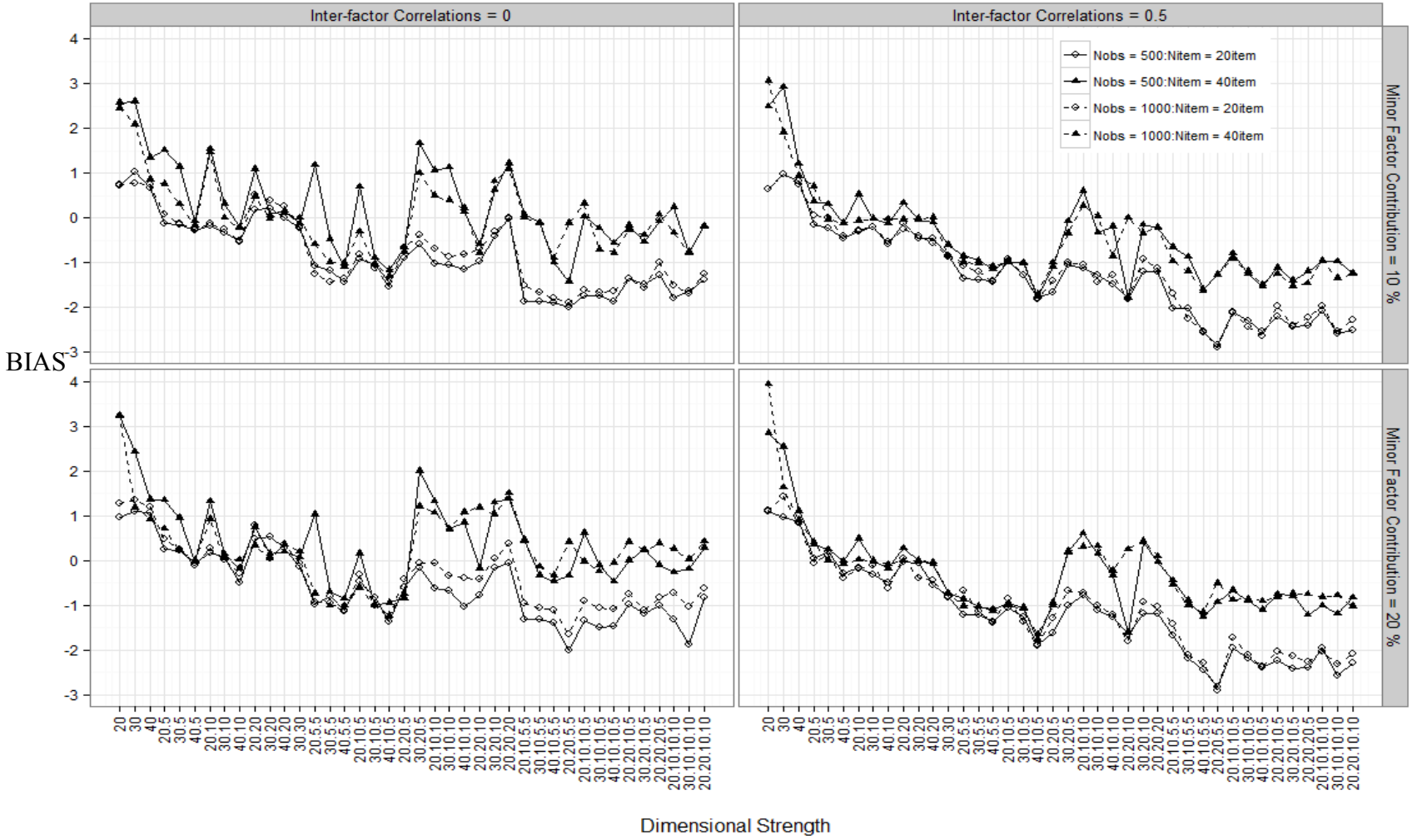


Figure 35. Bias with respect to the Quasi-true Number of Dimensions for the **Mplus MLR Unadjusted Likelihood Ratio Chi-Square Difference Test**

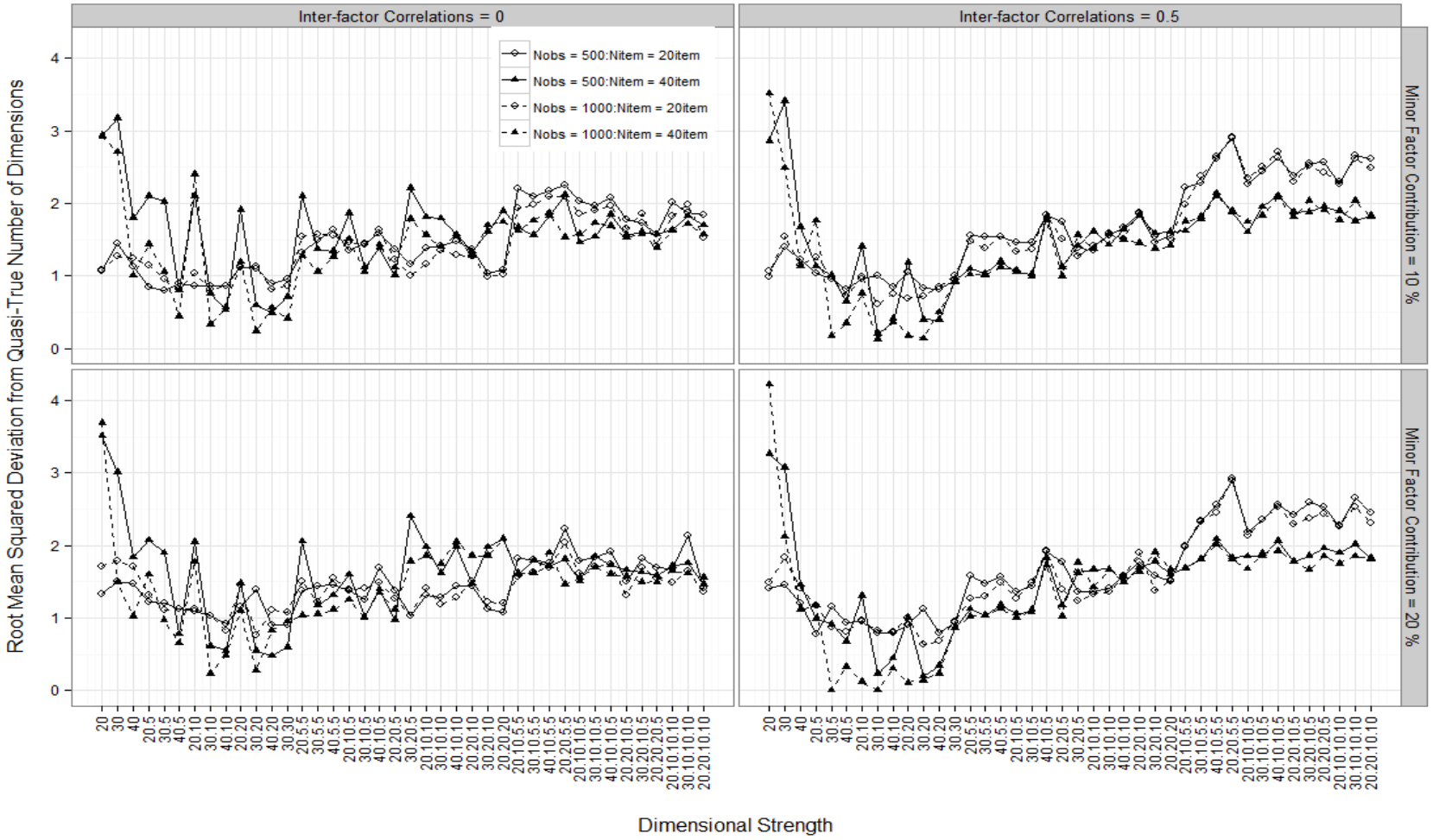


Figure 36. Root Mean Squared Deviation from the Quasi-true Number of Dimensions for the **Mplus MLR Unadjusted Likelihood Ratio Chi-Square Difference Test**

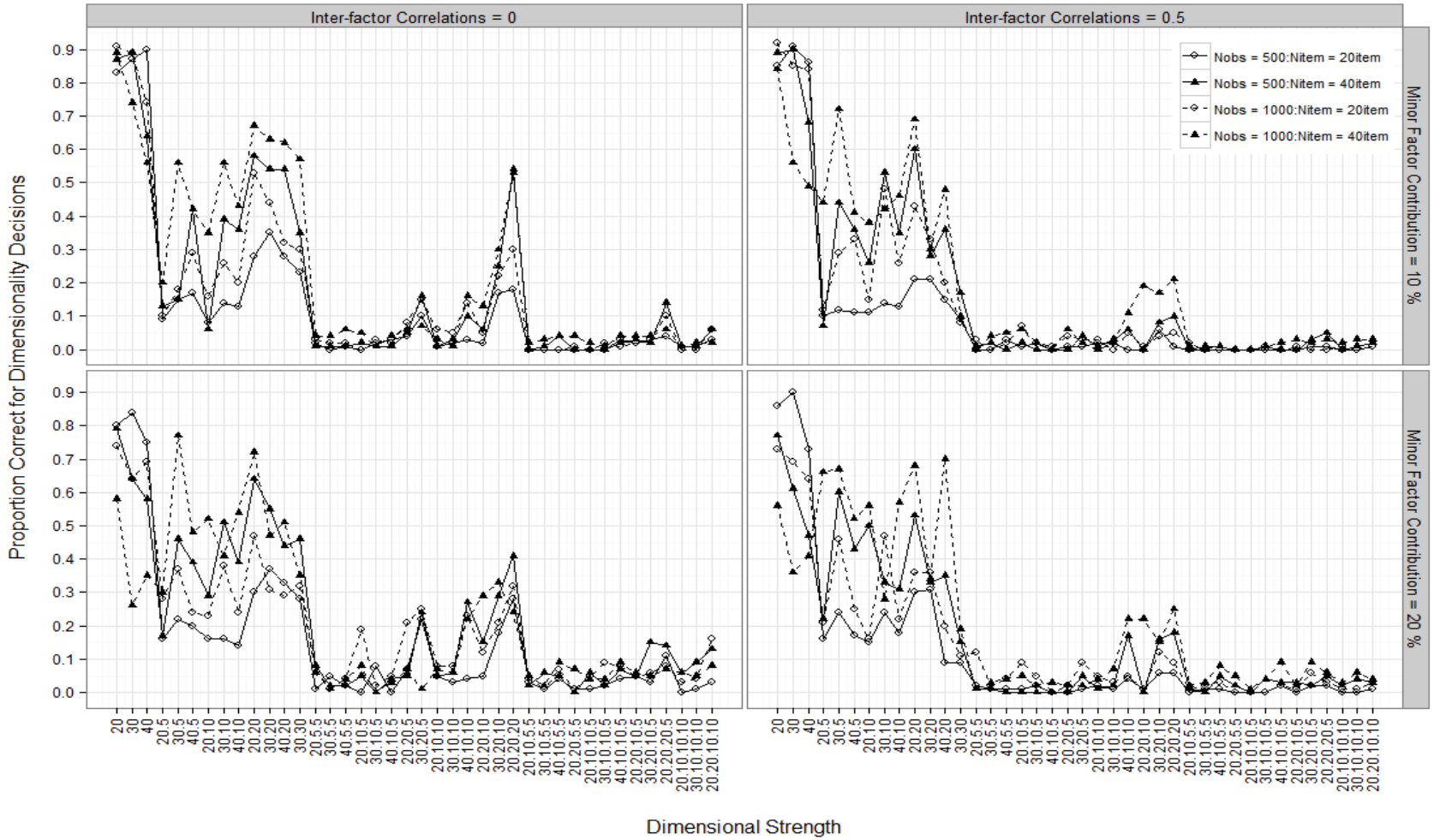


Figure 37. Proportion of Datasets in Which the Quasi-true Number of Dimensions is Correctly Identified by the **Mplus MLR Adjusted Likelihood Ratio Chi-Square Difference Test**

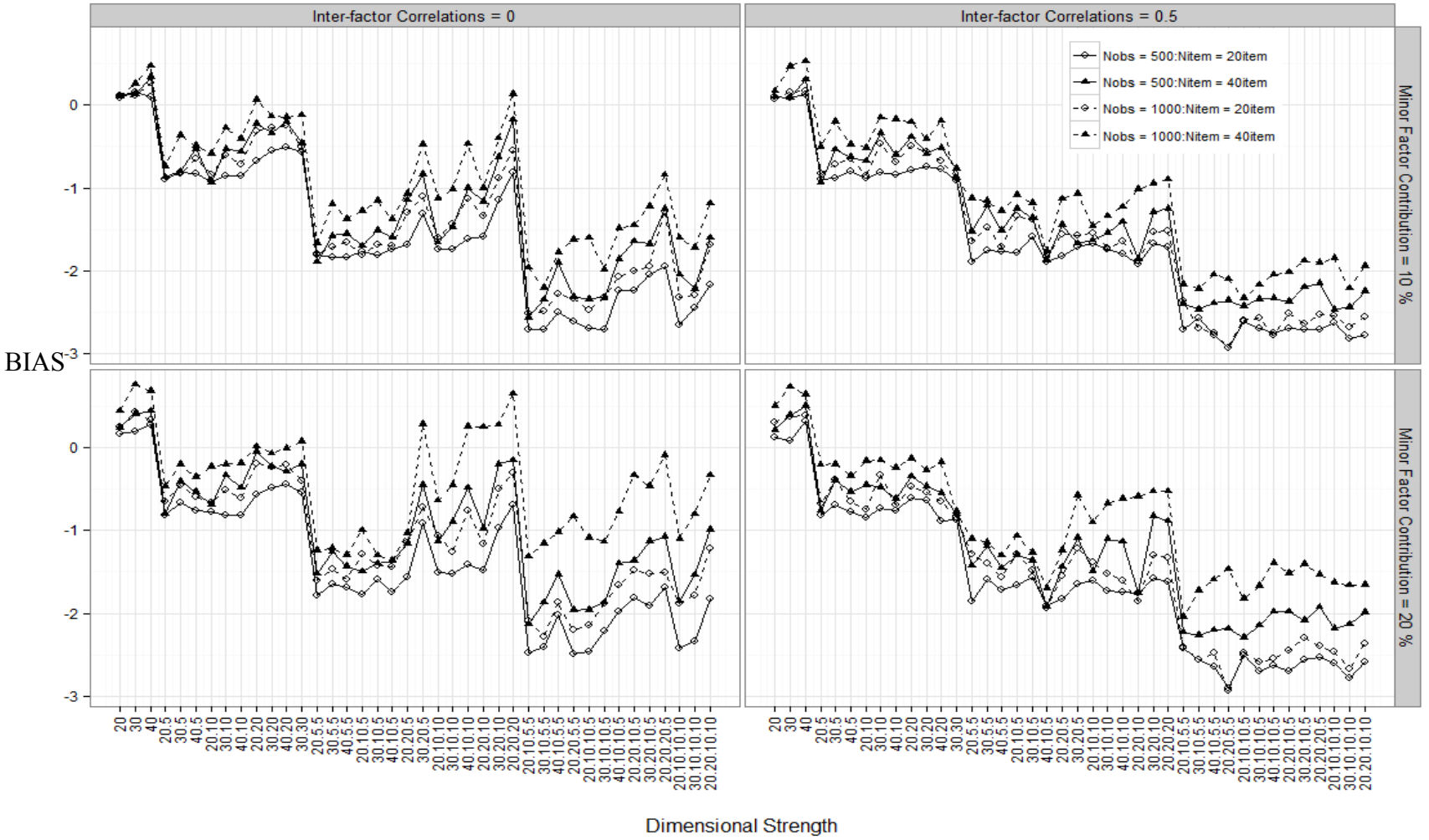


Figure 38. Bias with respect to the Quasi-true Number of Dimensions for the **Mplus MLR Adjusted Likelihood Ratio Chi-Square Difference Test**

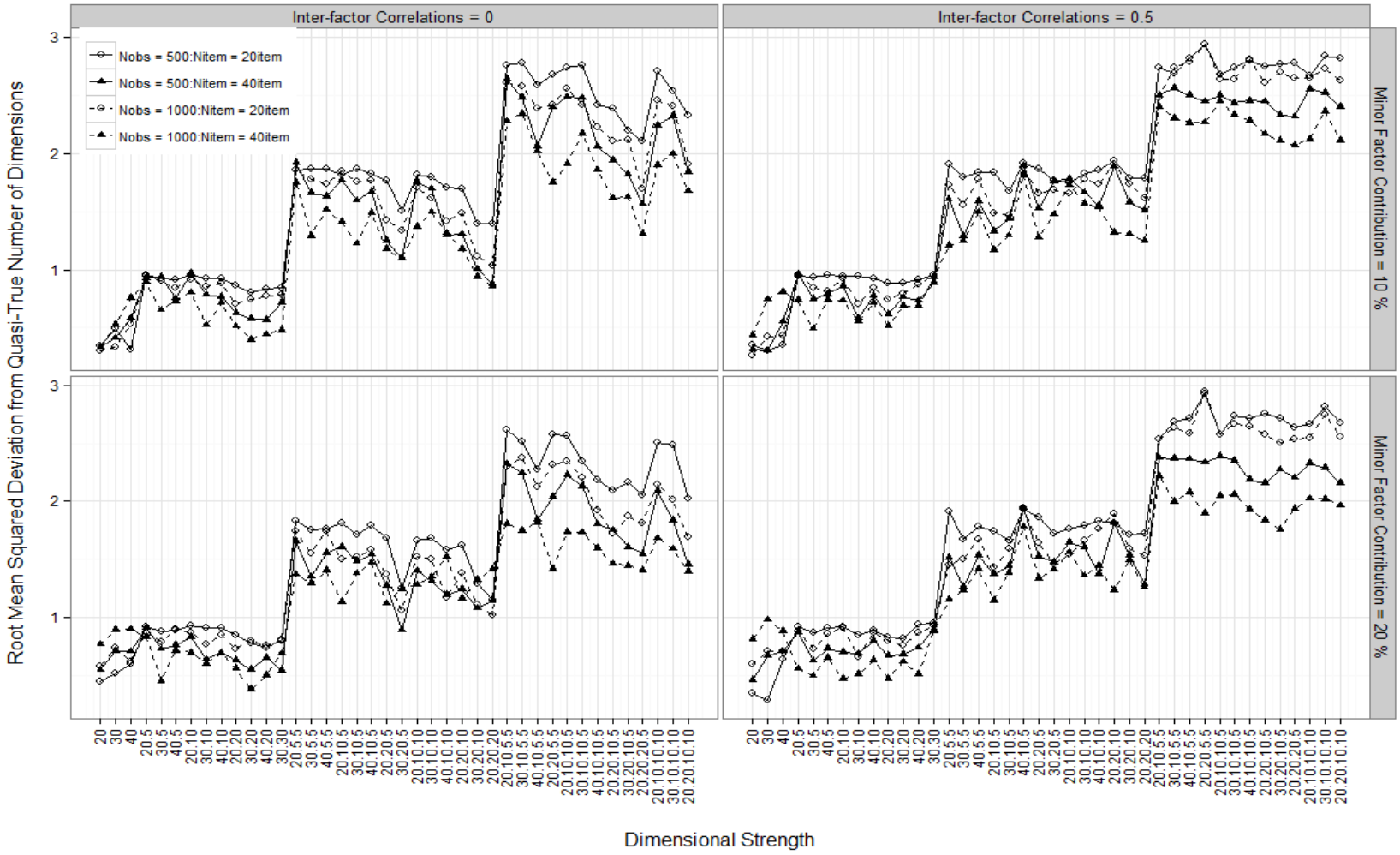


Figure 39. Root Mean Squared Deviation from the Quasi-true Number of Dimensions for the **Mplus MLR Adjusted Likelihood Ratio Chi-Square Difference Test**

The chi-square difference test with the adjusted log-likelihood values showed negative bias in almost all conditions, except some conditions in which there was one major dimension or when there were three strong major dimensions in the generating model. The magnitude of negative bias increased as the number of major dimensions in the generating model increased, and this was due to the tendency to select one- or two-dimensional models, particularly for the 20-item conditions.

Root Mean Square Error Approximation. The results for the RMSEA index obtained from Mplus WLS mean-adjusted and mean-and-variance adjusted chi-square statistics were very similar to each other. Therefore, the results for the RMSEA index obtained from Mplus WLS mean-adjusted chi-square statistics are only reported and shown in Figure 40, Figure 41, and Figure 42. A decision was reached using the RMSEA index for all replications in all conditions.

Figure 40 shows that the RMSEA index correctly identified the quasi-true number of dimensions in almost all conditions when there was only one major dimension and in some rare occasions when there were two or three very strong dimensions in the generating model. In most of the conditions, the RMSEA index never correctly identified the quasi-true number of dimensions. Figure 41 indicates that the bias with respect to the quasi-true number of dimensions was all negative in all conditions when the decision was incorrect. In most conditions, the RMSEA index tended to select one- or two-dimensional models.

The results are based on the standard cut-off value of 0.05 suggested in the literature for the RMSEA index. The simplest model with the lower bound of the confidence interval for the RMSEA index being smaller than 0.05 was selected for any replication. The results showed the cut-off value of 0.05 performed poorly for identifying the major dimensions in the generating model and tended to select one-dimensional models in most occasions. As a follow-up analysis, the current study also tried to identify whether there may be some other optimal cut-off values that maximized the accuracy of decisions with respect to the quasi-true number of dimensions selected. For demonstration purposes, a specific condition from the simulation study was randomly picked. In this condition, there were three major dimensions accounting for 30%, 10%,

and 5% of the total variance and 50 minor factors accounting for 20% of the variance; the inter-factor correlations were 0.5; the sample size was 1000; and the number of items was 40. For a simulated dataset in this specific condition, the RMSEA index has to be larger than a cut-off value for the two-dimensional solution and smaller than the cut-off value for the three-dimensional solution in order to correctly identify the quasi-true number of dimensions. As shown in Figure 43, if we plot the RMSEA indices from the two-dimensional solutions against the RMSEA indices from the three-dimensional solutions for 500 replications in this specific condition, the points are all lined up above a 45 degree line, because the RMSEA index from a two-dimensional solution is always expected to be higher than the RMSEA index from a three-dimensional solution for any replication. Because the RMSEA indices from the two-dimensional solutions were all below 0.05 in this condition, the proportion of selecting a three-dimensional model was zero when the standard cut-off value of 0.05 was used. If we chose a different cut-off value such as 0.025, then the three-dimensional solution would be selected for the replications at the upper-left area in Figure 43. If we could find such a value to maximize the number of replications at the upper-left corner, then it would be the optimal cut-off value that maximizes the accuracy of decisions with respect to the quasi-true number of dimensions for this specific condition. Table 39 shows the proportion of replications with correctly identified quasi-true numbers of dimensions across various cut-off values for this specific condition. The cut-off value of 0.029 is the optimal value in this case, and the three major dimensions can be correctly identified for about 31% of the simulated datasets in this condition.

An optimal cut-off value is found for each simulation condition and reported in Figure 44. In general, the optimal cut-off value for the RMSEA index was always smaller than 0.05 and varied across conditions. First, the optimal cut-off value varied for different sample sizes and numbers of items. As the sample size increased or number of items decreased, a smaller cut-off value optimized the decisions. Second, as there are more strong major dimensions in the generating model, a smaller cut-off value was found optimal. Third, as the variance accounted for by minor factors increased, a larger cut-off value was required to optimize the decisions. For most conditions, an optimal cut-off

value was found between 0.005 and 0.030 when the minor factors accounted for 10% and between 0.005 and 0.040 when the minor factors accounted for 20% of the variance.

Figure 45 shows the maximized proportions of correct decisions with respect to the number of major dimensions at the corresponding optimal cut-off RMSEA value for each condition. As seen in Figure 45, even choosing an optimal cut-off RMSEA value does not guarantee a high rate of correct decisions in finding the quasi-true number of dimensions. When there was one major dimension, the proportion corrects were all 100%, because the optimal value was selected as the maximum RMSEA value given across 500 replications for those conditions. When there were two major dimensions, the maximized proportion corrects varied between 30% and 80% with higher rates observed when there were 40 items. When there were three major dimensions, the maximized proportion corrects varied between 20% and 60% with higher rates observed when there were 40 items and no correlations among the major dimensions. When there were four major dimensions, the maximized proportion corrects varied between 10% and 40%.

Table 39. *The Proportion of Replications in Which the Quasi-True Number of Dimensions is Correctly Identified for Various Cut-off Values of the RMSEA Index*

Cut-off Value	Proportion Correct in Selecting Quasi-True Number of Dimensions
0.023	0.002
0.024	0.002
0.025	0.012
0.026	0.033
0.027	0.079
0.028	0.130
0.029	**0.306**
0.030	0.219
0.031	0.198
0.032	0.180
0.033	0.138
0.034	0.089
0.035	0.043
0.036	0.017

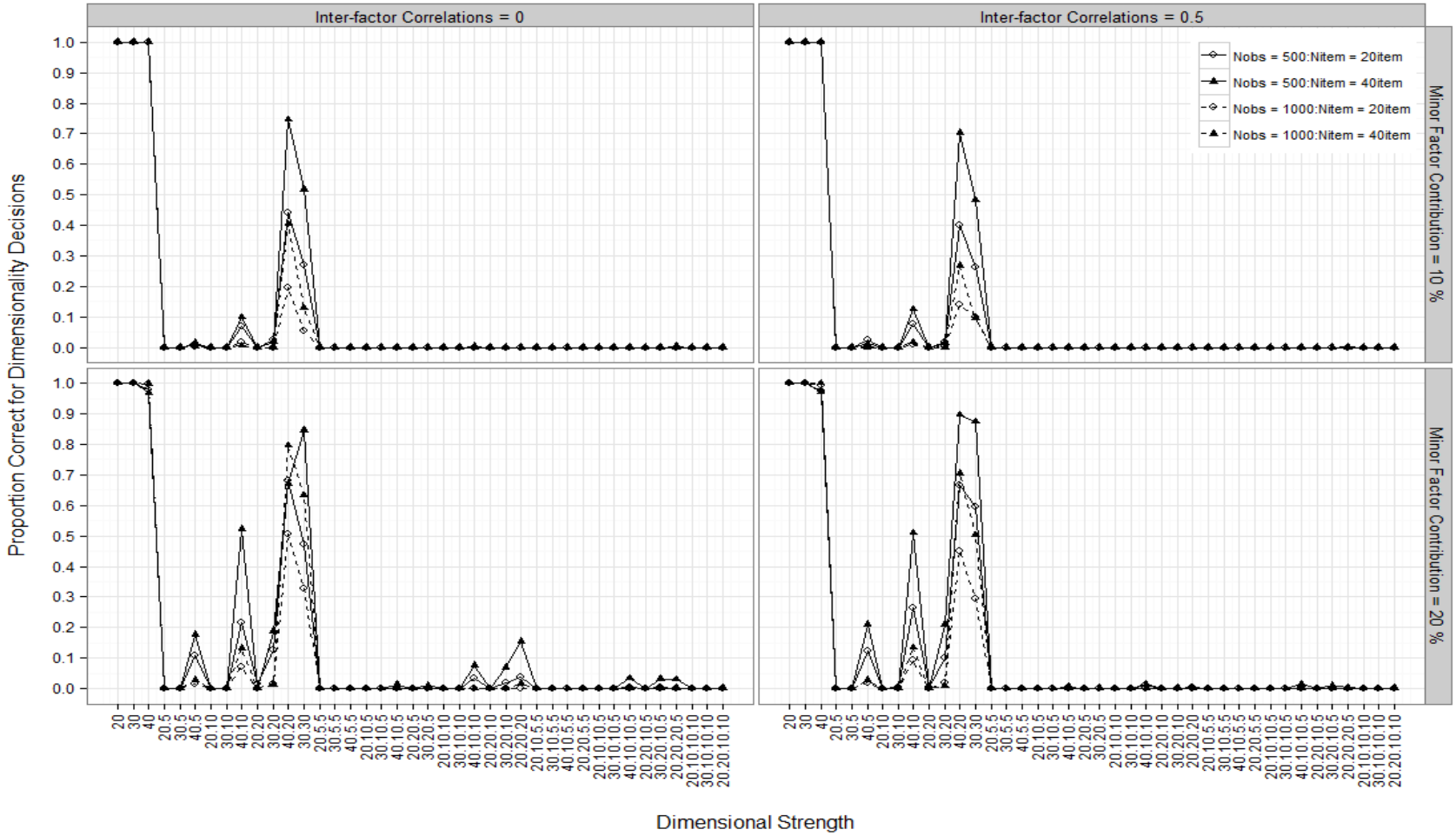


Figure 40. Proportion of Datasets in Which the Quasi-true Number of Dimensions is Correctly Identified by the **Mplus WLSM RMSEA Index**

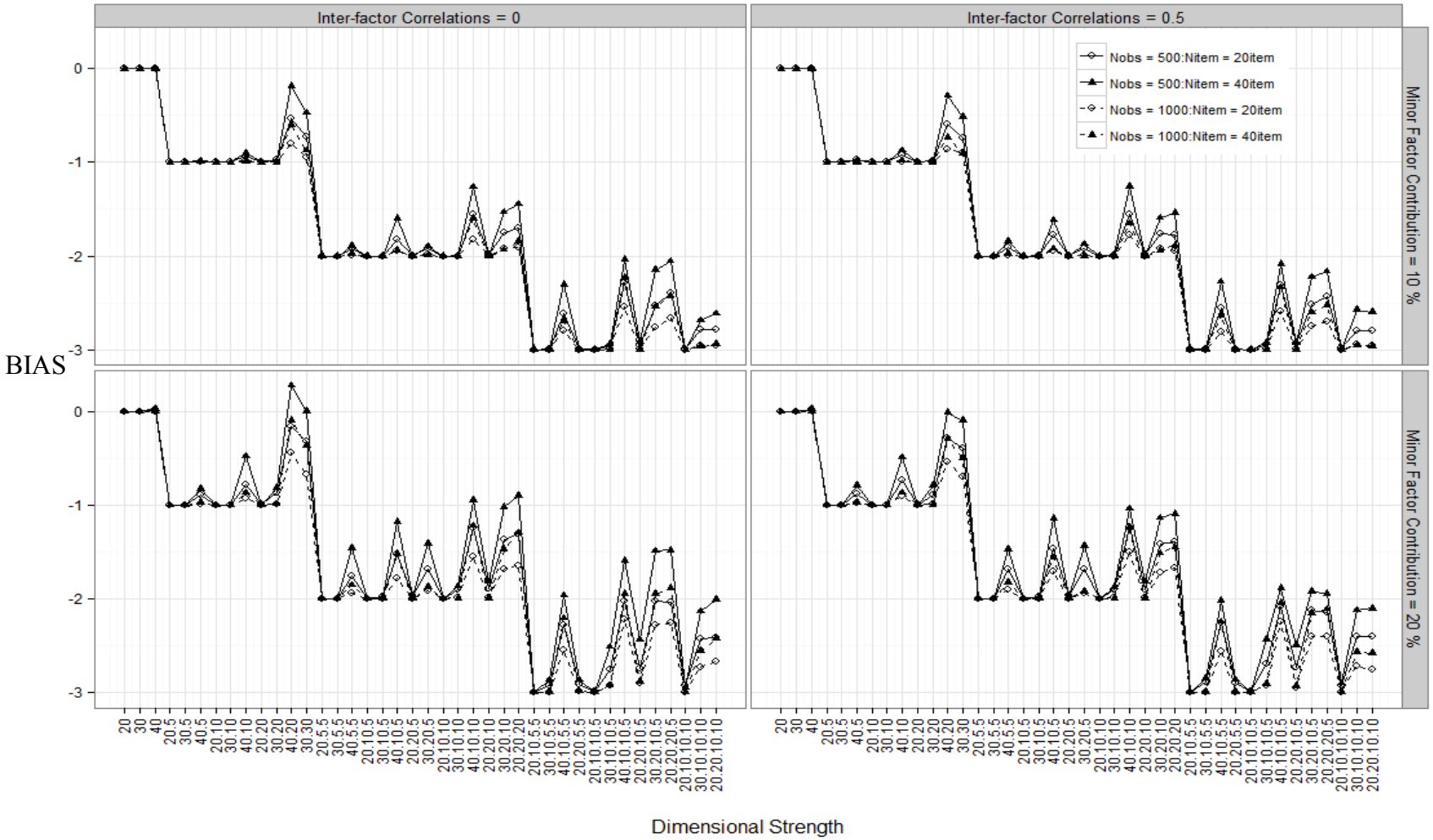


Figure 41. Bias with respect to the Quasi-true Number of Dimensions for the **Mplus WLSM RMSEA Index**

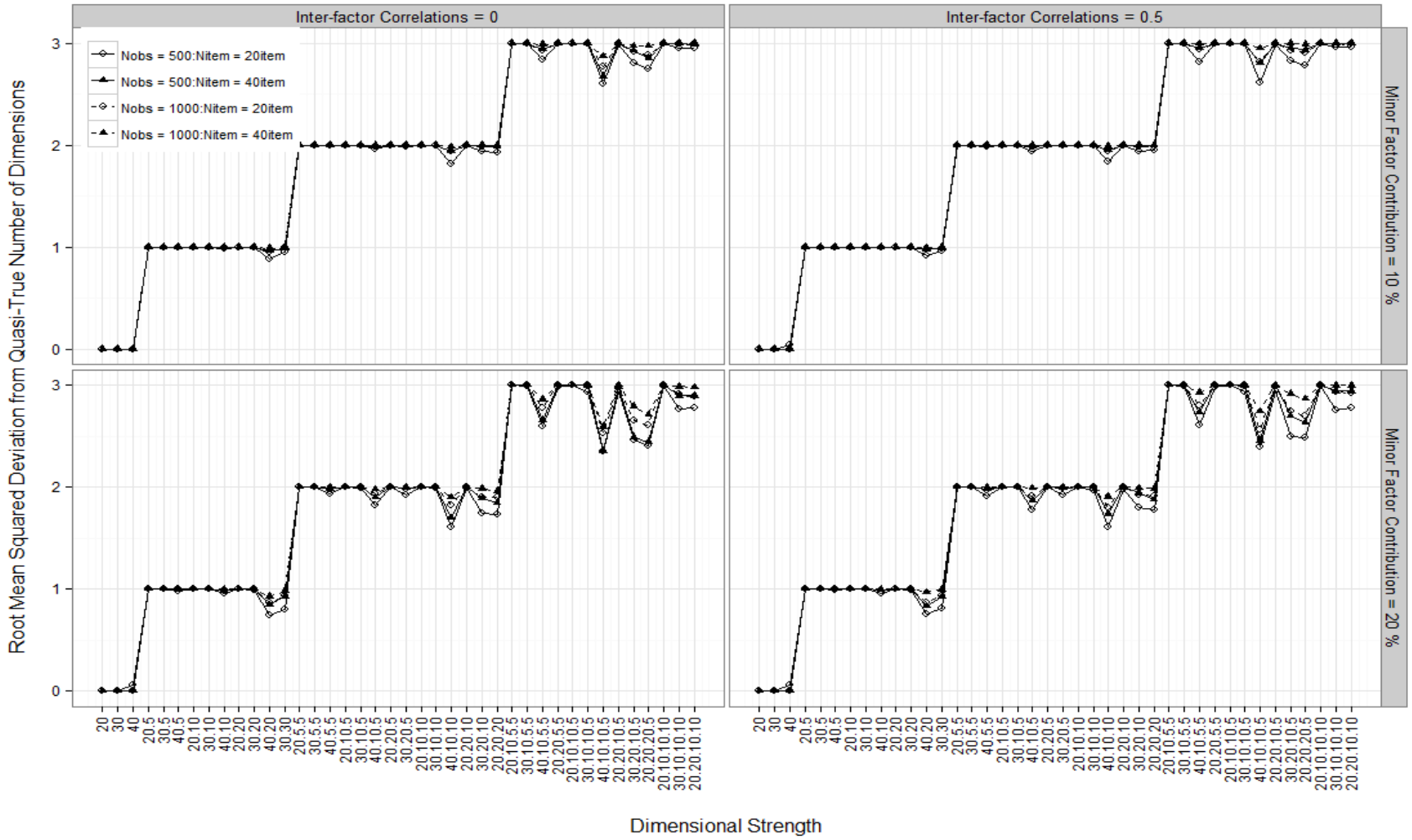


Figure 42. Root Mean Squared Deviation from the Quasi-true Number of Dimensions for the **Mplus WLSM RMSEA Index**

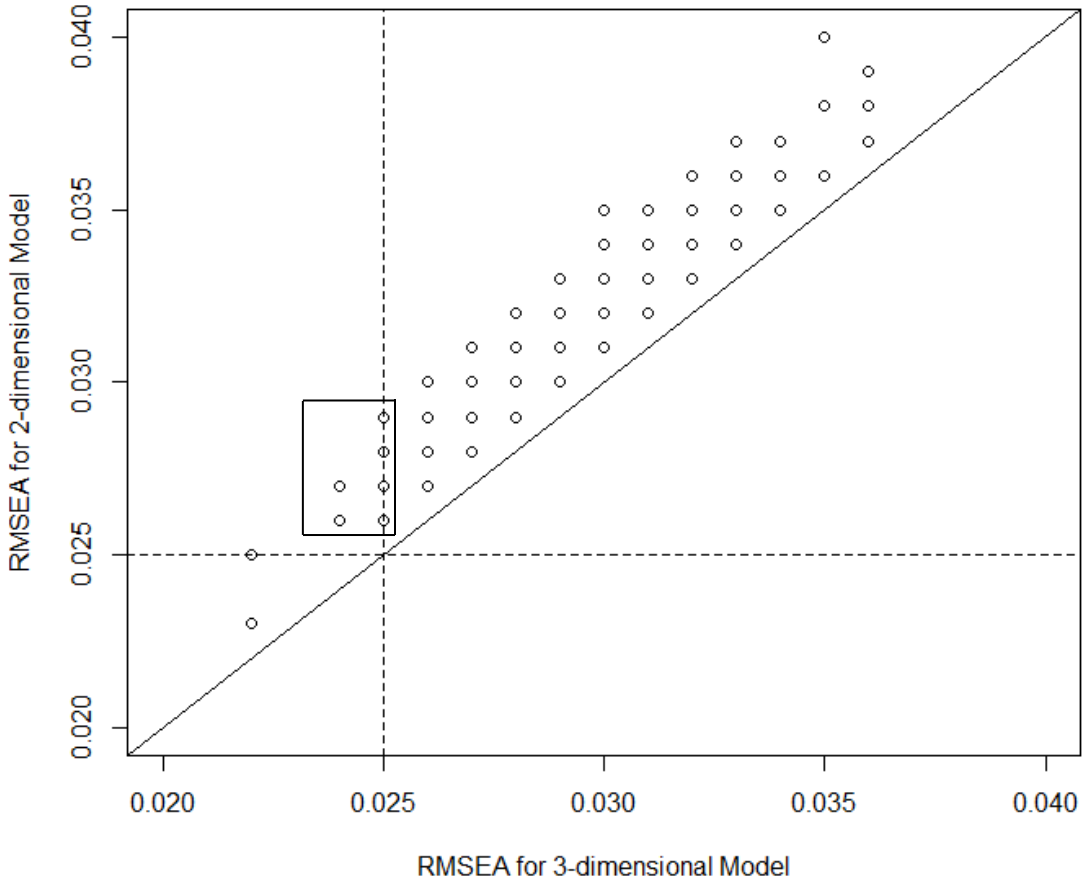


Figure 43. The Replications with Correctly Identified Quasi-True Number of Dimensions Using a Cut-off Value of 0.025 for the RMSEA Index

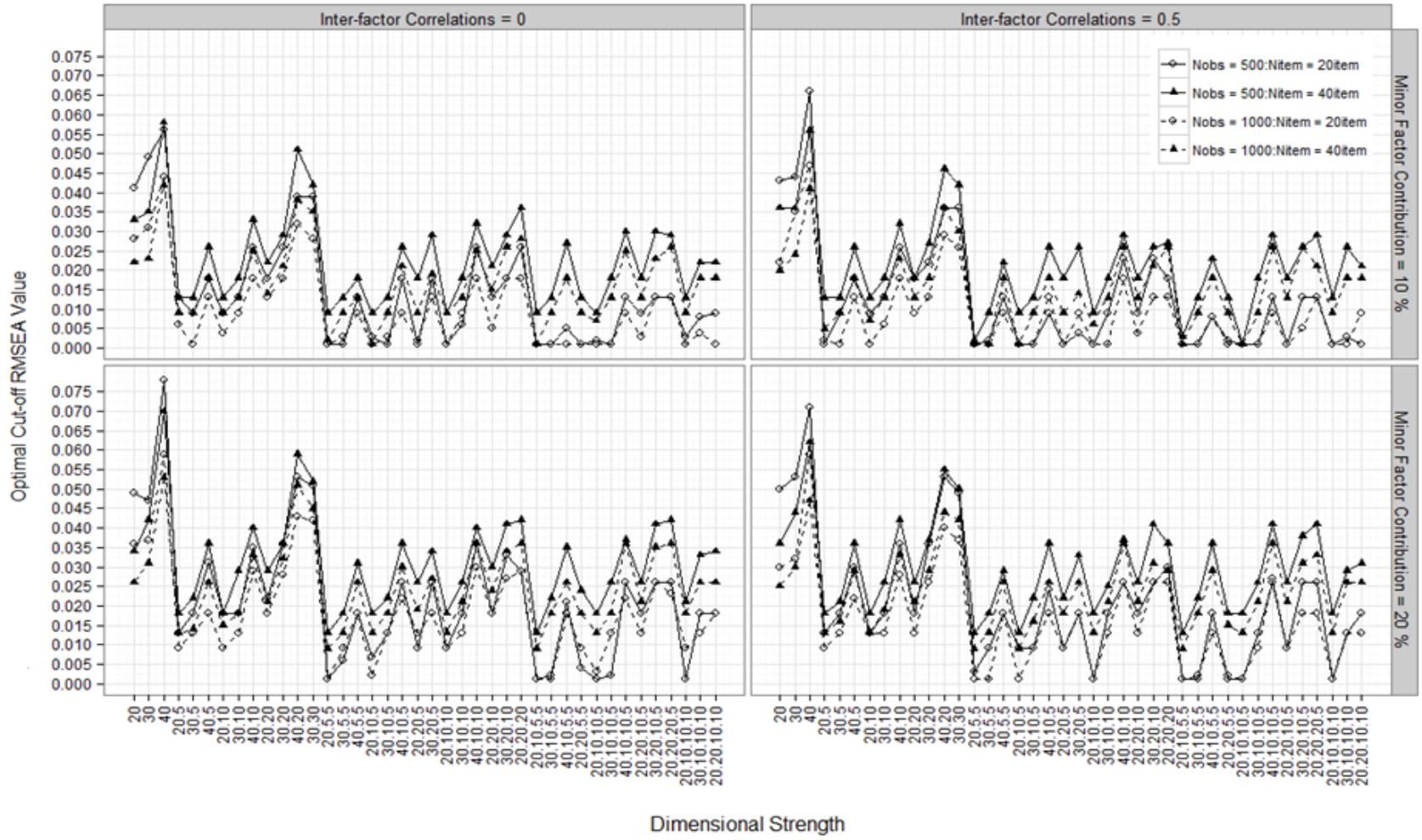


Figure 44. Optimal Cut-off Values for the Mplus WLSM RMSEA Index Across Simulation Conditions

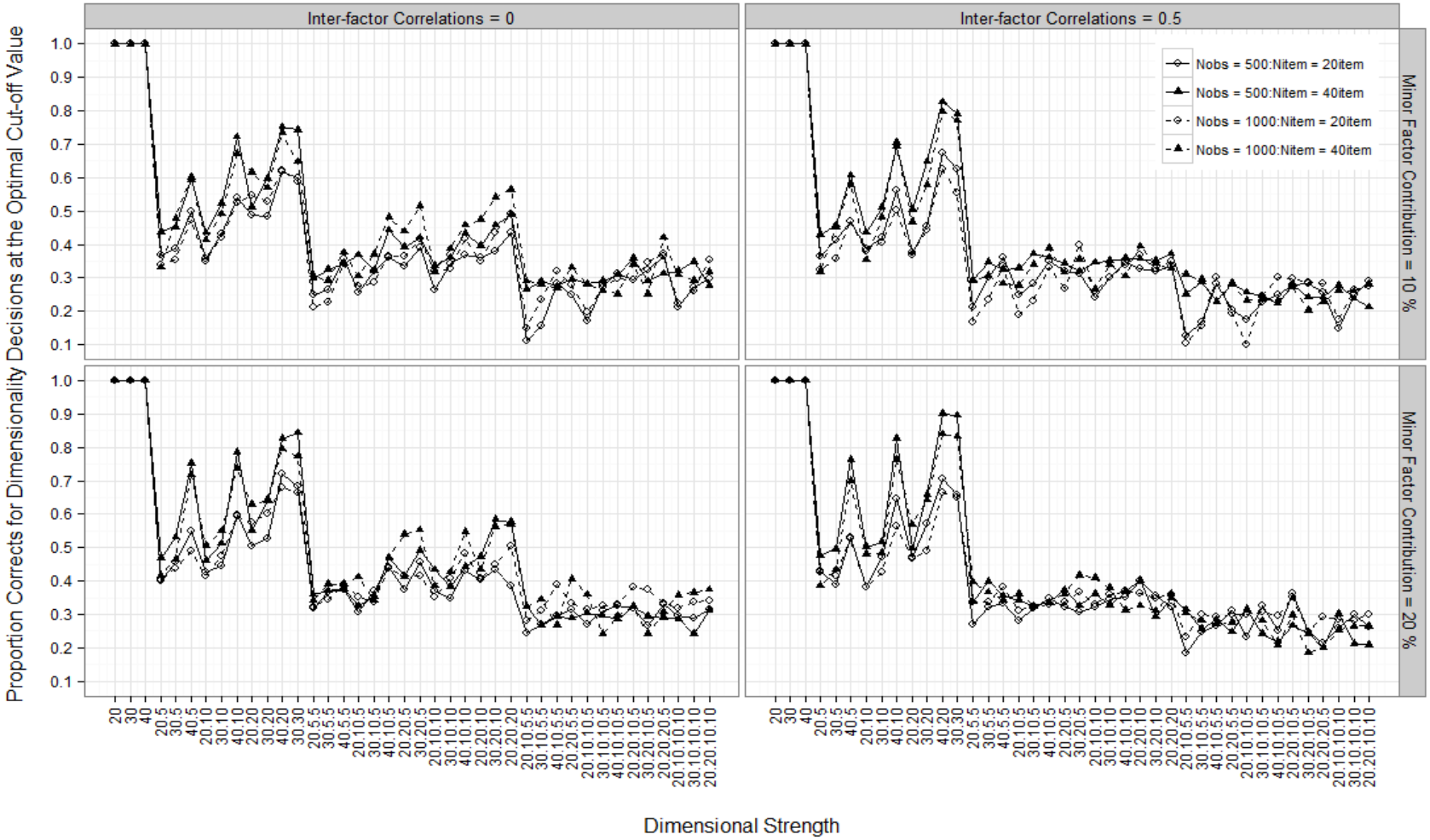


Figure 45. Maximized Proportions in Correctly Identifying Quasi-True Number of Dimensions at the Corresponding Optimal Cut-off Value for the **Mplus WLSM RMSEA Index**

Comparative Fit Index. The results for the CFI index obtained from Mplus WLS mean-adjusted and mean-and-variance adjusted chi-square statistics were almost identical to each other. Therefore, the results for the CFI index obtained from Mplus WLS mean-adjusted chi-square statistics are only reported and shown in Figure 46, Figure 47, and Figure 48. A decision was reached using the CFI index for all replications in all conditions.

Figure 46 shows that the CFI index correctly identified the quasi-true number of dimensions for almost all replications when there was one major dimension and never correctly identified the quasi-true number of dimensions in all other occasions. The pattern in Figure 46 and Figure 47 indicates that the CFI index selected one-dimensional models all the time regardless of the underlying structure using a cut-off value of 0.95. A similar process as for the RMSEA index was followed to identify an optimal cut-off value for the CFI index for each condition. Figure 49 shows the optimal cut-off values across the simulation conditions. In general, the optimal values for the 40-item conditions were slightly smaller than the optimal values for the 20-item conditions, but the differences were negligible within a range of 0.01. A cut-off value between 0.99 and 1 was found optimal in most conditions.

Figure 50 shows the maximized proportion of correct decisions with respect to the quasi-number of dimensions at the corresponding optimal value for each condition. The maximized proportion corrects varied between 20% and 60% for the conditions with two major dimensions, between 0% and 30% for the conditions with three major dimensions, and between 0% and 10% for the conditions with four major dimensions. Higher rates were generally observed when there were 40 items.

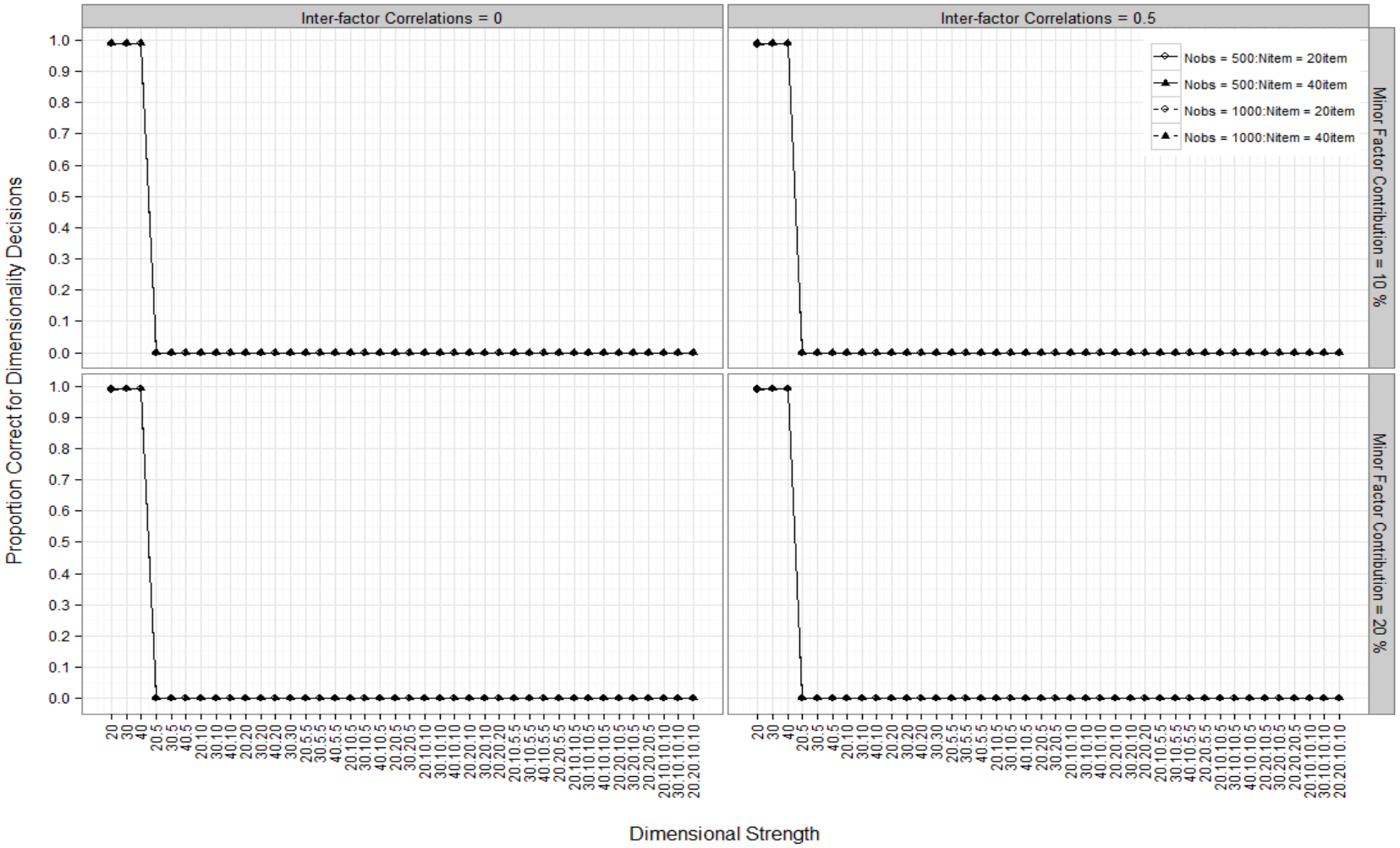


Figure 46. Proportion of Datasets in Which the Quasi-true Number of Dimensions is Correctly Identified by the **Mplus WLSM CFI Index**

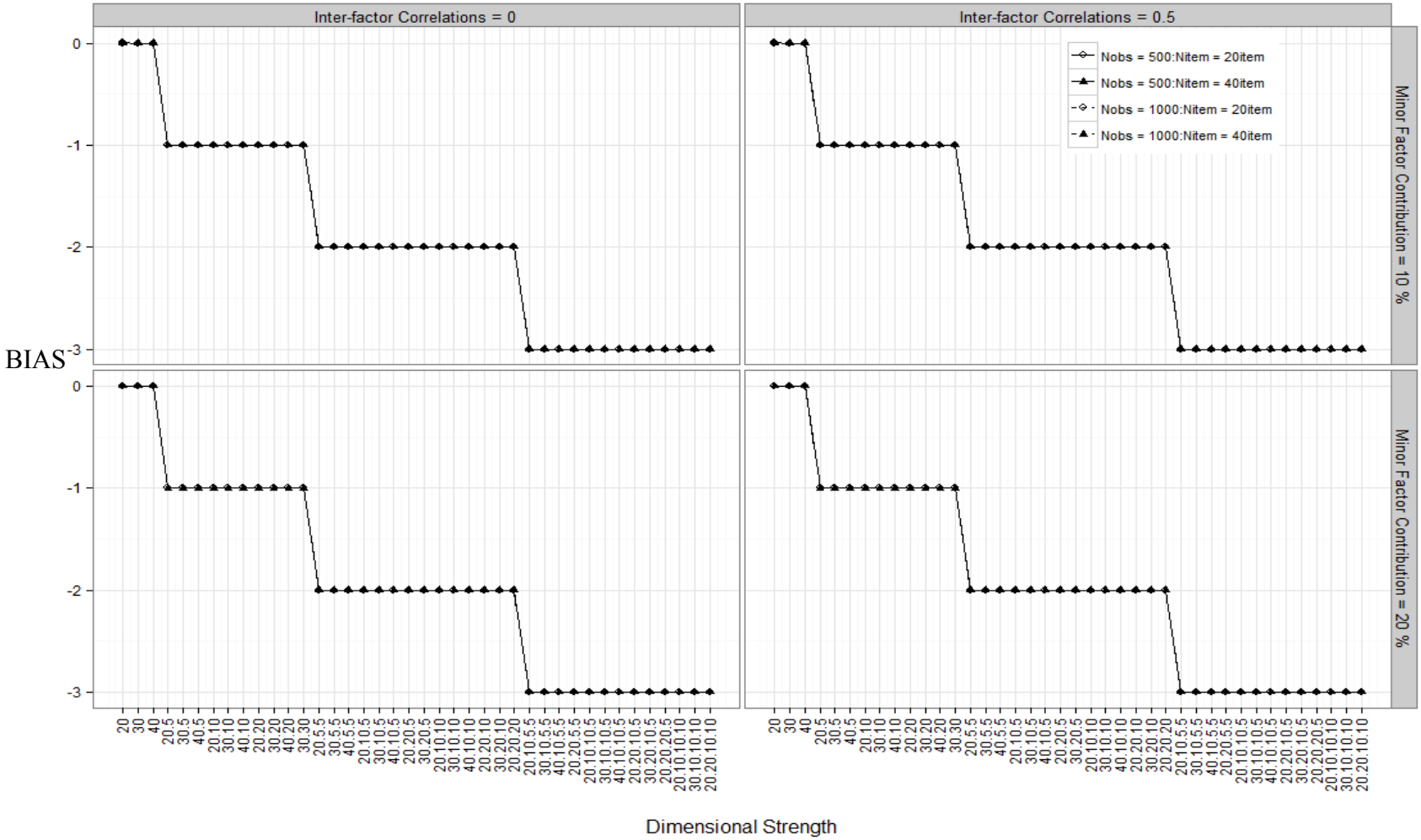


Figure 47. Bias with respect to the Quasi-true Number of Dimensions for the **Mplus WLSM CFI Index**

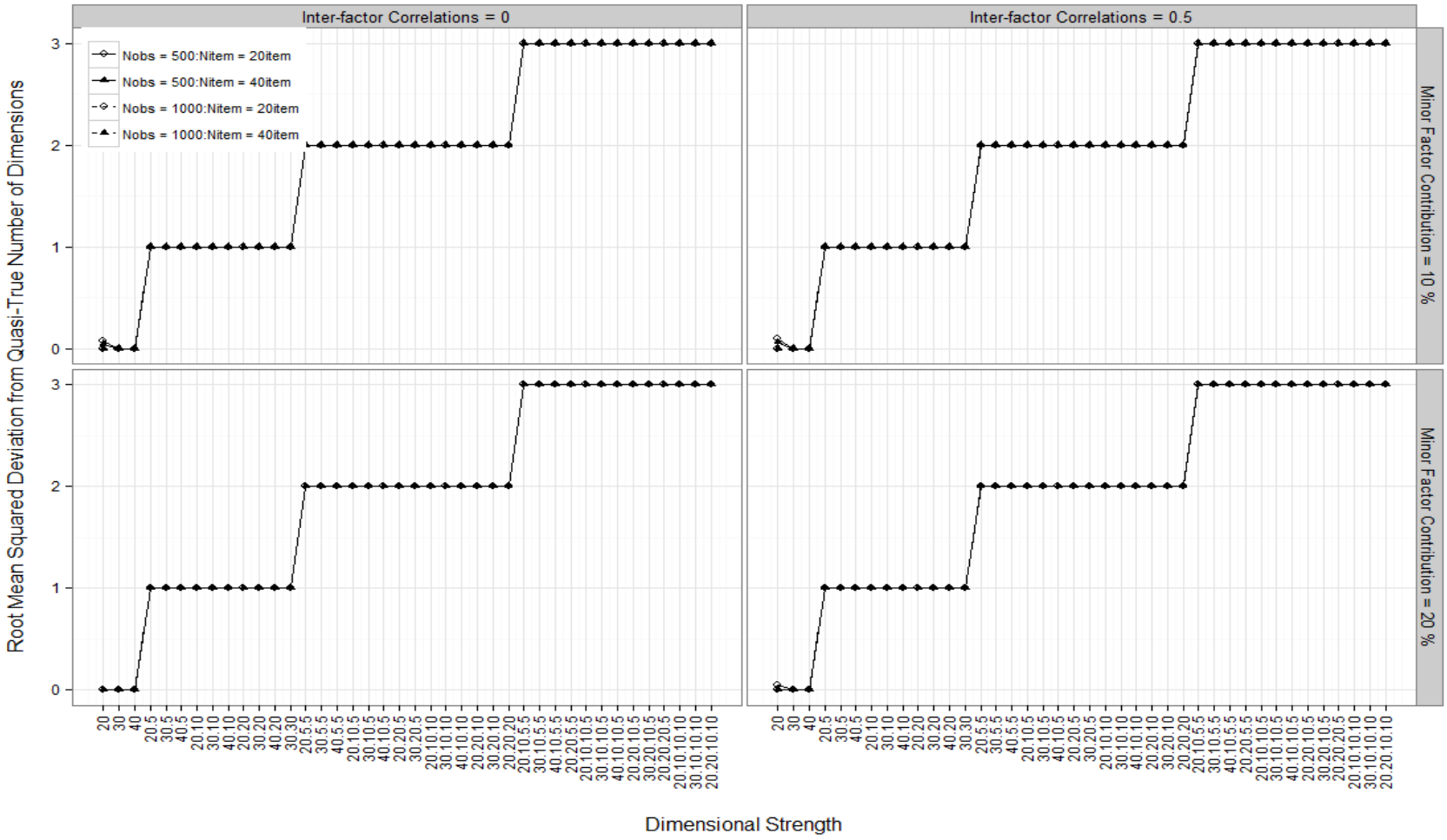


Figure 48. Root Mean Squared Deviation from the Quasi-true Number of Dimensions for the **Mplus WLSM CFI Index**

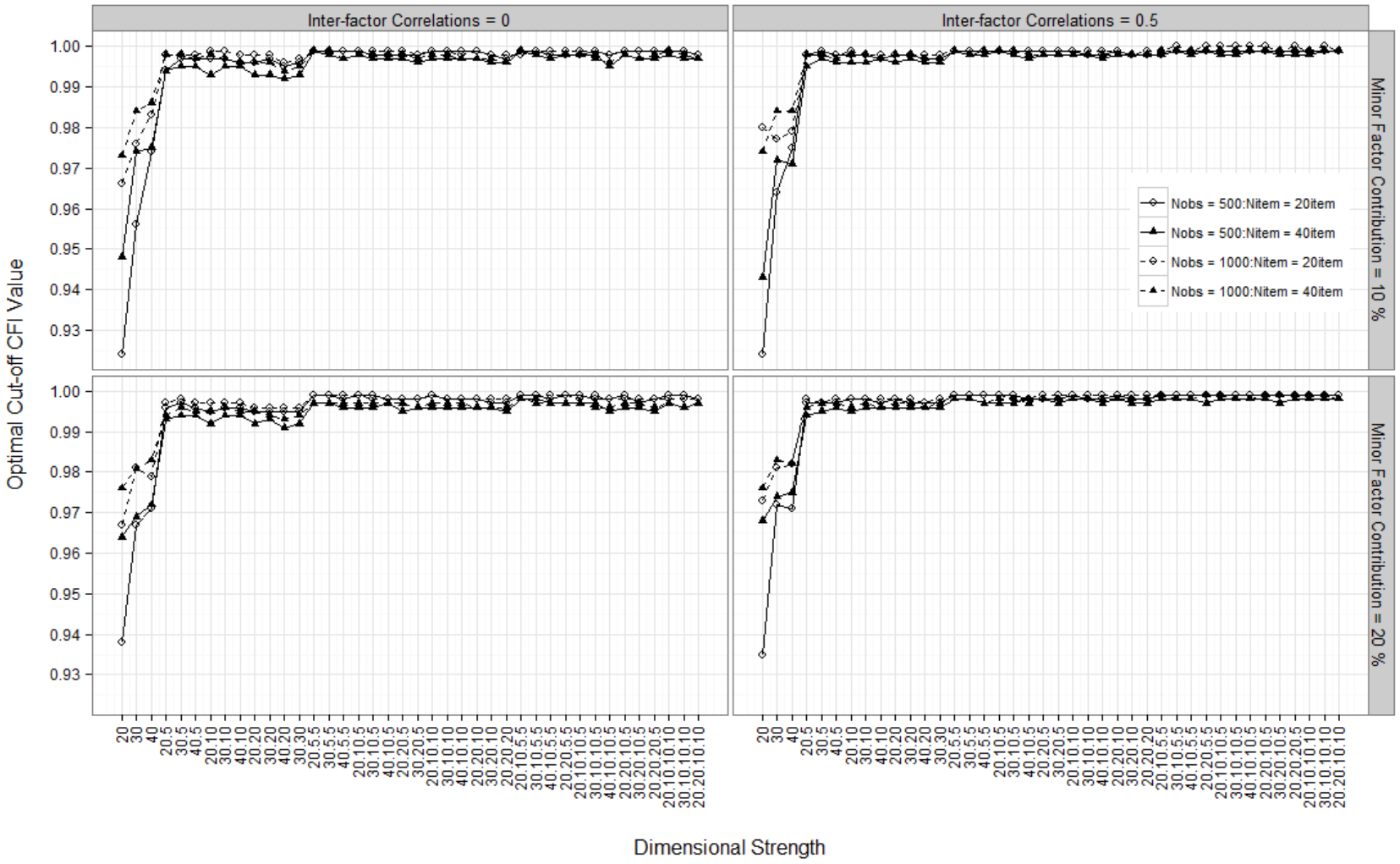


Figure 49. Optimal Cut-off Values for the **Mplus WLSM CFI Index** Across Simulation Conditions

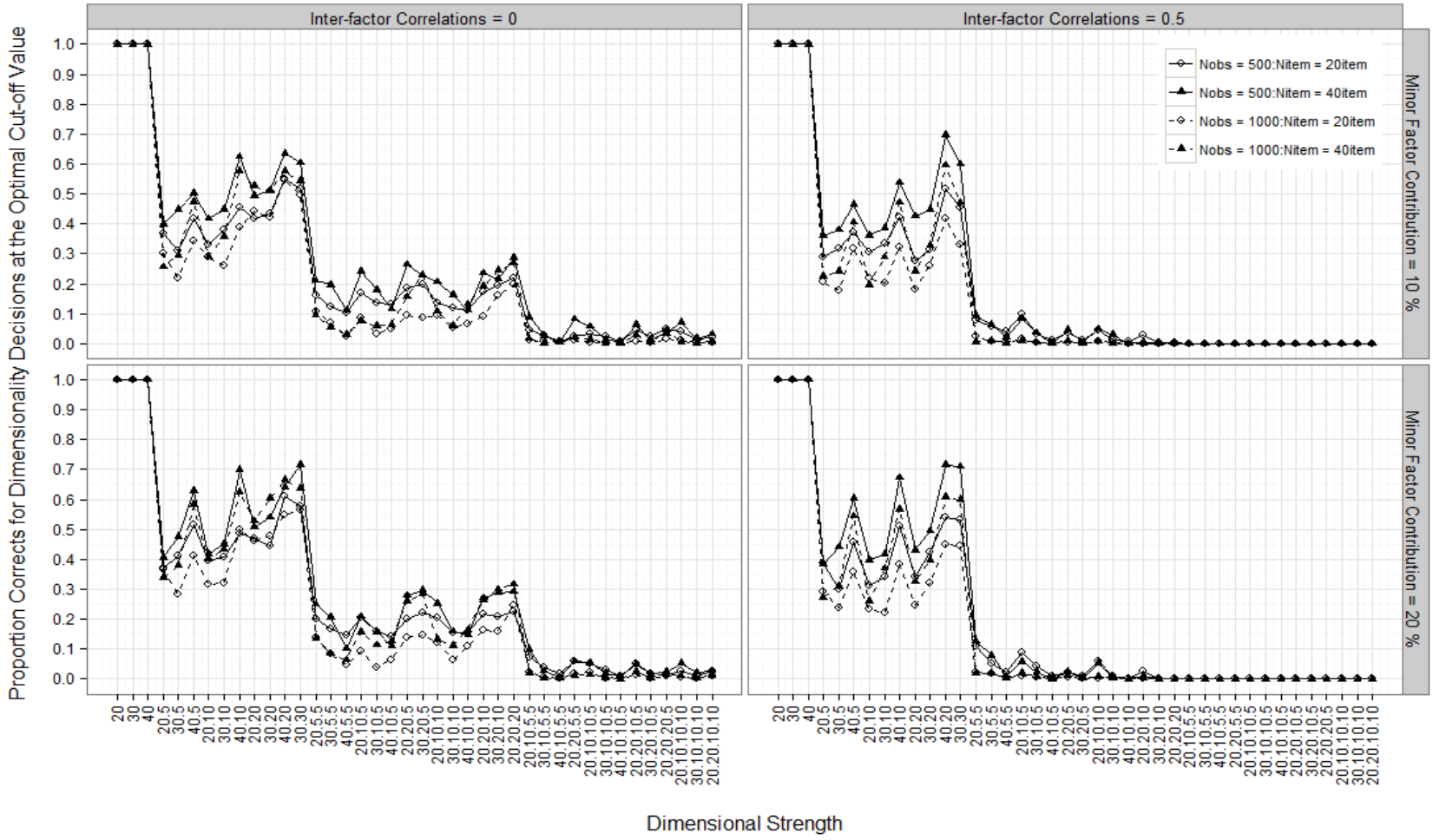


Figure 50. Maximized Proportions in Correctly Identifying Quasi-True Number of Dimensions at the Corresponding Optimal Cut-off Value for the **Mplus WLSM CFI Index**

Standardized Root Mean Squared Residual. The results for the SRMR index obtained from Mplus WLS are shown in Figure 51, Figure 52, and Figure 53. A decision was reached using the SRMR index for all replications in all conditions. Figure 51 shows that the SRMR index correctly identified the quasi-true number of dimensions in almost all conditions when there was one major dimension and the sample size was 1000 using a cut-off value of 0.07. The proportion of correct decisions with respect to the quasi-true number of dimensions varied between 10% and 90% for the conditions with one major dimension, and was particularly low when the sample size was 500. Success rates between 10% and 50% were observed for the conditions in which there were two major dimensions, the number of items was 40, and the inter-factor correlation was zero. In all other simulation conditions, the proportion of correct decisions with respect to the quasi-true number of dimensions was either lower than 20% or 0%. The patterns in Figure 52 and Figure 53 indicates that there was a tendency to select a one-dimensional model in many conditions, particularly when there were three or four major dimensions in the generating model.

The results suggested that a cut-off value of 0.07 might be unnecessarily high for the SRMR index. As a result of a similar procedure described before for the RMSEA index, an optimal cut-off value for the SRMR index was found for each simulation condition and these values are shown in Figure 54. The corresponding maximized proportion correct values are shown in Figure 55. Figure 54 indicates that the optimal cut-off values for the SRMR index are mainly dependent on the sample size and number of items. The optimal cut-off values for the conditions with one major dimension are too high, because the maximum SRMR value across all replications is reported in those conditions. So, the corresponding success rates in Figure 55 were 100% in those conditions. When there were two major dimensions and no correlation among the major dimensions, a cut-off value between 0.06 and 0.07 for the conditions with the sample size of 500 and number of items 40, between 0.05 and 0.06 for the conditions with the sample size of 500 and number of items 20, and between 0.04 and 0.05 for the conditions with the sample size of 1000 maximized the decisions. When there were three or four major dimensions and no correlation among the major dimensions, a cut-off value between 0.05

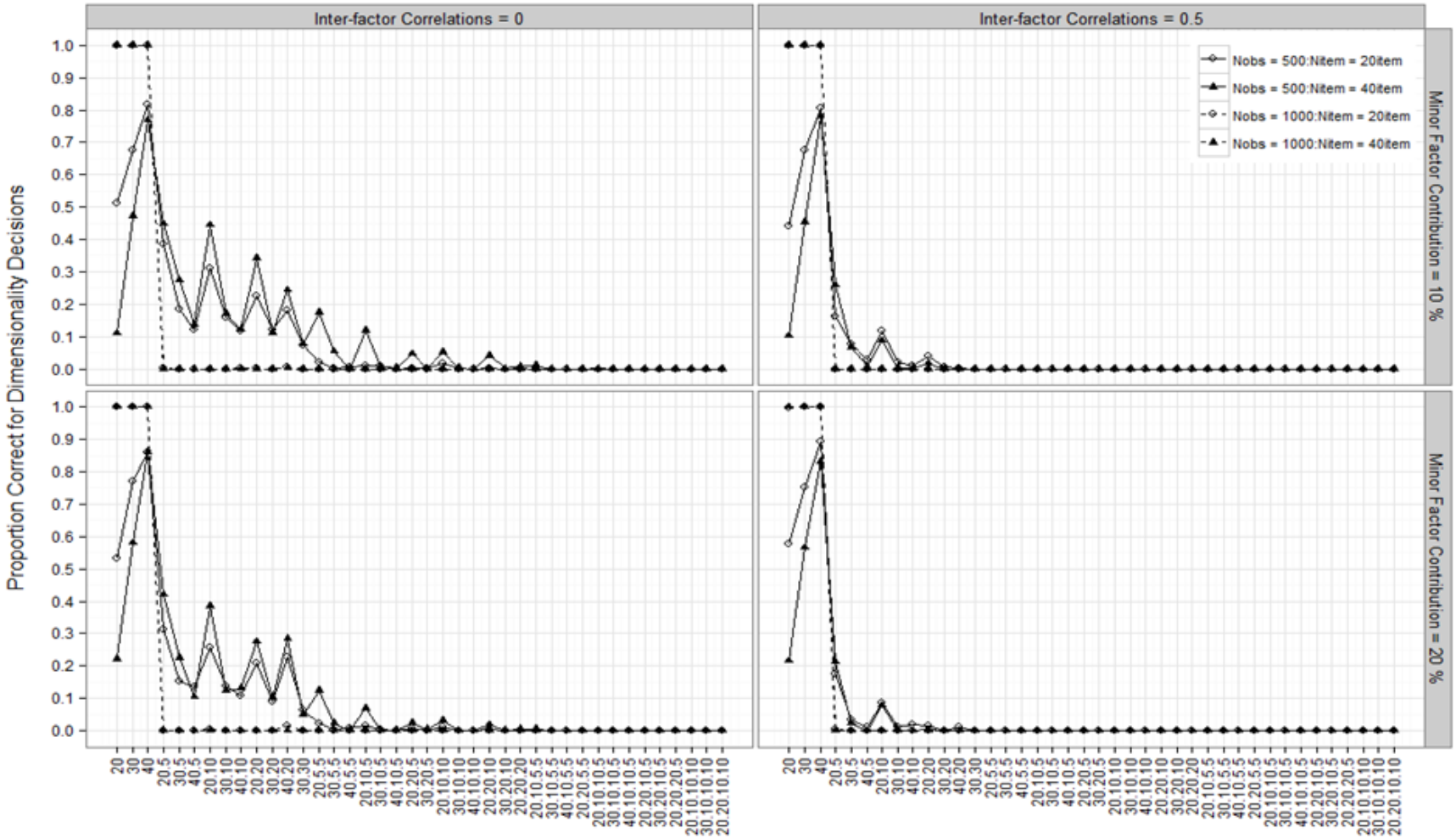


Figure 51. Proportion of Datasets in Which the Quasi-true Number of Dimensions is Correctly Identified by the **Mplus WLS SRMR Index**

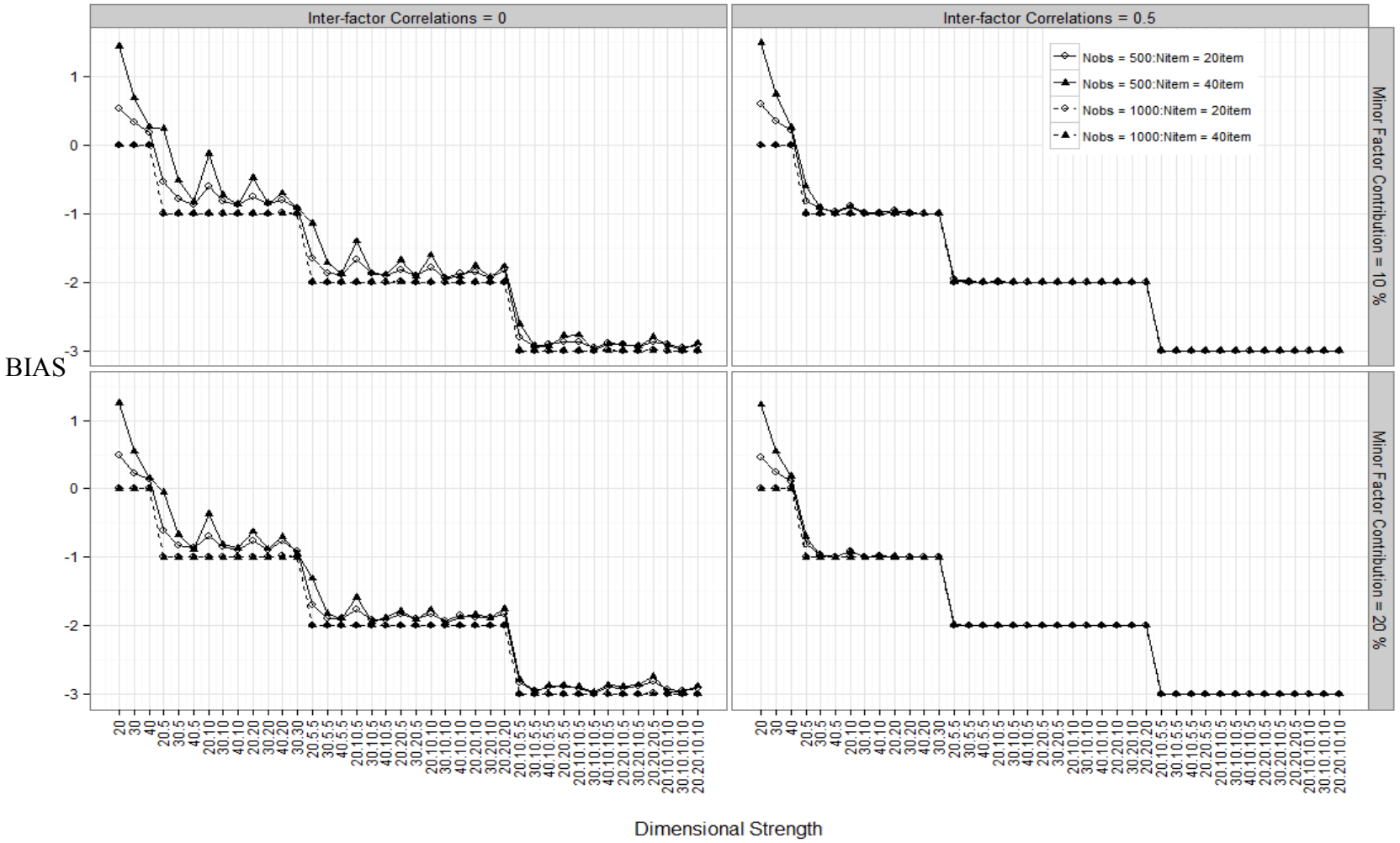


Figure 52. Bias with respect to the Quasi-true Number of Dimensions for the **Mplus WLS SRMR Index**

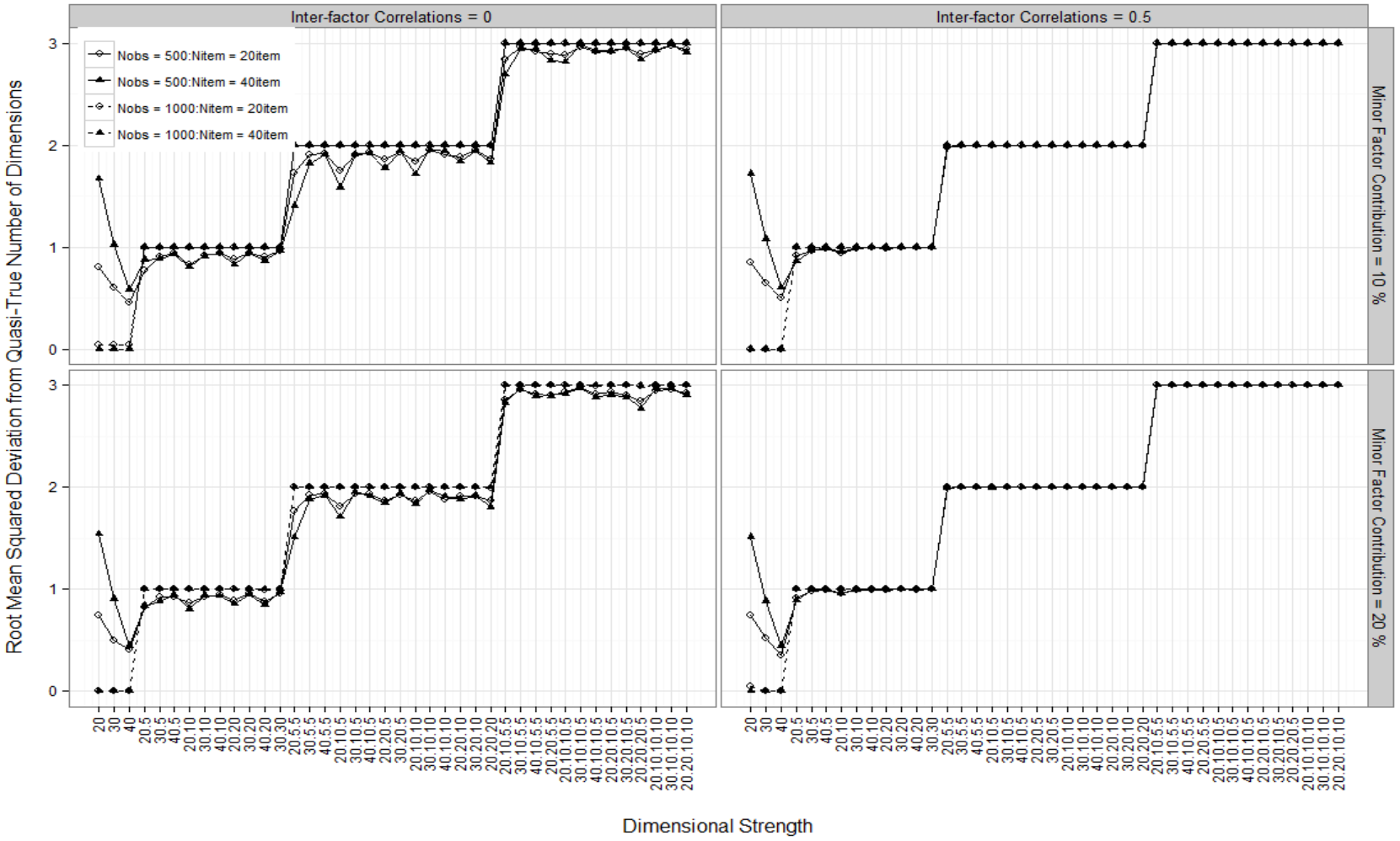


Figure 53. Root Mean Squared Deviation from the Quasi-true Number of Dimensions for the **Mplus WLS SRMR Index**

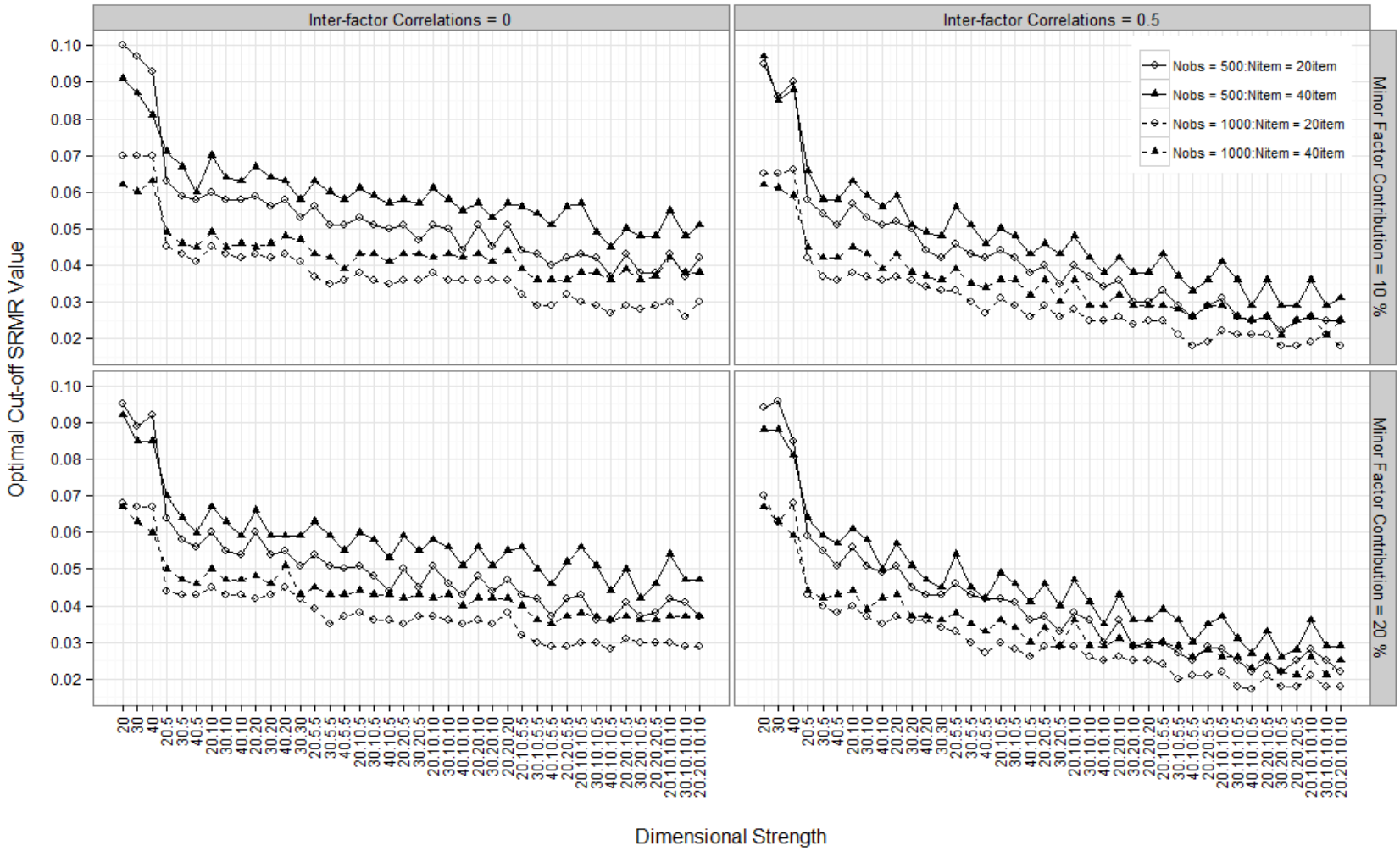


Figure 54. Optimal Cut-off Values for the **Mplus WLS SRMR Index** Across Simulation Conditions

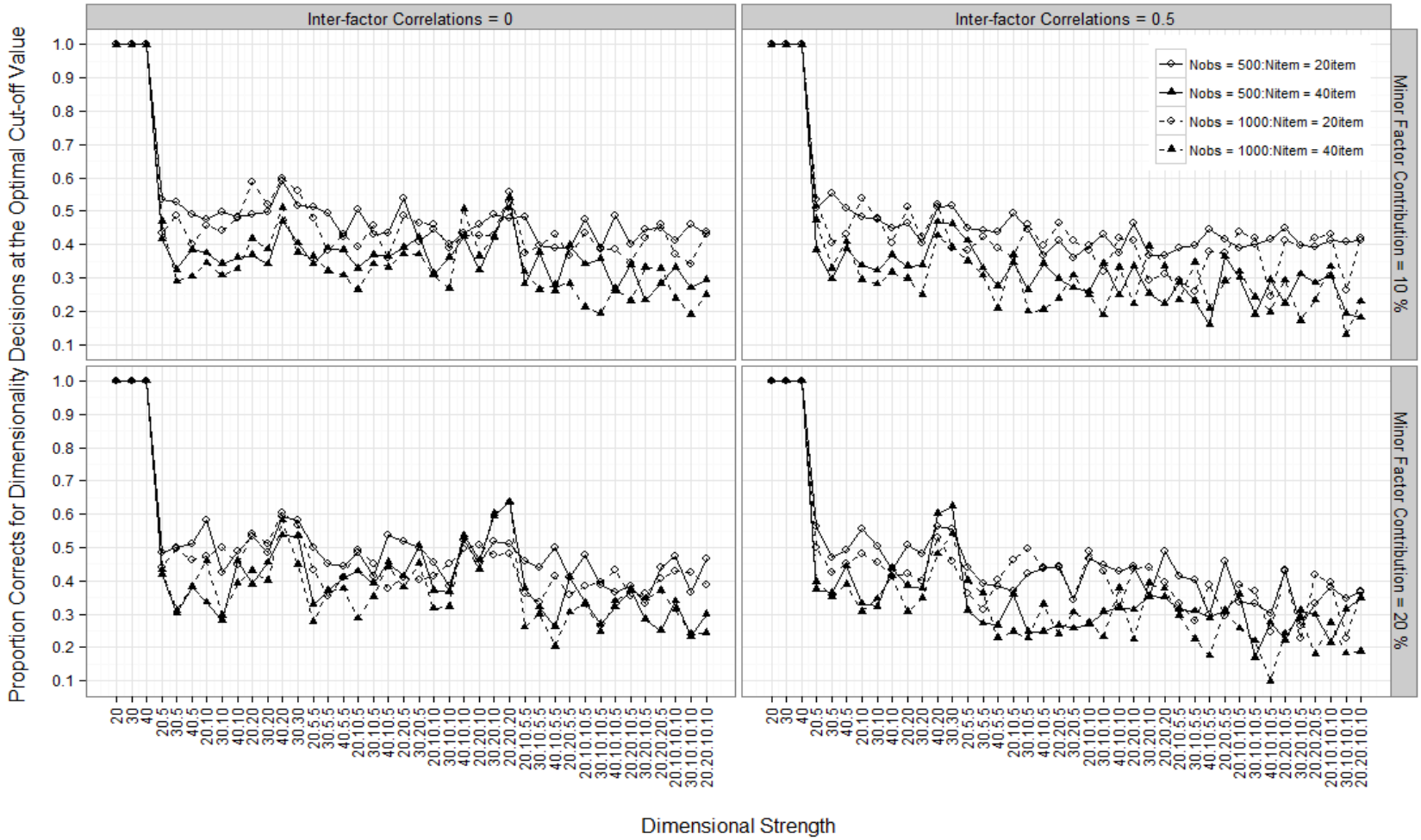


Figure 55. Maximized Proportions in Correctly Identifying Quasi-True Number of Dimensions at the Corresponding Optimal Cut-off Value for **the Mplus WLS SRMR Index**

and 0.06 for the conditions with the sample size of 500 and number of items 40, between 0.04 and 0.05 for the conditions with the sample size of 500 and number of items 20, and between 0.03 and 0.045 for the conditions with the sample size of 1000 maximized the decisions. The amount of variance accounted for by the minor factors did not seem to influence these intervals much, but the corresponding cut-off intervals were slightly smaller (around .01) when the correlation among the major dimensions increased to 0.5. The maximized proportions for the corresponding optimal values varied between 10% and 60% across conditions with higher rates in general observed for the conditions with 20-items.

Bayesian Information Criterion. The results for the BIC index based on the Mplus FIML estimation are presented in Table 40 and in Figure 56 through Figure 58. The models with up to six dimensions were fitted using Mplus MLR estimator, the BIC value was extracted for each model, and the model with the smallest BIC was found. In some replications, BIC was the smallest for the six-dimensional model and a decision was not reached. The proportion of those replications was 1% for only two conditions out of 640, and 0% for the rest of the conditions. In some other replications, the decision was not reached due to convergence issues at some point when fitting models with one to six dimensions, and Table 40 shows the proportions of no-decision replications due to convergence issues. The proportions were below 5% in most conditions when there were 20 items, and were below 10% in most conditions when there were 40 items. The replications with no-decision due to convergence problems were eliminated from further analysis. For other replications in which a decision was not reached after fitting the six-dimensional model, the number of suggested dimensions was assumed to be six for further computations.

Figure 56 shows the proportion of replications in which the quasi-true number of dimensions was correctly identified by BIC across simulation conditions. BIC was successful in correctly identifying the model with the quasi-true number of dimensions with a success rate above 90% when there was one major dimension. When there were two major dimensions, the success rate was between 10% and 70% for the conditions with the sample size of 1000 and 40 items. In all other conditions, the success rate was

Table 40. *Proportion of Datasets in Which the Dimensionality Decision Cannot be Reached Due to Convergence Problems by the Bayesian Information Criterion based on the Mplus FIML Estimation with the MLR Estimator*

Minor Factors	10 %				20 %			
Number of Items	20		40		20		40	
Sample Size	500	1000	500	1000	500	1000	500	1000
20	0.03	-	0.02	-	0.04	0.03	-	-
30	-	-	-	0.01	0.01	-	0.01	0.02
40	0.01	-	0.02	0.02	0.01	0.01	0.02	0.04
20.5	-	-	0.01	-	-	0.01	0.02	0.02
30.5	-	0.01	0.01	0.01	0.01	0.01	-	0.04
40.5	-	0.01	0.02	0.07	0.02	-	0.03	0.06
20.10	0.03	-	-	0.06	0.03	0.01	0.01	0.12
30.10	0.01	-	0.09	0.26	0.01	0.01	0.23	0.50
40.10	0.03	0.04	0.04	0.11	0.06	0.06	0.09	0.08
20.20	-	-	0.01	0.04	-	-	0.01	0.08
30.20	0.02	0.03	0.16	0.32	0.02	0.05	0.19	0.46
40.20	0.06	0.06	0.15	0.19	0.10	0.17	0.20	0.19
30.30	0.05	0.07	0.10	0.16	0.06	0.07	0.14	0.21
20.5.5	0.04	-	-	0.15	0.03	-	-	0.22
30.5.5	0.02	0.01	0.01	0.05	0.01	0.04	0.05	0.16
40.5.5	-	0.02	0.08	0.10	0.02	0.02	0.05	0.10
20.10.5	-	0.01	-	0.02	-	-	0.02	0.06
30.10.5	0.01	0.05	0.27	0.59	0.02	0.02	0.28	0.58
40.10.5	0.07	0.05	0.09	0.13	0.07	0.07	0.12	0.14
20.20.5	0.04	0.02	0.18	0.15	0.01	0.04	0.14	0.19
30.30.5	0.05	0.02	0.06	0.23	0.02	0.04	0.11	0.37
20.10.10	0.05	0.02	0.19	0.25	0.02	0.05	0.26	0.33
30.10.10	0.02	0.05	0.17	0.30	0.03	0.06	0.17	0.30
40.10.10	0.03	0.02	0.06	0.08	0.02	0.03	0.01	0.04
20.20.10	0.03	0.01	0.05	0.08	0.04	0.05	0.07	0.18
30.20.10	0.02	0.02	0.04	0.09	0.01	0.05	0.08	0.10
20.20.20	0.02	0.01	0.02	0.05	0.02	0.01	0.04	0.12
20.10.5.5	-	-	0.06	0.15	-	-	0.07	0.15
30.10.5.5	0.01	0.01	0.06	0.13	-	0.02	0.11	0.12
40.10.5.5	-	0.02	0.03	0.04	0.02	0.02	0.06	0.07
20.2/0.5.5	0.02	0.03	0.06	0.12	0.05	0.03	0.05	0.19
20.10.10.5	0.01	0.01	0.07	0.16	-	0.01	0.08	0.18
30.10.10.5	0.01	0.02	0.06	0.10	0.01	0.02	0.05	0.09
40.10.10.5	0.03	0.03	0.04	0.06	0.01	0.04	0.03	0.07
20.20.10.5	-	0.01	0.06	0.18	0.02	0.02	0.09	0.14
30.20.10.5	0.02	0.04	0.04	0.09	0.04	0.03	0.09	0.10
20.20.20.5	0.02	0.04	0.02	0.08	0.02	0.03	0.06	0.12
20.10.10.10	0.01	0.02	0.08	0.13	0.01	0.05	0.06	0.12
30.10.10.10	0.01	0.02	0.04	0.09	0.01	0.02	0.04	0.11
20.20.10.10	0.02	0.01	0.04	0.15	0.04	0.04	0.07	0.13

Note. Dashes indicate zero.

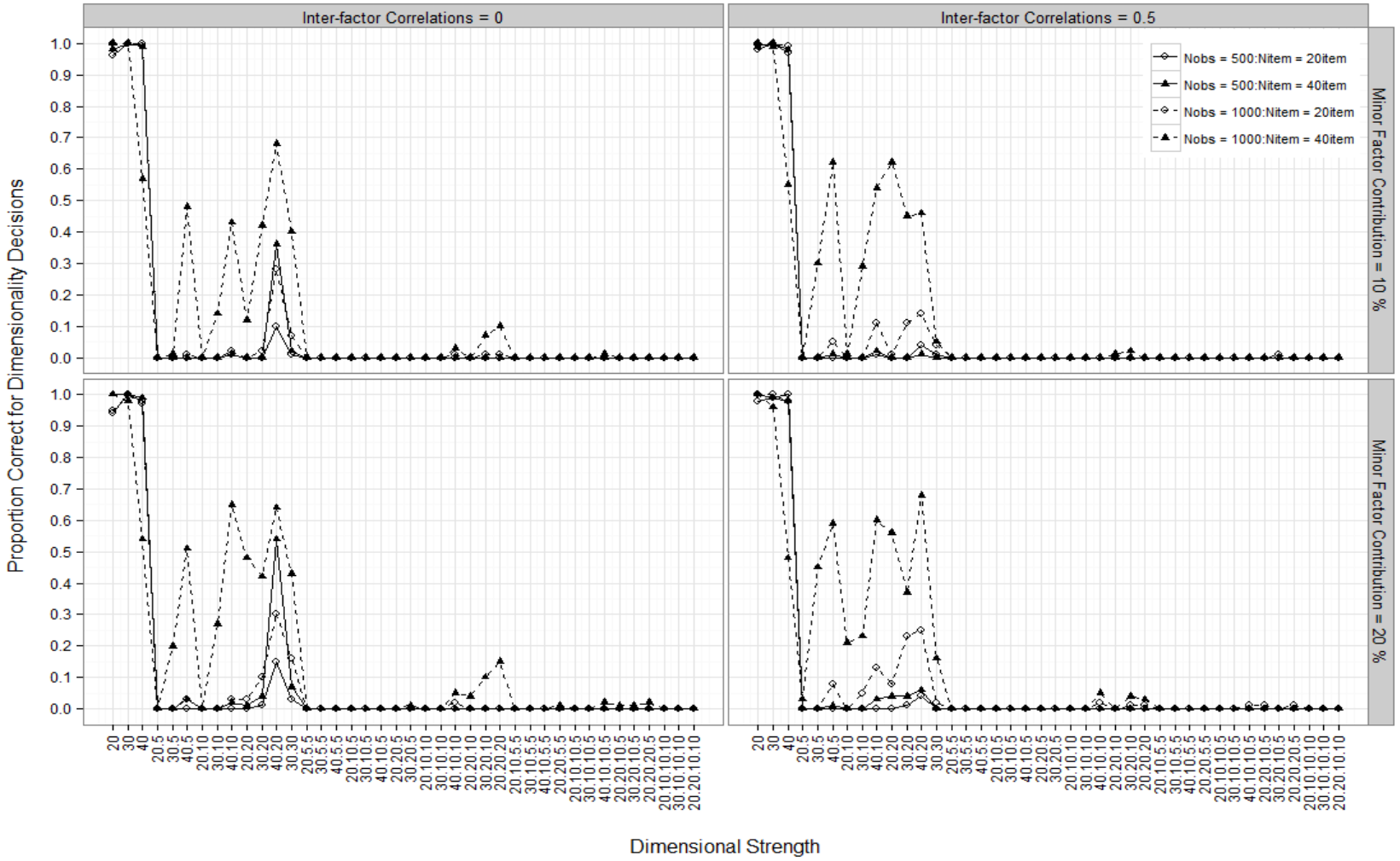


Figure 56. Proportion of Datasets in Which the Quasi-true Number of Dimensions is Correctly Identified by **BIC based on the Mplus MLR Estimator**

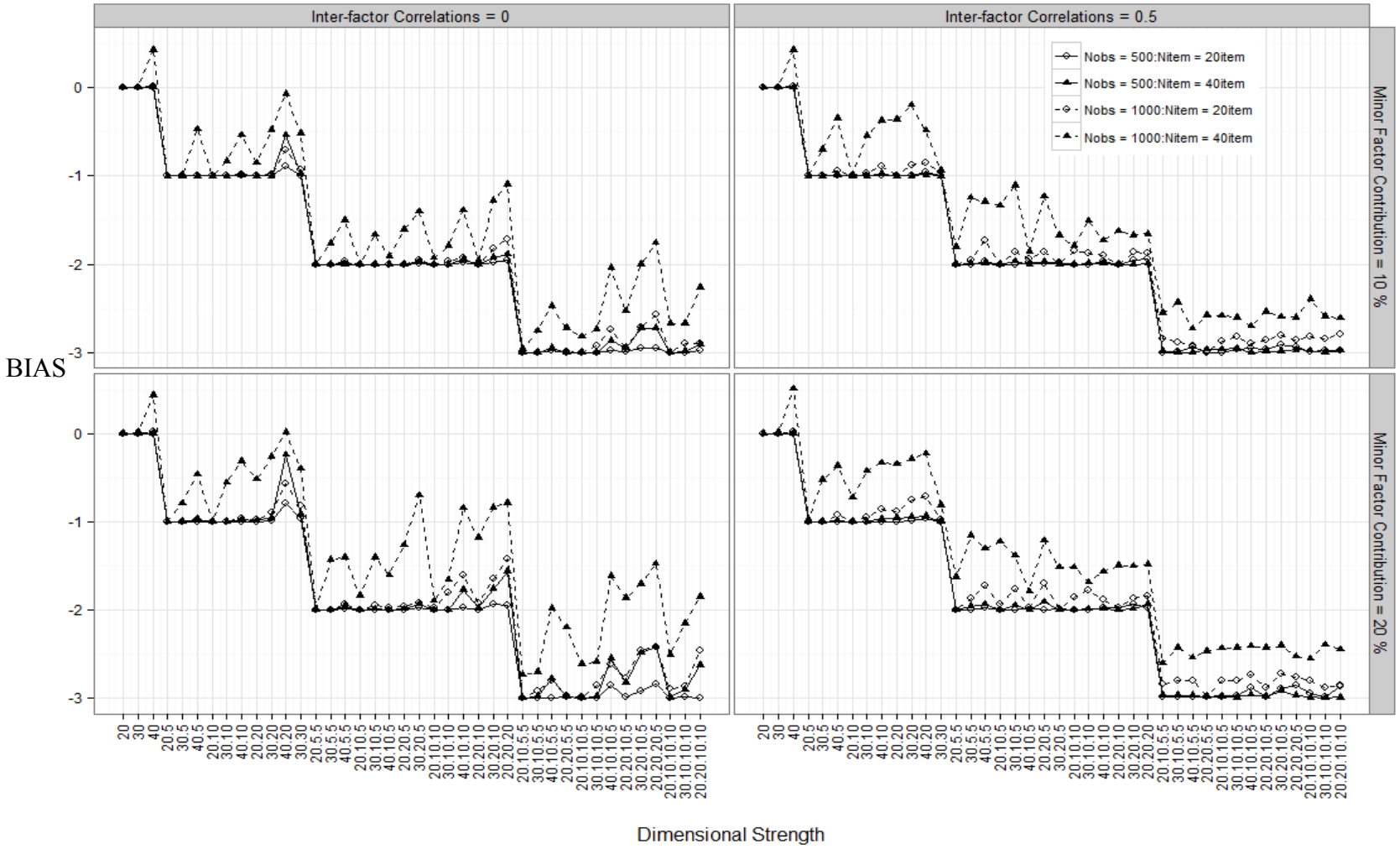


Figure 57. Bias with respect to the Quasi-true Number of Dimensions for **BIC based on the Mplus MLR Estimator**

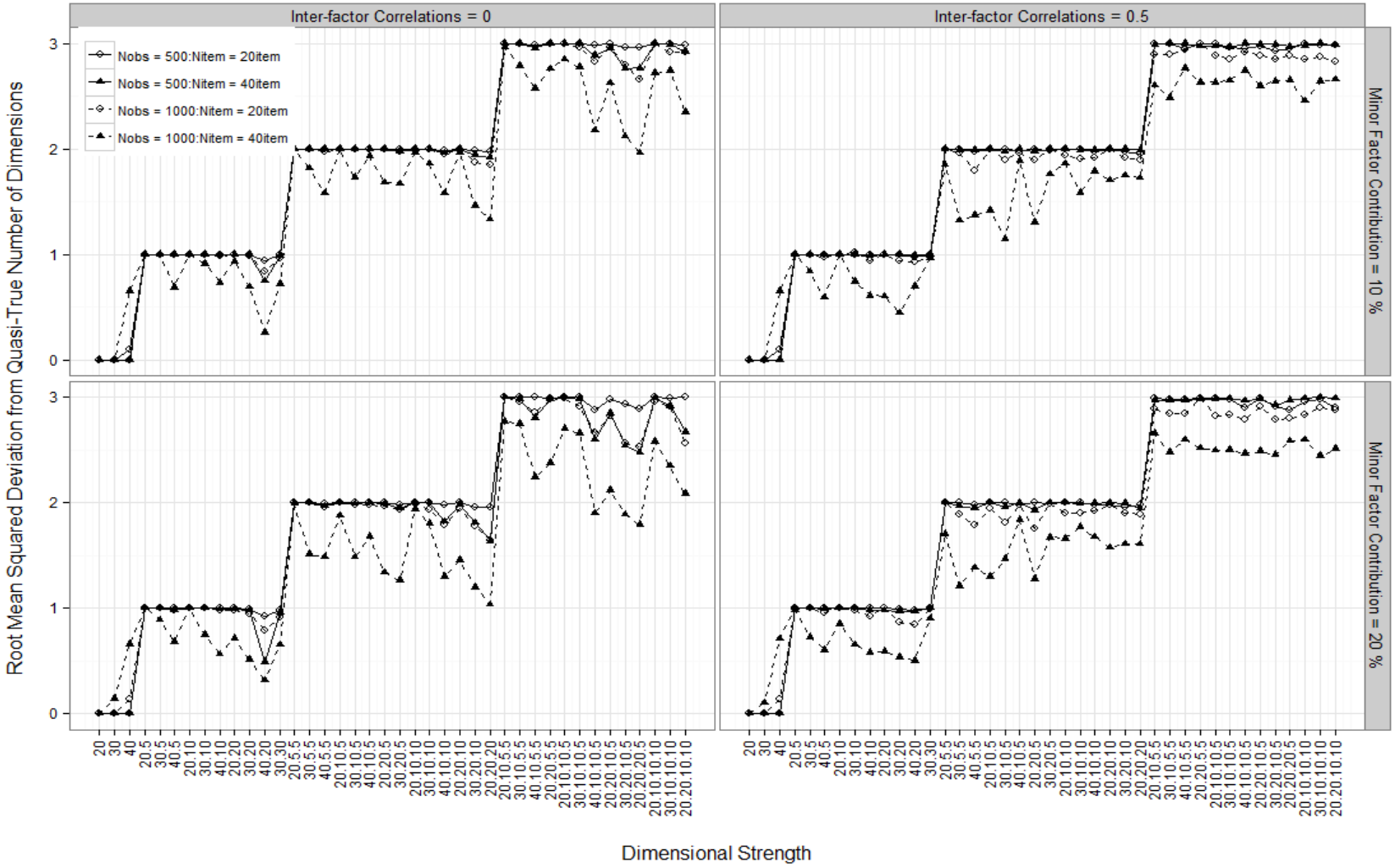


Figure 58. Root Mean Squared Deviation from the Quasi-true Number of Dimensions for **BIC based on the Mplus MLR Estimator**

very low, and BIC could not correctly identify the quasi-true number of dimensions for any replication in most conditions.

Figure 57 shows the bias with respect to the quasi-true number of dimensions for the decisions given by BIC, and indicates that BIC always tended to select the simpler models with less than the quasi-true number of dimensions in most conditions. Excluding the conditions in which the sample size was 1000 and the number of items was 40, BIC tended to select one-dimensional models almost all the time in most conditions regardless of the underlying structure. When the sample size was 1000 and the number of items was 40, BIC tended to select two-dimensional models in addition to one-dimensional models in some occasions.

Akaike Information Criterion. The results for the AIC index and corrected AIC index obtained from the Mplus MLR estimator were very similar to each other. The results for the corrected AIC index are reported and shown in Table 41 and Table 42 and in Figure 59 through Figure 61. The models with up to six dimensions were fitted using the Mplus MLR estimator, the AIC value was extracted for each model, and the model with the smallest AIC was found. In some replications, AIC was the smallest for the six-dimensional model and a decision was not reached. The proportions of no-decision replications after successfully fitting up to the six-dimensional model are reported in Table 41. The proportions were 0% for more than half of the conditions, and varied between 1% and 7% for the rest of the conditions, particularly with higher rates observed for the conditions in which the sample size was 1000, the number of items was 40, and the variance accounted for by minor factors was 20%. In some other replications, the decision was not reached due to convergence issues at some point when fitting models with one to six dimensions. The proportions of no-decision replications due to convergence issues are reported in Table 42, and varied between 0% and 61%. Higher rates were observed when the sample size was 1000 and the number of items was 40. The replications with no-decision due to convergence problems were eliminated from further analysis. For other replications in which a decision was not reached after fitting the six-dimensional model, the number of suggested dimensions was assumed to be six for further computations.

Figure 59 shows the proportion of correctly identifying the model with the quasi-true number of dimensions by the corrected AIC index across simulation conditions. When there was one major dimension, the proportions of correct decisions were above 80% for most conditions with the minor factors accounting for 10% of the variance, but between 10% and 80% for the conditions with minor factors accounting for 20% of the variance. When there were two major dimensions, the proportion corrects were all between 50% and 90% when the sample size was 1000 and the number of items was 40, and between 0% and 70% for other conditions. When there were three or four major dimensions, the proportion corrects were below 35% in some rare occasions and close to zero in most conditions. In general, the success rates for the AIC index increased as the sample size and number of items increased, and slightly decreased as the inter-factor correlations increased.

Figure 60 shows the bias with respect to the quasi-true number of dimensions for the decisions given by the corrected AIC index. When there was one major dimension in the generating structure, there was a positive bias between 0 and 1, indicating that AIC tended to select two-dimensional models for the replications in which one major dimension could not be correctly identified. When there were two major dimensions in the generating structure and the minor factors accounted for 10% of the variance, the bias was negative and between 0 and -1, particularly for the conditions where one of the two major dimensions is relatively weaker. When there were two major dimensions in the generating structure and the minor factors accounted for 20% of the variance, the bias was around zero for most of the 40-item conditions, and closer to -1 for the 20-item conditions. When there were three major dimensions in the generating structure, the bias was between -1 and -2 in most of the conditions. The only exceptions were the conditions where the sample size was 1000 and number of items was 40. In those conditions, the bias was much closer to zero and even positive when the minor factors accounted for 20% of the variance and all three major dimensions were very strong. When there were four major dimensions in the generating structure, the bias was between -2 and -3 for most of the conditions where minor factors accounted for 10% of the variance. When there were four major dimensions in the generating structure and minor factors accounted

for 20% of the variance, the bias was between 0 and -2 for most conditions when the inter-factor correlation was zero, and between -1 and -3 for most conditions when the inter-factor correlation was 0.5. In general, the corrected AIC tended to select one- or two-dimensional models when the quasi-true number of dimensions was not correctly identified, but the bias for the corrected AIC decreased as the sample size and number of items increased.

Table 41. *Proportion of Datasets in Which the Dimensionality Decision Cannot be Reached bythe Corrected Akaike Information Criterion based on the Mplus MLR Estimator*

Minor Factors	10 %				20 %			
Number of Items	20		40		20		40	
Sample Size	500	1000	500	1000	500	1000	500	1000
20	-	-	-	-	-	-	-	0.05
30	-	-	-	-	-	0.01	-	0.03
40	-	-	-	-	-	0.01	-	0.01
20.5	-	-	-	-	-	0.01	-	0.02
30.5	-	0.01	-	-	0.01	0.01	-	-
40.5	-	-	-	-	-	-	-	0.01
20.10	-	-	-	-	-	-	-	0.02
30.10	-	-	-	-	-	0.01	-	-
40.10	-	-	-	-	-	-	-	-
20.20	-	0.01	-	-	-	0.01	-	0.01
30.20	-	-	-	-	0.02	-	-	-
40.20	-	-	-	-	-	0.01	-	-
30.30	-	-	-	-	-	0.01	-	-
20.5.5	-	0.01	-	-	-	0.01	-	-
30.5.5	-	-	-	-	-	0.01	-	-
40.5.5	-	-	-	-	-	-	-	-
20.10.5	-	0.01	-	-	-	0.03	-	0.01
30.10.5	-	-	-	-	0.01	0.01	-	-
40.10.5	-	-	-	-	-	-	-	-
20.20.5	-	-	-	-	-	0.01	-	0.01
30.30.5	-	-	-	-	-	0.01	-	0.01
20.10.10	-	-	-	-	-	0.02	-	-
30.10.10	-	-	-	-	-	0.01	-	-
40.10.10	-	-	-	-	-	0.02	-	0.07
20.20.10	-	-	-	-	0.01	0.01	0.01	0.03
30.20.10	-	0.01	-	0.01	0.01	0.04	0.01	0.06
20.20.20	0.01	0.01	-	-	0.01	0.03	0.01	0.07
20.10.5.5	-	0.02	-	-	0.01	0.06	0.01	0.03
30.10.5.5	-	0.01	-	-	0.02	0.01	0.01	0.03
40.10.5.5	-	-	-	-	0.01	0.02	0.01	0.03
20.2/0.5.5	-	-	-	-	-	0.01	-	0.05
20.10.10.5	0.01	0.01	-	-	0.02	0.04	0.02	0.04
30.10.10.5	0.01	-	-	-	0.01	0.03	0.01	0.03
40.10.10.5	-	0.01	-	-	0.01	0.03	0.01	0.06
20.20.10.5	-	0.01	-	0.01	0.02	0.01	-	0.05
30.20.10.5	-	-	-	-	0.03	0.01	0.01	0.03
20.20.20.5	-	0.01	-	0.01	0.01	0.03	0.01	0.10
20.10.10.10	0.01	0.01	-	-	-	0.02	0.02	0.05
30.10.10.10	-	0.01	-	-	-	0.01	-	0.03
20.20.10.10	0.01	0.01	-	-	0.01	0.02	0.01	0.05

Note. Dashes indicate zero.

Table 42. *Proportion of Datasets in Which the Dimensionality Decision Cannot be Reached Due to Convergence Problems by the Corrected Akaike Information Criterion based on the Mplus MLR Estimator*

Minor Factors	10 %				20 %			
Number of Items	20		40		20		40	
Sample Size	500	1000	500	1000	500	1000	500	1000
20	0.04	-	0.02	-	0.05	0.05	0.01	0.04
30	-	-	-	0.01	0.01	0.03	0.03	0.05
40	0.02	-	0.02	0.03	0.02	0.02	0.06	0.06
20.5	-	-	0.01	0.01	-	0.05	0.07	0.07
30.5	-	0.02	0.01	0.02	0.01	0.03	0.04	0.06
40.5	-	0.01	0.02	0.10	0.06	0.07	0.09	0.08
20.10	0.03	0.01	-	0.06	0.05	0.04	0.04	0.18
30.10	0.02	0.06	0.13	0.29	0.07	0.19	0.33	0.53
40.10	0.09	0.11	0.16	0.15	0.17	0.13	0.26	0.10
20.20	-	0.03	0.02	0.06	0.02	0.04	0.03	0.11
30.20	0.05	0.06	0.26	0.35	0.04	0.26	0.33	0.47
40.20	0.19	0.16	0.25	0.23	0.25	0.33	0.35	0.22
30.30	0.11	0.17	0.20	0.19	0.20	0.25	0.30	0.31
20.5.5	0.04	-	-	0.16	0.05	0.02	0.02	0.32
30.5.5	0.04	0.06	0.02	0.06	0.12	0.18	0.07	0.19
40.5.5	0.01	0.02	0.10	0.11	0.05	0.05	0.08	0.13
20.10.5	-	0.02	-	0.03	0.01	0.04	0.06	0.08
30.10.5	0.02	0.11	0.32	0.61	0.09	0.16	0.35	0.61
40.10.5	0.10	0.12	0.14	0.18	0.16	0.20	0.26	0.24
20.20.5	0.06	0.08	0.23	0.15	0.06	0.12	0.24	0.20
30.30.5	0.05	0.07	0.08	0.30	0.09	0.18	0.22	0.46
20.10.10	0.10	0.05	0.21	0.37	0.07	0.20	0.40	0.47
30.10.10	0.07	0.10	0.19	0.38	0.13	0.17	0.30	0.41
40.10.10	0.09	0.05	0.06	0.10	0.15	0.10	0.06	0.09
20.20.10	0.06	0.03	0.07	0.12	0.11	0.15	0.21	0.22
30.20.10	0.03	0.06	0.04	0.12	0.05	0.14	0.11	0.16
20.20.20	0.03	0.03	0.03	0.07	0.05	0.08	0.06	0.14
20.10.5.5	0.01	0.02	0.07	0.15	0.03	0.07	0.11	0.23
30.10.5.5	0.02	0.04	0.08	0.14	0.01	0.06	0.16	0.16
40.10.5.5	0.01	0.04	0.03	0.05	0.05	0.09	0.08	0.11
20.2/0.5.5	0.06	0.05	0.09	0.15	0.12	0.13	0.18	0.21
20.10.10.5	0.02	0.02	0.09	0.19	0.02	0.08	0.12	0.25
30.10.10.5	0.03	0.03	0.09	0.12	0.06	0.05	0.07	0.12
40.10.10.5	0.06	0.05	0.05	0.07	0.08	0.09	0.08	0.11
20.20.10.5	0.02	0.04	0.09	0.21	0.03	0.10	0.13	0.20
30.20.10.5	0.03	0.07	0.05	0.12	0.12	0.10	0.11	0.15
20.20.20.5	0.03	0.07	0.03	0.11	0.07	0.13	0.09	0.14
20.10.10.10	0.01	0.05	0.10	0.17	0.04	0.09	0.08	0.19
30.10.10.10	0.03	0.04	0.06	0.12	0.04	0.10	0.08	0.13
20.20.10.10	0.04	0.02	0.05	0.19	0.08	0.11	0.12	0.16

Note. Dashes indicate zero.

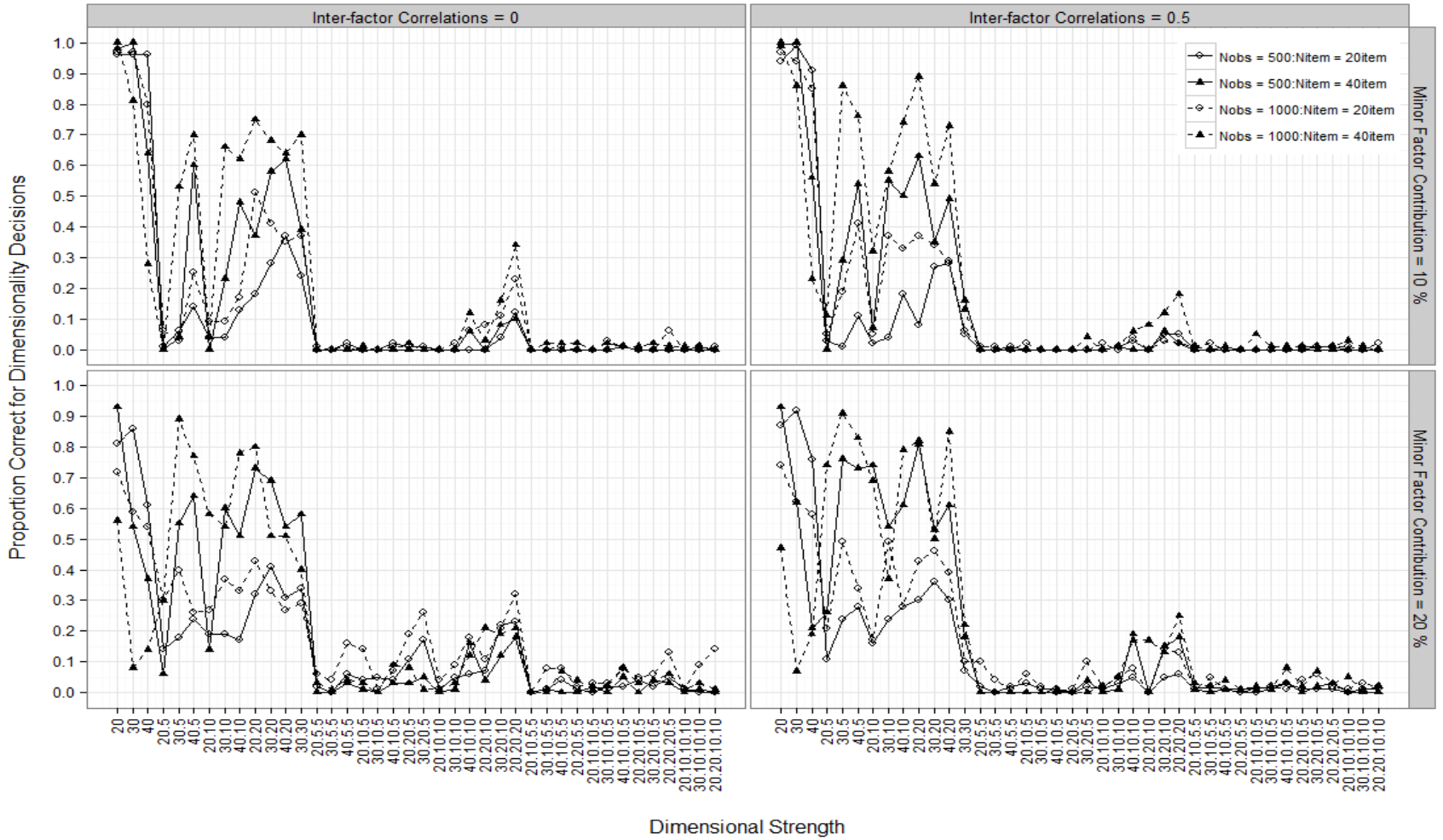


Figure 59. Proportion of Datasets in Which the Quasi-true Number of Dimensions is Correctly Identified by **the Corrected AIC based on the Mplus MLR Estimator**

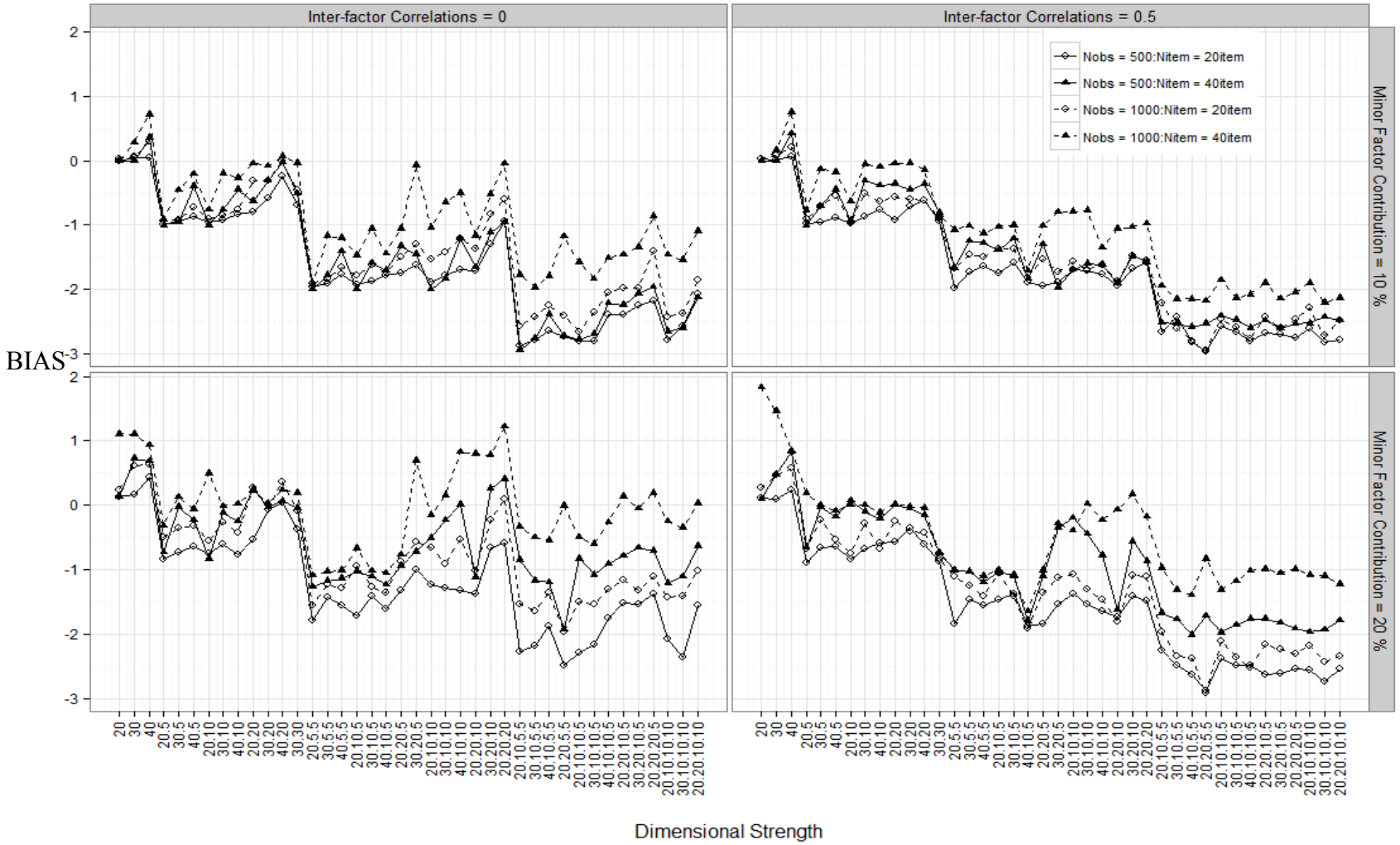


Figure 60. Bias with respect to the Quasi-true Number of Dimensions for **the Corrected AIC based on the Mplus MLR Estimator**

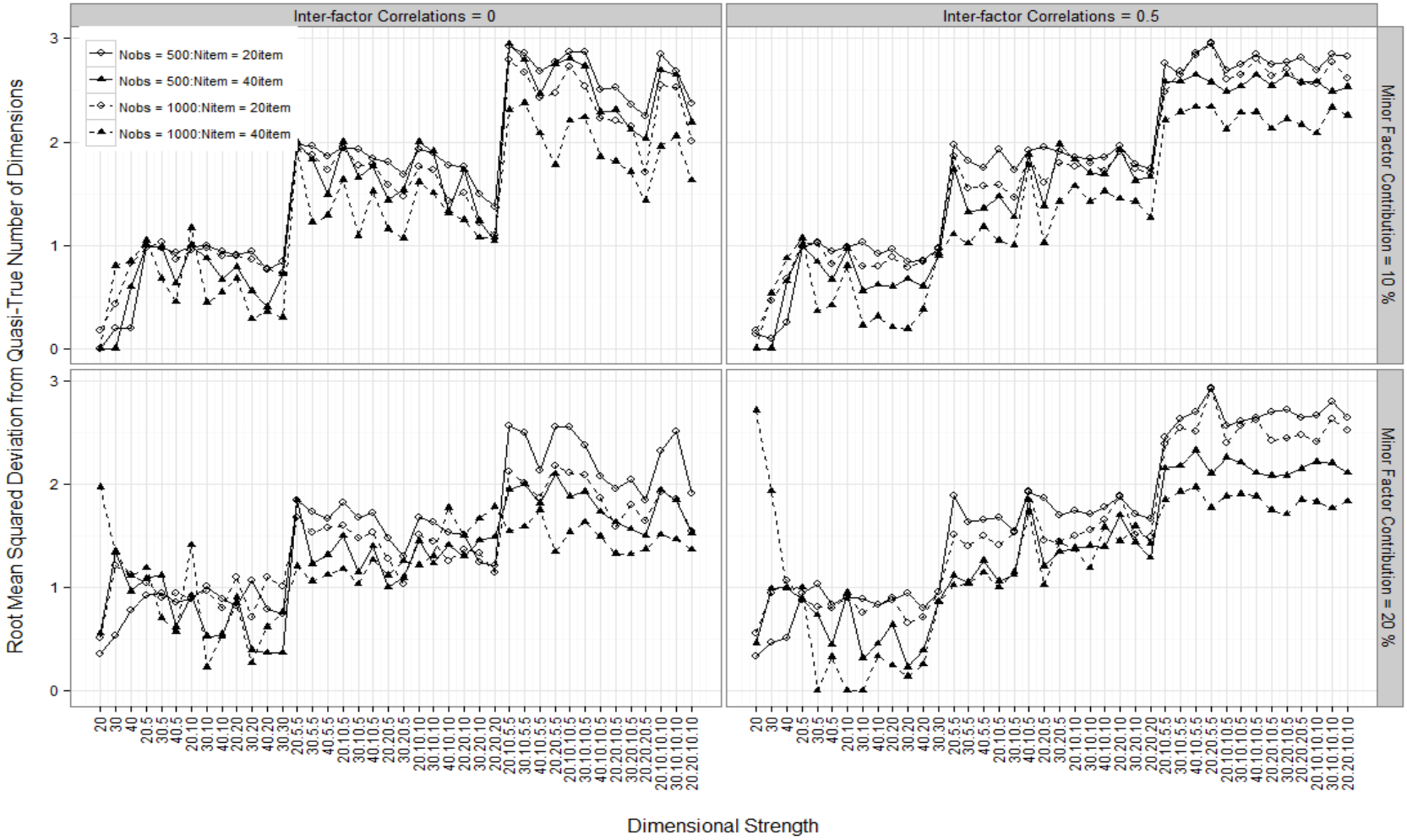


Figure 61. Root Mean Squared Deviation from the Quasi-true Number of Dimensions for **the Corrected AIC based on the Mplus MLR Estimator**

CHAPTER 5: DISCUSSION

The purpose of the current study was to examine the performance of analytical methods proposed in the literature to assess the dimensionality of latent structures underlying dichotomous item response data using both real and simulated data. In Study 1, the average, standard deviation, and range of the number of dimensions suggested by different approaches were investigated using sample datasets drawn from a very large real item response dataset treated as the population. Study 1 did not provide conclusive results because the true latent structure underlying the real data was unknown, but the results from Study 1 were informative in understanding the general behavior of each method in dimensionality assessment. Also, the results from Study 1 can be used to see how much the performances of analytical methods with the simulated data in Study 2 are consistent with the results from Study 1, which used real data. In Study 2, a comprehensive simulation study was run, and the performances of the analytical methods were evaluated using the number of major dimensions in the true generating model as a reference.

In this chapter, the contribution of the current study to the literature of dimensionality assessment is discussed. Next, some concluding suggestions about the analytical methods proposed in the literature are given for the practice of dimensionality assessment. Then, the limitations of the study and future directions for new research are discussed.

Impact and Contributions

The assumption of unidimensionality is one of the most crucial assumptions for the commonly used dichotomous IRT models in practice. A seminal paper by Hattie (1985) about assessing the unidimensionality assumption is one of the most cited papers in the field, and the literature has an extensive body of research concerned with testing the assumption of unidimensionality for dichotomous item response data. On the other hand, it is acknowledged that the assumption of unidimensionality is rarely met in practice. With the development and increasing use of multidimensional item response

theory models as well as the necessary computational tools, there is a need for analytical procedures to identify the number of latent traits with major influences on item responses. This decision process has been well emphasized in the factor analytic literature since the late 1930s, but studies in the factor analytic literature do not help much with the dimensionality assessment of dichotomous data because their focus is continuous measurement outcomes and dichotomous item responses need special treatment. In the literature, few studies go beyond the unidimensionality assumption, and the current study is one of them. This study provides additional information regarding the performances of different dimensionality assessment approaches in identifying multiple latent traits with major influences when the outcome is dichotomous.

From a methodological perspective, the current study has two main contributions. It is a well-known motto that "all models are wrong, but some of them are useful (Box & Draper, 1987, page 424)" to acknowledge model misspecification or model error in real practices. However, this fact is rarely integrated into the research design by most simulation studies, particularly in dimensionality assessment of dichotomous data. The first contribution of the current study is to acknowledge model misspecification in its research design. The presence of minor factors is not acknowledged by previous studies focusing on the dimensionality assessment of dichotomous item response data. Previous studies simulated datasets using a known latent structure with one, two, or three major dimensions and fitted the true generating models in analyzing the simulated datasets when assessing dimensionality. The model error was mostly ignored, and the performances of the dimensionality assessment approaches were investigated with such a condition that there was no model error. In the current study, the presence of minor factors was acknowledged and integrated into the designs of the two studies by following MacCallum's suggestions (2003) to examine the performance of different dimensionality assessment approaches under misspecified (imperfect) models. The second methodological contribution of the current study was to use fully complex structures when generating simulated datasets. The general tendency in previous studies was to use either simple or approximately simple structures in data generation. The current study

examined the performances of different dimensionality assessment procedures under such conditions that all items were multidimensional.

From a practical point of view, the current study provides some interesting and provoking results regarding the performances of some well-known and most commonly used practices under certain conditions. The results of the current study suggest that most of the methods proposed in the literature and available for practitioners are not necessarily useful tools in dimensionality assessment, particularly if the goal of dimensionality assessment is to identify the latent traits with major influences, when the underlying factor structure is complex and minor factors are present.

Summary Recommendations

The study examined the performances of several dimensionality assessment approaches. One of the objectives of the current study was to provide some insights regarding the accuracy of different approaches in identifying the latent traits with major influences underlying dichotomous item response data. A brief summary of conclusions about each method is given in this section.

Parallel Analysis and Revised Parallel Analysis. Both parallel analysis and revised parallel analysis are strongly recommended in the literature and seem to be the most viable alternatives among the methods that rely on eigenvalue examination. However, the results of the current study show that they perform very poorly in certain conditions and yield biased decisions regarding the number of major dimensions. Both parallel analysis and revised parallel analysis tend to select one-dimensional models in most conditions regardless of the number of major dimensions in the underlying latent structure. The poor results are actually not directly related to the methods, but due to the fact that the interpretations of the eigenvalues, particularly the first eigenvalue, are ambiguous under complex factor structures. When all items are multidimensional, the results indicate that the magnitude of the first eigenvalue is reflecting the amount of variance accounted for by the whole latent structure, not only the first latent dimension. Due to the large first eigenvalue, the remaining later eigenvalues were too small and not even larger than the random data eigenvalues when the underlying latent structure was factorially complex.

As a result of this fact, parallel analysis and revised parallel analysis tend to select one-dimensional models.

This fact was actually realized and published by Sir Thompson (1916) about a century ago. The results for the parallel analysis and revised parallel analysis in the current study are just a reiteration of his conclusions. If parallel analysis or revised parallel analysis support a one-dimensional model due to a large first eigenvalue, this does not necessarily indicate one dimension, and there may be many major latent dimensions overlapping in a complicated way. The results from Study 1 are also consistent with the results from Study 2 regarding the performance of parallel analysis and revised parallel analysis, as they also select one-dimensional models dominantly for the real datasets. Real datasets used in the current study were expected to have multidimensional natures to some unknown degree with several sub-components, but parallel analysis and revised parallel analysis dominantly favored one-dimensional models in most occasions. Similar observations also appeared in the literature. For instance, Reise and Hayiland (2005) analyzed a 25-item measure of cognitive problems and reported a first eigenvalue of 13.29 while the second eigenvalue was 1.5, which may suggest a very strong general factor. However, they also showed that seven factors could be extracted from the data and these factors were interpretable. Similarly, Smith and Reise (1998) analyzed a 23-item measure of stress reaction and reported a very large 9.59 to 0.97 ratio of the first to second eigenvalues, but they also showed that five correlated factors could be extracted and interpreted.

It seems that this fact is not well emphasized/realized in the literature when evaluating the performance of parallel analysis, and this may be due to the fact that most simulation studies used simple structures to generate data. The practitioners should be very cautious when using parallel analysis, revised parallel analysis, or any other method that relies on the magnitude of the eigenvalues in dimensionality assessment unless there is a strong justification that the underlying latent structure is perfectly simple.

DETECT. DETECT was the only procedure that provided reasonable hit rates for the conditions in which the quasi-true number of dimensions was three or four. However, it is unclear whether or not the high hit rates in those conditions are due to the systematic bias

toward selecting three- and four-dimensional models. The reason for this suspicion is the decisions given by DETECT for the conditions where there were one or two major dimensions in the true generating model. In Study 2, DETECT always tended to select about four dimensions on average for the conditions with one major dimension, and three or four dimensions on average for the conditions with two major dimensions. Similarly, when there were four major dimensions, DETECT tended to select three dimensions on average when the number of items was 20, and four dimensions on average when the number of items was 40. The results from Study 1 with real data also supported this suspicion. The average number of dimensions selected by DETECT across 500 replications was always between 3.5 and 4 for all conditions of the sampling study in Study 1, although DETECT selected five dimensions at the population level for all occasions. Therefore, there might be a systematic bias in selecting three and four dimensions for the DETECT procedure regardless of the underlying latent structure when the structure is factorially complex. Practitioners should be aware of this fact if DETECT is used for dimensionality assessment.

Chi-Square Statistics. In the current study, eight different chi-square statistics were examined. In fact, chi-square statistics are not appropriate for dimensionality assessment if the goal is to identify the major latent dimensions, because they try to find the "true" model and are very sensitive to minor factors when there is sufficient statistical power as is also shown again in the current study. However, if a practitioner is willing to use chi-square statistics in dimensionality assessment for some reason, the most viable option among the others seems to be the mean-and-variance adjusted chi-square statistics from Mplus WLS estimation (WLSMV estimator), as it provided the best hit rates in identifying the quasi-true number of dimensions and was less sensitive to minor factors given the sample sizes considered in the study.

The approximate chi-square statistic (Achi) based on NOHARM estimation is not recommended for any use, because it tends to select one-dimensional models in most conditions regardless of the underlying structure. The approximate log-likelihood ratio chi-square test (ALR) based on NOHARM estimation is also not recommended for any use, because it was not very successful in correctly identifying the model with the quasi-

true number of dimensions. Also, the NOHARM ALR statistic selected one-dimensional models almost all the time when there was not sufficient power, and showed a large amount of positive bias in most conditions when there is sufficient power indicating sensitivity to minor factors. Mean-adjusted and mean-and-variance adjusted chi-square statistics had low proportions of correctly selecting the models with the quasi-true number of dimensions, and they were more sensitive to minor factors as shown by large amounts of positive bias in many conditions.

The chi-square difference tests with adjusted and unadjusted log-likelihoods are logically different, because they test relative improvement in model fit rather than exact fit. They are also not recommended because the hit rates were low particularly for the conditions where the number of major dimensions was three or four. Also, when they could not correctly identify the quasi-true number of dimensions, the bias in given decisions was negative (under-prediction) in most conditions. The chi-square difference test with unadjusted log-likelihoods showed positive bias(over-prediction) in some conditions.

Model Fit Indices. Six different types of fit indices are considered in the current study. These are the RMSEA, CFI, and SRMR indices based on the Mplus WLS estimation, and AIC, corrected AIC, and BIC based on the Mplus FIML estimation. The performance of the CFI and SRMR were evaluated using the cut-off values of 0.95 and 0.07, respectively. For the RMSEA index, the confidence interval was used to make decisions rather than point estimates. The smallest model was selected such that the lower bound of the associated RMSEA confidence interval was smaller than 0.05.

The results of the current study suggest that using the standard cut-off values proposed in the literature for the RMSEA, CFI, and SRMR indices may not be optimal for dimensionality assessment purposes. These indices tend to select one-dimensional models in most conditions using the standard cut-off values. The results from Study 2 indicate that a smaller value for the RMSEA and SRMR indices and a larger value for the CFI index may optimize the accuracy of decisions in selecting the quasi-true number of dimensions. For instance, a cut-off value between 0.99 and 1.00 for the CFI index was required to optimize the decisions in most conditions. However, the results also suggest

that choosing an optimal cut-off value is not a straightforward process, particularly for the RMSEA and SRMR indices, and there is not a unique cut-off value that optimizes the decisions in all conditions. Optimal cut-off values for the RMSEA and SRMR indices seem to be a function of sample size, number of items, number of major dimensions in the latent structure and the variance accounted for by these major dimensions. Yet, choosing an optimal cut-off value for those indices does not guarantee high rates of accuracy in identifying the number of major dimensions. Among the CFI, RMSEA, and SRMR indices, SRMR seems to be the most suitable one for further investigation as the optimal values for the SRMR index are not influenced much by the underlying factor structure and seem to be a function of sample size and number of items.

The BIC index never correctly identified the quasi-true number of dimensions in most conditions, and tended to select one- or two-dimensional models in many occasions regardless of the underlying factor structure. This is consistent with the findings of the previous studies in the IRT literature, as they reported that BIC tends to select simpler underparameterized models in different contexts. The results for AIC and corrected AIC were very similar to each other, and provided better results than BIC in terms of accuracy and bias in selecting the quasi-true number of dimensions. Specifically, AIC seems to be relatively successful in correctly identifying the quasi-true number of dimensions when there are one or two major dimensions in the latent structure, but not very successful when there are three or four major dimensions.

Limitations and Future Research

The current study has several limitations related to the design and data analysis, and the findings should be interpreted in the light of those limitations. The first limitation of the study is the high rates of non-convergent replications in some conditions for some particular methods related to Mplus WLS and FIML estimations. In those conditions, the results are based on the converged replications and may be biased. Another limitation for the dimensionality assessment criteria related to the Mplus FIML estimation is the number of quadrature points used in the estimation. Due to the limited computational resources and time constraints, the number of quadrature points was restricted when fitting higher dimensional models. In Study 2, seven quadrature points when fitting one-

dimensional models, five quadrature points when fitting two- and three-dimensional models, four quadrature points when fitting four-dimensional models, and three quadrature points when fitting five- and six-dimensional models were used to approximate the integration for the FIML estimation. As the number of quadrature points decreases, it is likely that the approximation error increases for the estimated log-likelihood values. This may directly affect the proportion of non-convergent replications and indirectly affect the performance of any method that relies on the log-likelihood value. Future research may use as many quadrature points as possible for fitting higher dimensional models and fix the number of quadrature points across models, if possible, to reduce the effects of approximation error in log-likelihood values on the related methods.

The second limitation of the study is the type of factor structure used in generating the datasets. A complex factor structure with all items loading on all dimensions is used for the current study. Consequently, the findings of the study are limited to complex factor structures. One could argue that simpler factor structures would be more realistic and expected in certain fields such as psychology. In such conditions, some of the methods may provide better results for simple or approximately simple structures than what has been reported in the current study. An important limitation is that there is no baseline condition (perfect simple structure) to compare the results. If the amount of factor complexity could be added as another variable to manipulate in the simulation study, the findings would be more conclusive. The current study already has 640 simulation conditions, and another independent variable that manipulates the amount of factor complexity would make it difficult to implement. Future research may run some follow-up simulations with simple structures or approximately simple structures using a similar design to compare the results with the current study. It would not be surprising to see better performances for some of the methods that performed poorly in the current study.

The third limitation of the study is the response model used to generate data. For the simulation study, the way the datasets were generated was equivalent to the compensatory two-parameter normal ogive multidimensional item response model. First, a corresponding three-parameter multidimensional IRT model with the guessing

parameter included can be used for future research. But, better results are not expected since none of the multidimensional IRT software can estimate the guessing parameter. So, an additional model misspecification would make the results worse, unless the true guessing parameter values are provided by the users in data analysis. Also, future research can use non-compensatory multidimensional IRT models to generate data and study the performance of dimensionality assessment approaches under the non-compensatory multidimensional IRT models. The dimensionality assessment of non-compensatory data structures is a relatively unexplored area. Again, however, the results would be worse due to additional model misspecification, because the current software can only fit compensatory models. Finally, the sample size and the number of items manipulated in the study limit the generalizability of the findings with the sample sizes of 500 and 1000, and the number of items 20 and 40.

Conclusions

The concepts of *validity* and *validation* are the most fundamental concepts in educational and psychological testing. Messick (1995) highlighted six aspects to support the validity of a test: the content, substantive, structural, generalizability, external, and consequential aspects. Among these different aspects of validity, special attention and emphasis was given to the structural aspect or construct-related validity evidence (Embretson, 1983; Messick, 1988, 1995). Lord and Novick (1968) stated that the most important characteristic of a test is its construct validity. Similarly, Messick (1988) argued that even though the construct-related evidence may not be the whole of validity, there can be no validity without it. In the absence of evidence regarding what the test scores mean, it is not possible to judge the appropriateness, meaningfulness, and usefulness of score inferences.

The different aspects of validity are also reflected in the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999) as different types of validity evidence. Five sources of evidence are described to support the validity argument: evidence based on test content, evidence based on response processes, evidence based on internal structure, evidence based on relationships with other variables, and evidence based on the consequences of testing. As emphasized at the very beginning, several

standards have been established for test developers and test users to provide evidence regarding the internal structure of the test responses or the structural aspect of the validity argument.

Dimensionality analysis is a standard procedure for providing evidence regarding the internal structure of a set of items. There are many different approaches in the literature proposed for dimensionality analyses. The current study aimed to provide some insight for the performance of different dimensionality assessment approaches with misspecified models when the underlying latent structure was factorially complex. The current study does not provide support for a particular approach in using dimensionality assessment, since most of the methods performed poorly in most conditions; however, it provides some insight for practitioners.

REFERENCES

Ackerman, T. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement*, 113-127.

Ackerman, T. A. (2005). Multidimensional item response theory modeling. In A. Maydeu-Olivares, & J. J. McArdle, *Contemporary Psychometrics* (pp. 3-25). New Jersey: Lawrence Erlbaum Associates, Inc.

Ackerman, T. A. (April, 1988). *An explanation of differential item functioning from a multidimensional perspective.* Paper Presented at the Annual Meeting of American Educational Research Association ED306281.

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov, & F. Csaki (Ed.), *Second International Symposium on Information Theory*, (pp. 267-281). Budapest.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52*(3), 317-332.

American Educational Research Association, American Psychological Association, & National Council for Measurement in Education. (1999). *Standards for Educational and Psychological Testing.* Washington, DC: APA.

Ansley, T.M.; Forsyth, R.A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two dimensional data. *Applied Psychological Measurement*, 39-48.

Asparouhov, T., & Muthen, B. (2007). Computationally efficient estimation of multilevel high-dimensional latent variable models. *Proceedings of the 2007 Joint Statistical Meetings, Section on Statistics in Epidemiology* (pp. 2531-2535). Alexandria, VA: American Statistical Association.

Asparouhov, T., M. (2010). *Simple second order chi-square correction.* Retrieved from Mplus Technical Appendices: http://www.statmodel.com/techappen.shtml

Bentler, P. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238-246.

Bentler, P. M. (2006). EQS 6 Structural Equations Program Manual. Encino, CA: Multivariate Software, Inc.

Berger, M. F., & Knol, D. L. (December, 1990). *On the Assessment of Dimensionality in Multidimensional Item Response Theory Models.* Department of Education, University of Twente, Netherlands. Research Report 90-8. ED 329584.

Bock, D., Gibbons, R., & Muraki, E. (1988). Full information item factor analysis. *Applied Psychological Measurement*, 261-280.

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

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-Information Item Factor Analysis. *Applied Psychological Measurement, 12*(3), 261.

Bolt, D. M. (1999). Evaluating the Effects of Multidimensionality on IRT True-Score Equating. *Applied Measurement in Education, 12*(4), 383-407.

Bolt, D. M. (2005). Limited and Full Information Estimation of Item Response Theory Models. In A. Maydeu-Olivares, & J. J. McArdle, *Contemporary Psychometrics* (pp. 27-72). New Jersey: Lawrence Erlbaum Associates, Inc.

Box, G.E.P. & Draper, N.R. (1987). *Empirical Model-Building and Response Surfaces*. Wiley.

Breithaupt, K. J. (1996). A Comparison of the Approximate Chi-Square and DIMTEST in Conditions of Pseudo-Guessing and Correlated Factors. *Dissertation Abstract International, 34*(5).

Browne, M. (2000). Cross-validation Methods. *Journal of Mathematical Psychology, 44*, 108-132.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen, & J. S. Long, *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Buja, A., & Eyuboglu, N. (1992). Remarks on Parallel Analysis. *Multivariate Behavioral Research, 27*(4), 509-540.

Burnham, K., & Anderson, D. (2002). *Model selection and multi-model inference: a practical information-theoretic approach.* New York: Springer.

Cai, L., Thissen, D., & du Toit, S. (2011). IRTPRO for Windows [Computer software]. Lincolnwood, IL: Scientific Software International.

Camilli, G., Wang, M.-m., & Fesq, J. (1995). The Effects of Dimensionality on Equating the Law School Admission Test. *Journal of Educational Measurement, 32*(1), 79-96.

Carroll, J. B. (1945). The Effect of Difficulty and Chance Success on Correlations Between Items or Between Tests. *Psychometrika, 10*(1), 1-19.

Cattell, R. B. (1958). Extracting the correct number of factors in factor analysis. *Educational and Psychological Measurement, 18*(4), 791-838.

Cattell, R. B. (1966). The Scree Test For the Number of Factors. *Multivariate Behavioral Research, 1*(2), 245-276.

Cattell, R. B., & Vogelmann, S. (1977). A Comprehensive Trial of the Scree and KG Criteria for Determining the Number of Factor Scores. *The Journal of Multivariate Behavioral Research, 12*, 289-325.

Cho, S.-J., Li, F., & Bandalos, D. (2009). Accuracy of the Parallel Analysis Procedure With Polychoric Correlations. *Educational and Psychological Measurement, 69*(5), 748-759.

Comrey, A. L. (1978). Common Methodological Problems in Factor Analysis. *Journal of Consulting and Clinical Psychology, 46*, 648-659.

Cook, L. L., Dorans, N. J., Eignor, D. R., & Petersen, N. S. (April, 1983). *An Assessment of the Relationship Between the Assumption of Unidimensionality and the Quality of IRT True-Score Equating.* Paper Presented at the Annual Meeting of the American Educational Research Association, Montreal, Quebec. ED235190.

Cook, L. L., Eignor, D. R., & Taft, H. L. (1988). A Comparative Study of the Effects of Recency of Instruction on the Stability of IRT and Conventional Item Parameter Estimates. *Journal of Educational Measurement, 25*(1), 31-45.

Coovert, M. D., & McNelis, K. (1988). Determining the number of common factors in factor analysis: a review and a program. *Educational and Psychological Measurement, 48*, 687-692.

Crawford, A. V., Green, S. B., Levy, R., Lo, W.-J., Scott, L., Svetina, D., & Thompson, M. S. (2010). Evaluation of Parallel Analysis Methods for Determining the Number of Factors. *Educational and Psychological Measurement, 70*(6), 885-901.

Crawford, C. B. (1975). Determining the Number of Interpretable Factors. *Psychological Bulletin, 82*(2), 226-237.

Crawford, C. B., & Koopman, P. (1979). Note: Inter-rater reliability of scree test and mean square ratio test of number of factors. *Perceptual and Motor Skills, 49*, 223-226.

Cudeck, R., & Henly, S. (1991). Model selection in covariance structures analysis and the "problem" of sample size: a clarification. *Psychological Bulletin, 109*(3), 512-519.

De Ayala, R. J. (1992). The Influence of Dimensionality on Cat Ability Estimation. *Educational and Psychological Measurement, 52*(3), 513-527.

De Champlain, A. F. (1993). Assessing Test Dimensionality Using Two Approximate Chi-Square Statistics. *Dissertation Abstract International, 54*(6).

De Champlain, A. F. (1996). The Effect of Multidimensionality on IRT True-Score Equating for Subgroups of Examinees. *Journal of Educational Measurement, 33*(2), 181-201.

De Champlain, A., & Gessaroli, M. E. (1998). Assessing the Dimensionality of Item Response Matrices with Small Sample Sizes and Short Test Lengths. *Applied Measurement in Education, 11*(3), 231-253.

de Winter, J., Dodou, D., & Wieringa, P. (2009). Exploratory Factor Analysis with Small Sample Sizes. *Multivariate Behavioral Research, 44*, 147-181.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological), 39*(1), 1-38.

Dinno, A. (2009). Exploring the Sensitivity of Horn's Parallel Analysis to the Distributional Form of Random Data. *Multivariate Behavioral Research, 44*, 362-388.

Dorans, N. J., & Kingston, N. M. (1985). The Effects of Violations of Unidimensionality on the Estimation of Item and Ability Parameters and on Item Response Theory Equating of the GRE Verbal Scale. *Journal of Educational Measurement, 22*(4), 249-262.

Drasgow, F., & Lissak, R. I. (1983). Modified Parallel Analysis: A Procedure for Examining the Latent Dimensionality of Dichotomously Scored Item Responses. *Journal of Applied Psychology, 68*(3), 363-373.

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement, 7*, 189-199.

Embretson, S. (1983). Construct Validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*(1), 179-197.

Emretson, S., & Reise, S. (2000). *Item response theory for psychologists.* Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(3), 272-299.

Fava, J. L., & Velicer, W. F. (1992). The effects of overextraction on factor and component analysis. *Multivariate Behavioral Research, 27*(3), 387-415.

Fava, J. L., & Velicer, W. F. (1996). The effects of underextraction in factor and component analyses. *Educational and Psychological Measurement, 56*(6), 907-929.

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

Finch, H., & Habing, B. (2007). Performance of DIMTEST- and NOHARM-Based Statistics for Testing Unidimensionality . *Applied Psychological Measurement, 31*(4), 292-307.

Finch, H., & Monahan, P. (2008). A Bootstrap Generalization of Modified Parallel Analysis for IRT Dimensionality Assessment. *Applied Measurement in Education, 21*, 119-140.

Finch, H., Stage, K., & Monahan, P. (2008). Comparison of Factor Simplicity Indices for Dichotomous Data- DETECT R, Bentler's Simplicity Index, and the Loading Simplicity Index. *Applied Measurement in Education*, 41-64.

Folk, V. G., & Green, B. F. (1989). Adaptive Estimation When the Unidimensionality Assumption of IRT is Violated. *Applied Psychological Measurement, 4*, 373-389.

Ford, J. K., MacCallum, R. C., & Tait, M. (1986). The application of exploratory factor analysis in applied psychology: a critical review and analysis. *Personnel Psychology, 39*(2), 291-314.

Fraser, C., & McDonald, R. P. (1988). NOHARM: Least Squares Item Factor Analysis. *Multivariate Behavioral Research, 23*, 267-269.

Froelich, A. G. (2000). Assessing Unidimensionality of Test Items and Some Asymptotics of Parametric Item Response Theory. *Dissertation Abstract International, 61*(10).

Froelich, A. G., & Stout, W. F. (2003). A New Bias Correction Method for the DIMTEST Procedure. *Unpublished Manuscript*. Retrieved 5 10, 2011, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.8.8941&rep=rep1&type=pdf

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2011). Performance of Velicer's Minimum Average Partial Factor Retention Method with Categorical Variables. *Educational and Psychological Measurement, 71*(3), 551-570.

Gessaroli, M. E., De Champlain, A. F., & Folske, J. C. (March, 1997). *Assessing Dimensionality Using a Likelihood-Ratio Chi-Square Test Based on a Non-linear Factor Analysis of Item Response Data.* Paper Presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Gessaroli, M., & De Champlin, A. (1996). Using an approximate chi-square statistic to test the number of dimensions underlying the responses to a set of items. *Journal of Educational Measurement*, 157-179.

Ghosh, J., & Tapas, S. (2001). Model selection: an overview. *Current Science, 80*(9), 1135-1144.

Gibbons, R., & Hedeker, D. (1992). Full information item bi-factor analysis. *Psychometrika, 57*(3), 423-436.

Glorfeld, L. W. (1995). An Improvement on Horn's Parallel Analysis Methodology for Selecting the Correct Number of Factors to Retain. *55*(3), 377-393.

Gorsuch, R. L. (1983). *Factor Analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Green, S. B., Levy, R., Thompson, M. S., Lu, M., & Lo, W. (2012). A Proposed solution to the problem with using completely random data to assess the number of factors with parallel analysis. *Educational and Psychological Measurement, 72*(3), 357-374.

Guttman, L. (1954). Some necessary and sufficient conditions for common factor analysis. *Psychometrika, 19*, 149-161.

Hakstian, A. R., & Muller, V. J. (April, 1972). *Some Empirical Findings Concerning the Number of Factors Problem.* Paper presented at the annual meeting of American Educational Research Association, Chicago, IL. (ERIC Document Reproduction Service No. ED061270) .

Hakstian, A., Rogers, W., & Cattell, R. (1982). The behavior of number-of-factor rules with simulated data. *Multivariate Behavioral Research, 17*, 193-219.

Hambleton, R., & Rovinelli, R. (1986). *Assessing the dimensionality of a set of test items. Applied Psychological Measurement, 10*(3), 287-302.

Harrison, D. A. (1986). Robustness of IRT Parameter Estimation to Violations of the Unidimensionality Assumption. *Journal of Educational and Behavioral Statistics, 11*(2), 91-115.

Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research*, 49-78.

Hattie, J. (1985). Methodology Review: Assessing Unidimensionality of Tests and Items. *Applied Psychological Measurement, 9*(2), 139-164.

Hattie, J., Krakowski, K., Rogers, H. J., & Swaminathan, H. (1996). An Assessment of Stout's Index of Essential Unidimensionality. *Applied Psychological Measurement, 20*(1), 1-14.

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor Retention Decisions in Exploratory Factor Analysis: A Tutorial on Parallel Analysis. *Organizational Research Methods, 7*(2), 191-205.

Henson, R. K., & Roberts, J. K. (February, 2001). *A Meta-Analytic Review of Exploratory Factor Analysis Reporting Practices in Published Research.* Paper Presented at the Annual Meeting of the Southwest Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. ED449227).

Hong, S. (1999). Generating correlation matrices with model error for simulation studies in factor analysis: A combination of the Tucker-Koopman-Linn model and Wijsman's algorithm. *Behavior Research Methods: Instruments & Computers, 31*(4), 727-730.

Horn, J. L. (1965). A Rationale and Test for the Number of Factors in Factor Analysis. *Psychometrika, 30*(2), 179-185.

Hu, L., & Bentler, P. (1999). Cutoff criterion for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Hu, L., & Bentler, P. M. (1999). Cutoff criterion for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Humphreys, L. G. (1964). Number of cases and number of factors: an example where N is very large. *Educational and Psychological Measurement, 24*, 457-466.

Humphreys, L. G., & Ilgen, D. R. (1969). Note on a Criterion for the Number of Common Factors. *Educational and Psychological Measurement, 29*(3), 571-578.

Humphreys, L. G., & Montanelli, R. G. (1975). An Investigation of the Parallel Analysis Criterion for Determining the Number of Common Factors. *Multivariate Behavioral Research*, 193-205.

Joreskog, K. G. (1962). On the Statistical Treatment of Residuals in Factor Analysis. *Psychometrika, 27*(4), 335-354.

Jurs, S., Zoski, K., & Mueller, R. (April, 1993). *Using Linear Regression to Determine the Number of Factors to Retain in Factor Analysis and the Number of Issues to Retain in Delphi Studies and Other Surveys.* Paper Presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA. (ERIC Document Reproduction Service No. ED361344).

Kaiser, H. F. (1960). The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement, 20*(1), 141-151.

Kaiser, H. F., & Hunka, S. (1973). Some Empirical Results with Guttman's Stronger Lower Bound for the Number of Common Factors. *Educational and Psychological Measurement, 33*(1), 99-102.

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

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

Kim, H. R. (1996). A New Index of Dimensionality - DETECT. *J. Korea Soc. Math. Educ. Ser. B: Pure Appl. Math., 3*(2), 141-153.

Kim, H.-R. (1995). New Techniques for the Dimensionality Assessment of Standardized Test Data. *Dissertation Abstracts International, 55*(12 (AAT 9512427)).

Kirisci, L., & Hsu, T. (1995). The Robustness of BILOG to Violations of the Assumption of Unidimensionality of Test Items and Normality of Ability Distributions. *Paper presented at annual meeting of the National Council on Measurement in Education* . San Francisco.

Kirisci, L., Hsu, T.-c., & Yu, L. (2001). Robustness of Item Parameter Estimation Programs to Assumptions of Unidimensionality and Normality. *Applied Psychological Measurement, 25*(2), 146-162.

Kline, R. B. (2005). *Principles and Practice of Structural Equation Modeling* (Second ed.). New York: The Guilford Press.

Knol, D. L., & Berger, M. P. (1991). Empirical Comparison Between Factor Analysis and Multidimensional Item Response Models. *Multivariate Behavioral Research, 26*(3), 457-477.

Lau, C.-M. A. (1997). Robustness of a Unidimensional Computerized Mastery Testing Procedure With Multidimensional Testing Data. *Dissertation Abstract International, 57*(7).

Lautenschlager, G. J. (1989). A Comparison of Alternatives to Conducting Monte Carlo Analyses for Determining Parallel Analyses Criteria. *Multivariate Behavioral Research, 24*(3), 365-395.

Levy, R., & Svetina, D. (2010). A Generalized Dimensionality Discrepancy Measure for Dimensionality Assessment in Multidimensional Item Response Theory. *British Journal of Mathematical and Statistical Psychology, 64*, 208-232.

Li, F., Cohen, A. S., Kim, S.-H., & Cho, S.-J. (2009). Model selection methods for mixture dichotomous IRT models. *Applied Psychological Measurement, 33*(5), 353-373.

Linn, R. L., & Harnisch, D. L. (1981). Interactions between Item Content and Group Membership on Achievement Test Items. *Journal of Educational Measurement, 18*(2), 109-118.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, New Jersey: Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores.* The United States of America: Information Age Publishing Inc.

Lorenzo-Seva, U., Timmerman, M. E., & Kiers, H. A. (2011). The Hull method for selecting the number of common factors. *Multivariate Behavioral Research, 46*(2), 340-364.

MacCallum, R. C. (2003). Working with Imperfect Models. *Multivariate Behavioral Research, 38*(1), 113-139.

MacCallum, R. C., & Tucker, L. R. (1991). Representing Sources of Error in the Common-Factor Model: Implications for Theory and Practice. *Quantitative Methods in Psychology, 109*(3), 502-511.

MacCallum, R. C., Tucker, L. R., & Briggs, N. E. (2001). An alternative perspective on parameter estimation in factor analysis and related methods. In R. Cudeck, S. du Toit, & D. Sorbom, *Structural Equation Modelin: Present and Future* (pp. 39-57). Lincolnwood, IL: SSI.

MacCallum, R. C., Widaman, K. F., Preacher, K. J., & Hong, S. (2001). Sample size in factor analysis: The role of model error. *Multivariate Behavioral Research, 36*, 611-637.

Maydeu-Olivares, A. (2001). Multidimensional Item Response Theory Modeling of Binary Data: Large Sample Properties of NOHARM Estimates. *Journal of Educational and Behavioral Statistics, 26*(1), 51-71.

McDonald, R. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology, 34*, 100-117.

McDonald, R. P. (1967). Nonlinear Factor Analysis. *Psychometric Monographs, No. 15*.

McDonald, R. P. (1999). *Test Theory: A Unified Treatment.* New Jersey: Lawrence Erlbaum Associates, Inc.

Meng, X.-L., & Schilling, S. (1996). Fitting Full-Information Item Factor Models and an Empirical Investigation of Bridge Sampling. *Journal of American Statistical Association, 91*(435), 1254-1267.

Messick, S. (1988). The Once and Future Issues of Validity: Assessing the Meaning and Consequences of Measurement. In H. Wainer, & H. I. Braun, *Test Validity* (pp. 33-46). New Jersey: Lawrence Erlbaum Associates.

Messick, S. (1995). Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry Into Score Meaning. *American Psychologist, 50*(9), 741-749.

Mulaik, S. A. (2010). *Foundations of Factor Analysis* (2nd ed.). Chapman & Hall/CRC.

Mumford, K. R., Ferron, J. M., Hines, C. V., Hogarty, K. Y., & Kromney, J. D. (April. 2003). *Factor Retention in Exploratory Factor Analysis: A Comparison of Alternative Methods.* Paper Presented at the Annual Meeting of the American Educational Research Association, Chicago, IL. (ERIC Document Reproduction Service No. ED476430).

Muthén, L., & Muthén, B. (1998-2010). Mplus User's Guide. Sixth Edition. Los Angeles, CA: Muthén & Muthén.

Nandakumar, R. (April, 1991). *Assessing Dimensionality of a Set of Items: Comparison of Different Approaches.* Paper Presented at the Annual Meeting of American Educational Research Association, Chicago, IL.

Nandakumar, R., & Stout, W. (1993). Refinements of Stout's Procedure for Assessing Latent Trait Unidimensionality. *Journal of Educational Statistics, 18*(1), 41-68.

Nandakumar, R., & Stout, W. (1993). Refinements of Stout's Procedure for Assessing Latent Trait Unidimensionality. *Journal of Educational and Behavioral Statistics*, 41-68.

Nandakumar, R., & Yu, F. (1996). Empirical Validation of DIMTEST on Nonnormal Ability Distributions. *Journal of Educational Measurement, 33*(3), 355-368.

Nandakumar, R., Yu, F., & Zhang, Y. (2011). A Comparison of Bias Correction Adjustments for the DETECT Procedure. *Applied Psychological Measurement, 35*(2), 127-144.

Nasser, F., Benson, J., & Wisenbaker, J. (2002). The Performance of Regression-Based Variations of the Visual Scree for Determining the Number of Common Factors. *Educational and Psychological Measurement, 62*(3), 397-419.

Nichol, P. E. (2011). Recovery of Multidimensional Item Response Theory Data Structures Using Multivariate Analyses. *Dissertation Abstract International, 72*(8).

Oshima, T. C., & Miller, M. D. (1990). Multidimensionality and IRT- Based Item Invariance Indexes: The Effect of Between- Group Variation in Trait Correlation. *Journal of Educational Measurement, 27*(3), 273-283.

Patarapichayatham, C., Kamata, A., & Kanjanawasee, S. (2011). Evaluation of model selection strategies for cross-level two-way differential item functioning analysis. *Educational and Psychological Measurement, 72*(1), 44-51.

Piccone, A. V. (2009). A comparison of three computational procedures for solving the number of factors problem in exploratory factor analysis. *Dissertation Abstracts International, 71*(4).

Preacher, K. J., & MacCallum, R. C. (2002). Exploratory factor analysis in behavior genetic research: Determinants of good recovery. *Behavior Genetics, 32*, 153-161.

Preacher, K. J., Zhang, G., Kim, C., & Mels, G. (2013). Choosing the optimal number of factors in exploratory factor analysis: A model selection perspective. *Multivariate Behavioral Research, 48*(1), 28-56.

Preinerstorfer, D., & Formann, A. K. (2012). Parameter recovery and model selection in mixed Rasch models. *British Journal of Mathematical and Statistical Psychology, 65*(2), 251-262.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25*, 111-163.

Reckase, M. D. (1979). Unifactor Latent Trait Models Applied to Multifactor Tests: Results and Implications. *Journal of Educational Statistics, 4*(3), 207-230.

Reckase, M. D. (2009). *Multidimensional Item Response Theory.* New York: Springer.

Reckase, M. D. (April, 1990). *Unidimensional Data from Multidimensional Tests and Multidimensional Data From Unidimensional Tests.* Paper Presented at the Annual Meeting of the American Educational Research Association, Boston. ED 318758.

Reckase, M. D., & McKinley, R. L. (1982, July). *Some Latent Trait Theory in a Multidimensional Latent Space.* Paper Presented at the Item Response Theory and Computerized Adaptive Testing Conference, Wayzata, MN. ED264265.

Reise, S., & Hayiland, M. (2005). Item response theory and the measurement of clinical change. *Journal of Personality Assessment, 84*, 228-238.

Reise, S., Morizot, J., & Hays, R. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research, 16*, 19-31.

Revelle, W., & Rocklin, T. (1979). Very Simple Structure: An Alternative Procedure for Estimating the Optimal Number of Factors. *Multivariate Behavioral Research, 14*, 403-414.

Roussos, L. A., & Ozbek, O. Y. (2006). Formulation of the DETECT Population Parameter and Evaluation of DETECT Estimator Bias. *Journal of Educational Measurement, 43*(3), 215-243.

Schilling, S., & Bock, R. D. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika, 70*(3), 1-23.

Seraphine, A. (2000). The Performance of DIMTEST When Latent Trait and Item Difficulty Distributions Differ. *Applied Psychological Measurement*, 82-94.

Seraphine, A. E. (1994). A power study of three procedures for the assessment of unidimensionality. *Dissertation Abstract International, 56*(11).

Smith, L., & Reise, S. (1998). Gender differences on negative affectivity: an IRT study of differential item functioning on the multidimensional personality questionnaire stress reaction scale. *Journal of Personality and Social Psychology, 75*, 1350-1362.

Spencer, S. G. (2004). The Strength of Multidimensional Item Response Theory in Exploring Construct Space That is Multidimensional and Correlated. *Unpublished Doctoral Dissertation*. Retrieved 4 12, 2011, from http://contentdm.lib.byu.edu/ETD/image/etd646.pdf

Steiger, J. H. and Lind, J. (1980) *Statistically-based tests for the number of common factors*. Paper presented at the Annual Spring Meeting of the Psychometric Society, Iowa City.

Stocking, M. L., & Eignor, D. R. (December, 1986). *The Impact of Different Ability Distributions on IRT Preequating.* Educational Testing Service, Princeton, New Jersey. Research Report 143. ED 281864.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, 589-617.

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55*(2), 293-325.

Stout, W., Nandakumar, R., & Habing, B. (1996). Analysis of Latent Dimensionality of Dichotomously and Polytomously Scored Test Data. *Behaviormetrika, 23*(1), 37-65.

Svetina, D. (2011). Assessing Dimensionality in Complex Data Structures: A Performance Comparison of DETECT and NOHARM Procedures. *Dissertation Abstract International, 72*(7).

Tate, R. (2003). A Comparison of Selected Empirical Methods for Assessing the Structure of Responses to Test Items. *Applied Psychological Measurement, 27*(3), 159-203.

Tate, R. (2009). Test Dimensionality. In G. Tindal, & T. M. Haladyna, *Large Scale Assessment Programs for All Students: Validity, Technical adequacy, and Implementation* (pp. 155-181). New Jersey: Lawrence Erlbaum Associates, Inc.

Thompson, B., & Daniel, L. G. (1996). Factor analytic evidence for the construct validity of scores: a historical overview and some guidelines. *Educational and Psychological Measurement, 56*(2), 197-208.

Thomson, G. H. (1916). A hierarchy without a general factor. *British Journal of Psychology, 8*(3), 270-393.

Tran, U., & Forman, A. (2009). Performance of Parallel Analysis in Retrieving Unidimensionality in the Presence of Binary Data. *Educational and Psychological Measurement*, 50-61.

Tucker, L. R., Koopman, R. F., & Linn, R. L. (1969). Evaluation of factor analytic research procedures by means of simulated correlation matrices. *Psychometrika, 34*(4).

Turner, N. E. (1998). The Effect of Common Variance and Structure Pattern on Random Data Eigenvalues: Implications for the Accuracy of Parallel Analysis. *Educational and Psychological Measurement, 58*(4), 541-568.

Velicer, W. F. (1976). Determining the Number of Components from the Matrix of Partial Correlations. *Psychometrika, 41*(3), 321-327.

Velicer, W. F., Eaton, C. A., & Fava, J. L. (2000). Construct explication through factor or component Analysis: a review and evaluation of alternative procedures for determining the number of factors or components. In R. D. Goffin, & E. Helmes, *Problems and Solutions in Human Assessment.* New York: Springer.

Walker, C. M., Azen, R., & Schmitt, T. (2006). Statistical Versus Substantive Dimensionality: The Effect of Distributional Differences on Dimensionality Assessment Using DIMTEST. *Educational and Psychological Measurement, 66*(5), 721-738.

Waller, N. (2001). MicroFACT 2.0: A microcomputer factor analysis program for ordered polytomous data and mainframe size problems [Computer software and manual]. St. Paul, MN: Assessment Systems Corporation.

Wang, M. (April, 1986). *Fitting a Unidimensional Model to Multidimensional Item Response Data.* Paper Presented at the ONR Contractors Conference. Gatlinburg, TN: To appear as ONR Technical Report 87-1. University of Iowa.

Way, W. D., Ansley, T. N., & Forsyth, R. A. (1988). The Comparative Effects of Compensatory and Noncompensatory Two-Dimensional Data on Unidimensional IRT Estimates. *Applied Psychological Measurement, 12*(3), 239-252.

Weiss, D. J. (1971). Further Considerations in Applications of Factor Analysis. *Journal of Counseling Psychology, 1*, 85-92.

Weng, L.-J., & Cheng, C.-P. (2005). Parallel Analysis with Unidimensional Binary Data. *Educational and Psychological Measurement, 65*(5), 697-716.

Western, B. (1999). Bayesian analysis for sociologists: an introduction. *Sociological Methods & Research, 28*(1), 7-34.

Whittaker, T. A., Chang, W., & Dodd, B. G. (2012). The performance of IRT model selection methods within mixed-format tests. *Applied Psychological Measurement, 36*(3), 159-180.

Wood, J. M., Tataryn, D. J., & Gorsuch, R. L. (1996). Effects of under- and overextraction on principal axis factor analysis with varimax rotation. *Psychological Methods, 1*(4), 354 - 365.

Wood, R., Wilson, D., Gibbons, R., Schilling, S., Muraki, E., & Bock, D. (2003). TESTFACT[Computer software and manual]. In M. du Toit, *IRT from SSI.* Lincolnwood, IL: Scientific Software International.

Wothke, W. (1993). Nonpositive Definite Matrices in Structural Modeling. In K. Bollen, & S. Long, *Testing Structural Equation Models.* Newbury Park, CA: Sage.

Yu, C. (2002). *Evaluating cut-off criteria of model fit indices for latent variable models with binary and continuous outcomes.* Doctoral Dissertation, University of California. Retrieved March 6, 2013, from http://statmodel2.com/download/Yudissertation.pdf

Zeng, J. (2010). Development of a Hybrid Method for Dimensionality Identification Incorporating an Angle-Based Approach. *Dissertation Abstract International, 71*(5).

Zhang, J. (1996). Some fundamental issues in item response theory with applications. *Dissertation Abstract International, 57*(11).

Zhang, J., & Stout, W. (1999). The Theoretical DETECT Index of Dimensionality and Its Application to Approximate Simple Structure. *Psychometrika, 64*(2), 213-249.

Zoski, K. W., & Jurs, S. (1996). An Objective Counterpart to the Visual Scree Test for Factor Analysis: The Standard Error Scree. *Educational and Psychological Measurement, 56*(3), 443-451.

Zucchini, W. (2000). An introduction to model selection. *Journal of Mathematical Psychology, 44*, 41-61.

Zumbo, B. D. (2007). Validity: Foundational Issues and Statistical Methodology. In C. R. Rao, & S. Sinharay, *Handbook of Statistics: Psychometrics* (Vol. 26, pp. 45-80). Amsterdam: Elsevier.

Zwick, W. R., & Velicer, W. F. (1982). Factors Influencing Four Rules for Determining the Number of Components to Retain. *Multivariate Behavioral Research, 17*, 253-269.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of Five Rules for Determining the Number of Components to Retain. *Psychological Bulletin, 99*(3), 432-442.

APPENDIX A: Running Average Plots for Study 1 and Study 2

Figure 1A. Running Average First Eigenvalue across Four Subtests

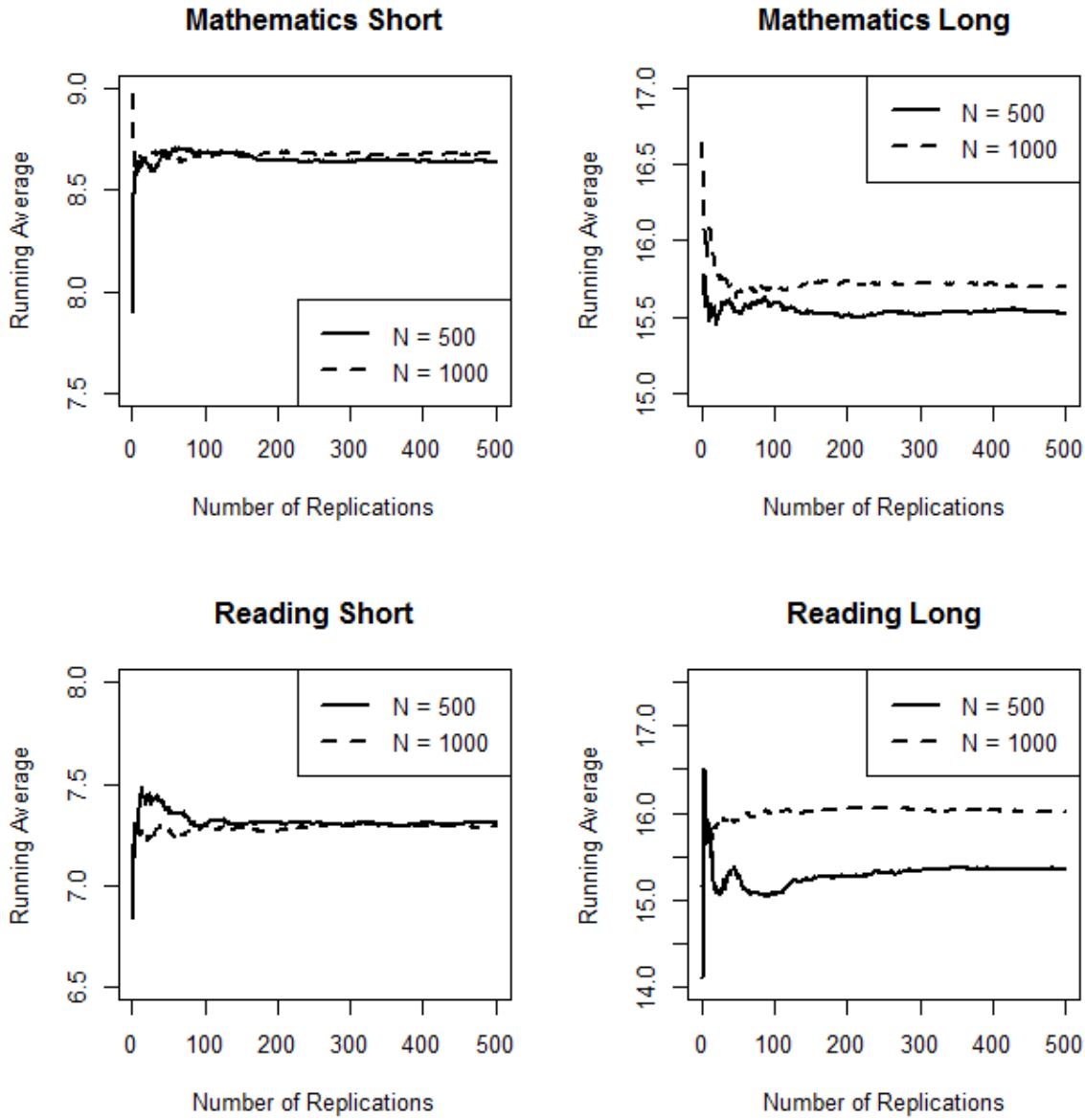


Figure 2A. Running Average Second Eigenvalue across Subtests

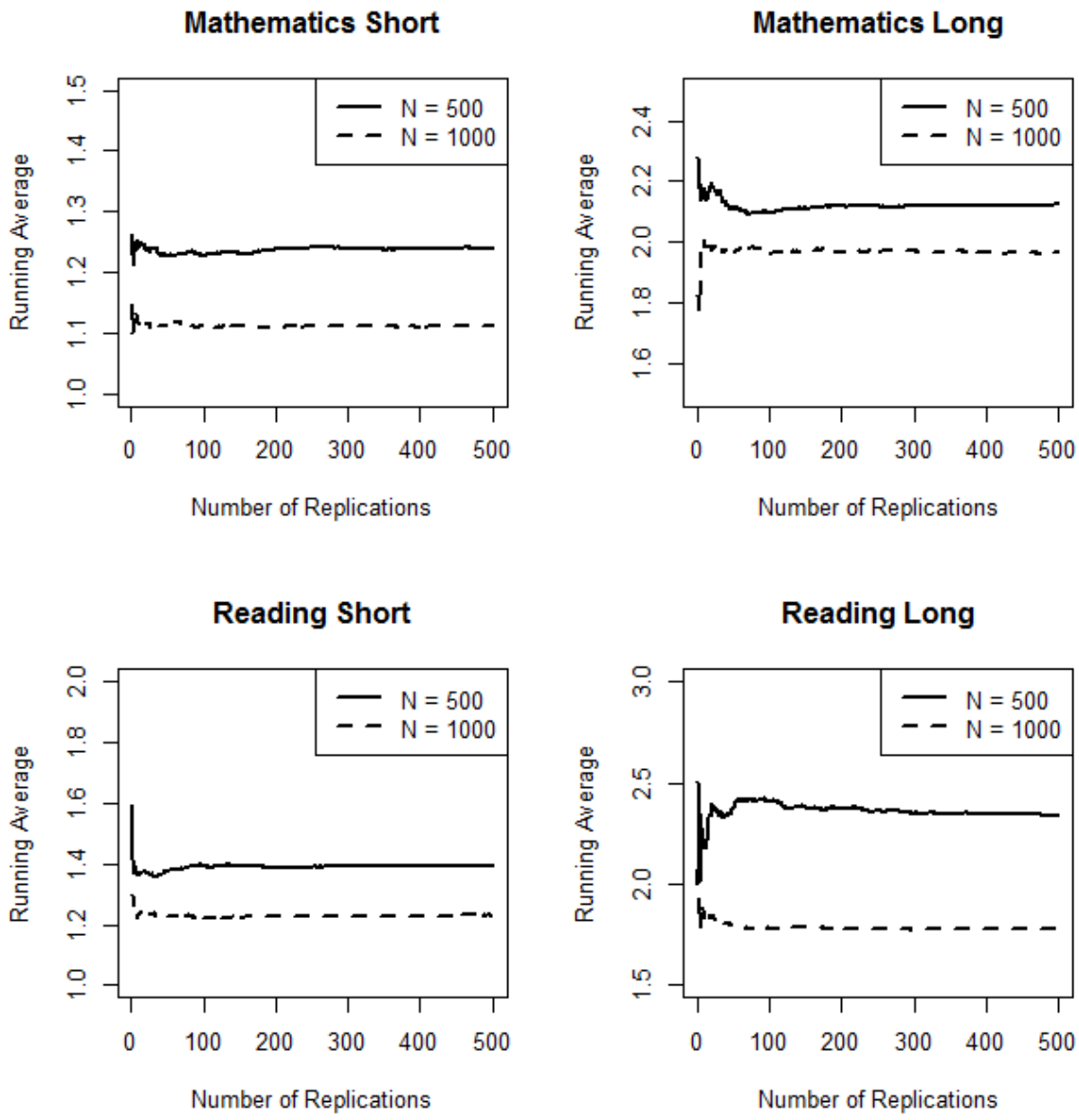


Figure 3A. Running Average Third Eigenvalue across Four Subtests

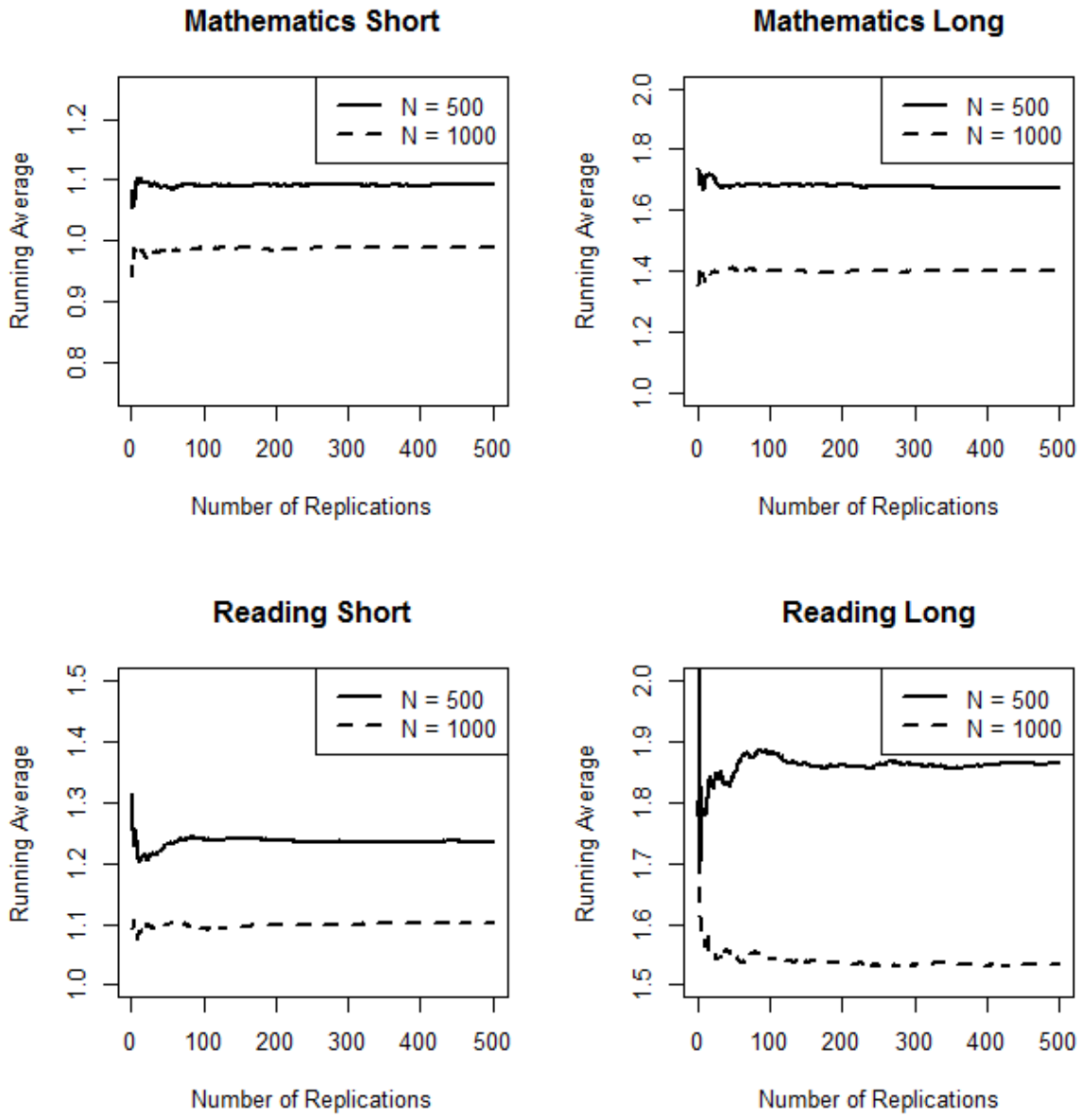


Figure 4A. Running Average RMSEA (WLSM) for One-dimensional Model across Four Subtests

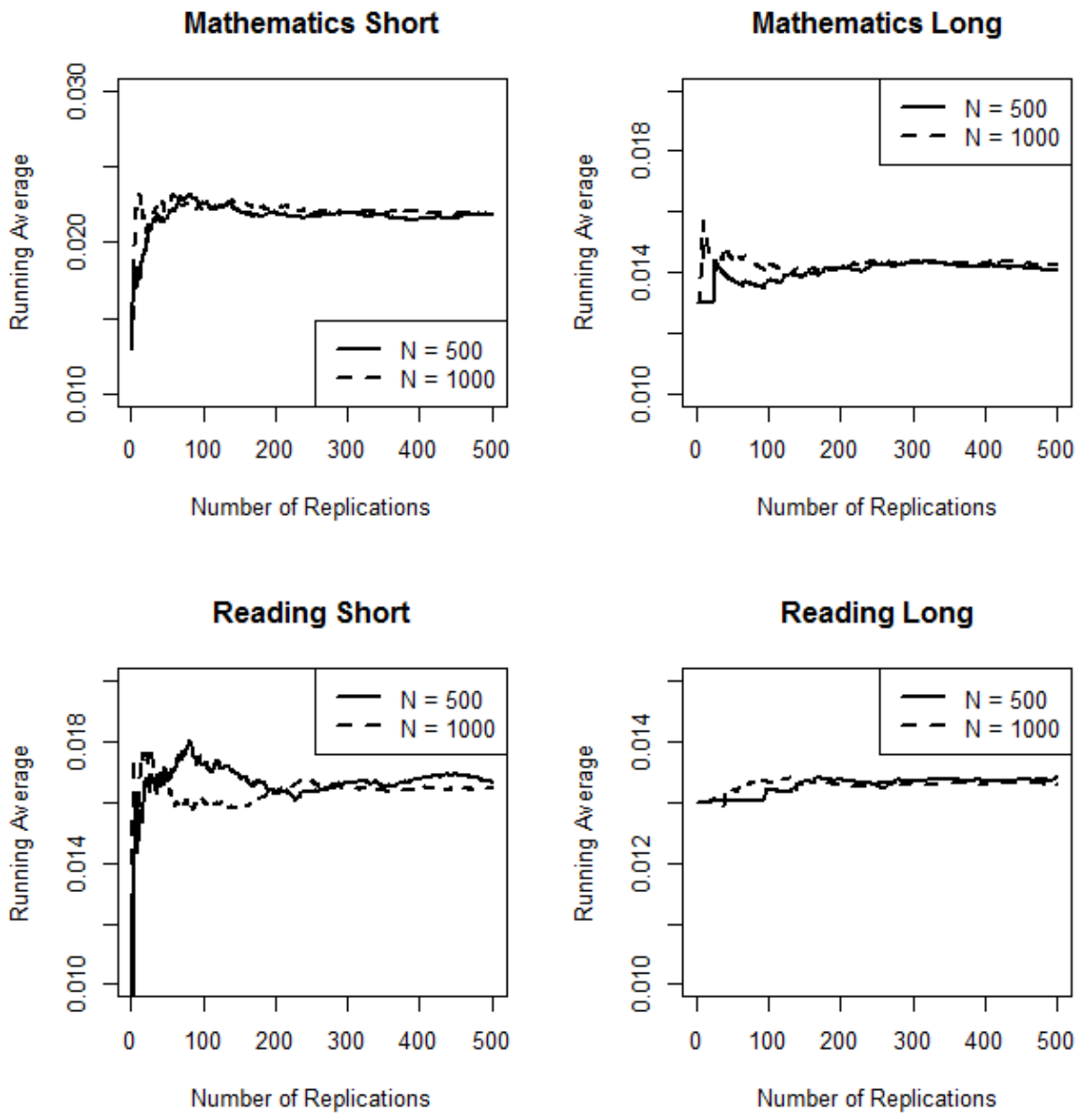


Figure 5A. Running Average RMSEA (WLSM) for Two-dimensional Model across Four Subtests

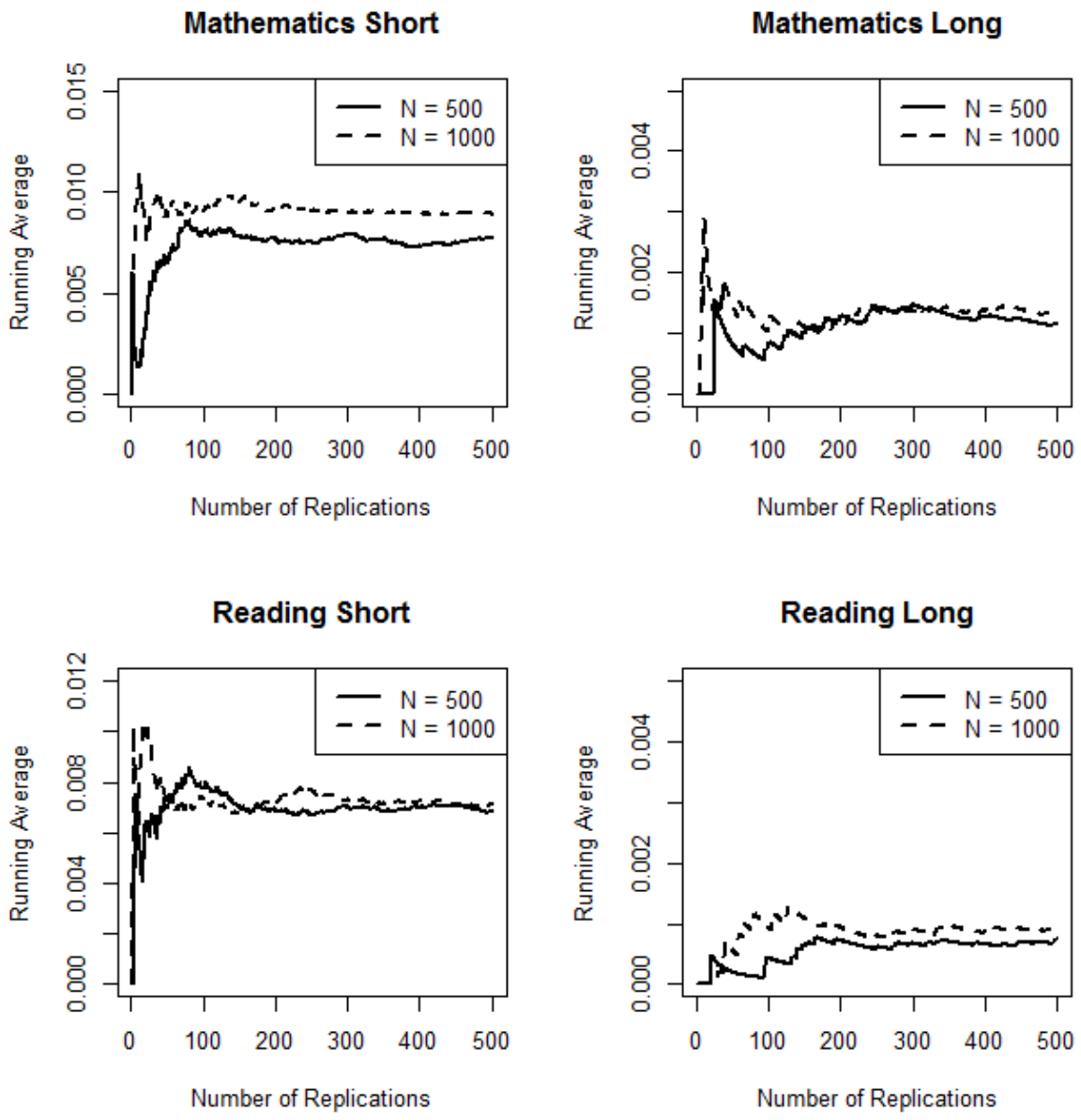


Figure 6A. Running Average RMSEA (WLSM) for Three-dimensional Model across Four Subtests

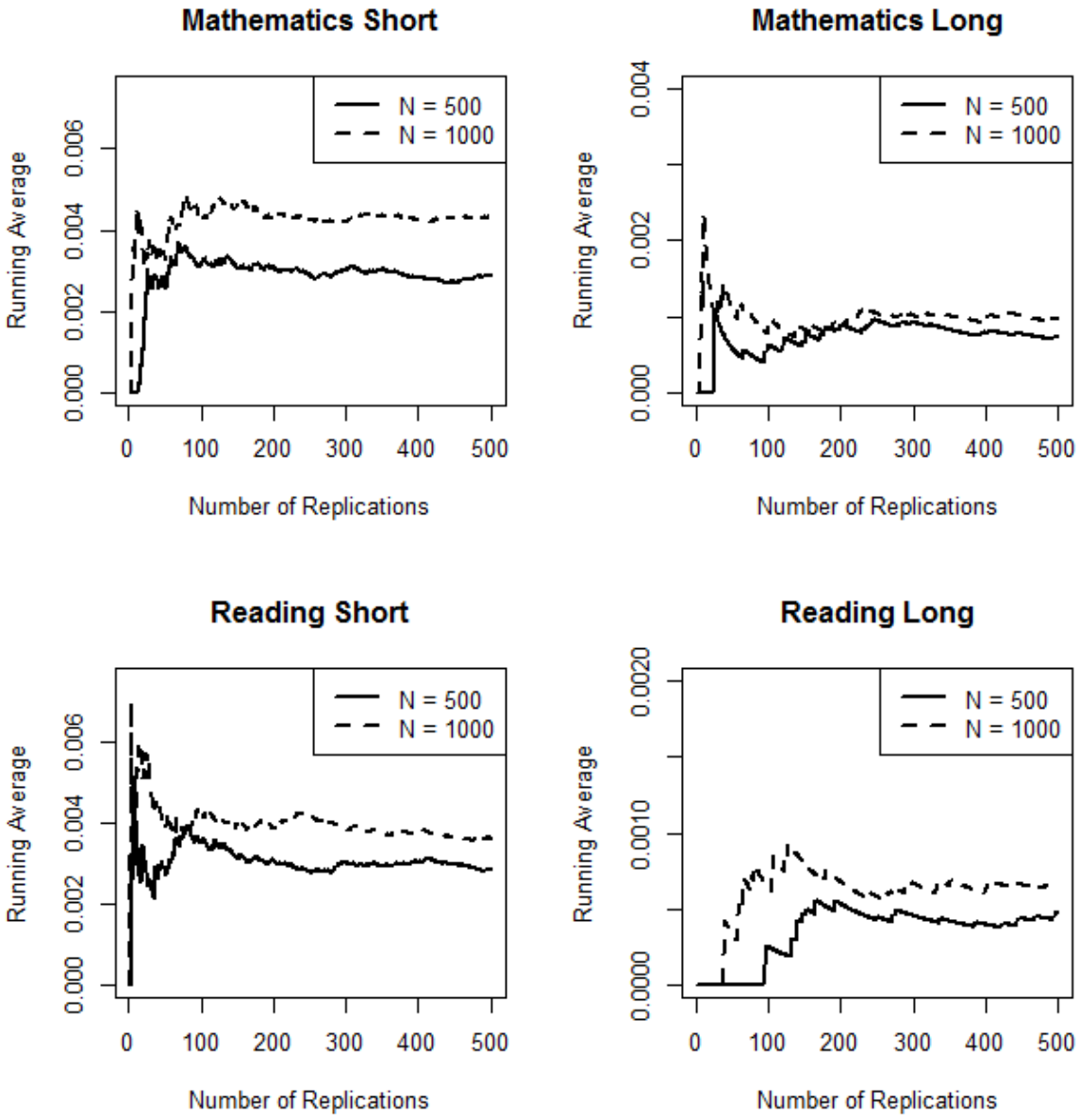


Figure 7A. Running Average CFI (WLSM) for One-dimensional Model across Four Subtests

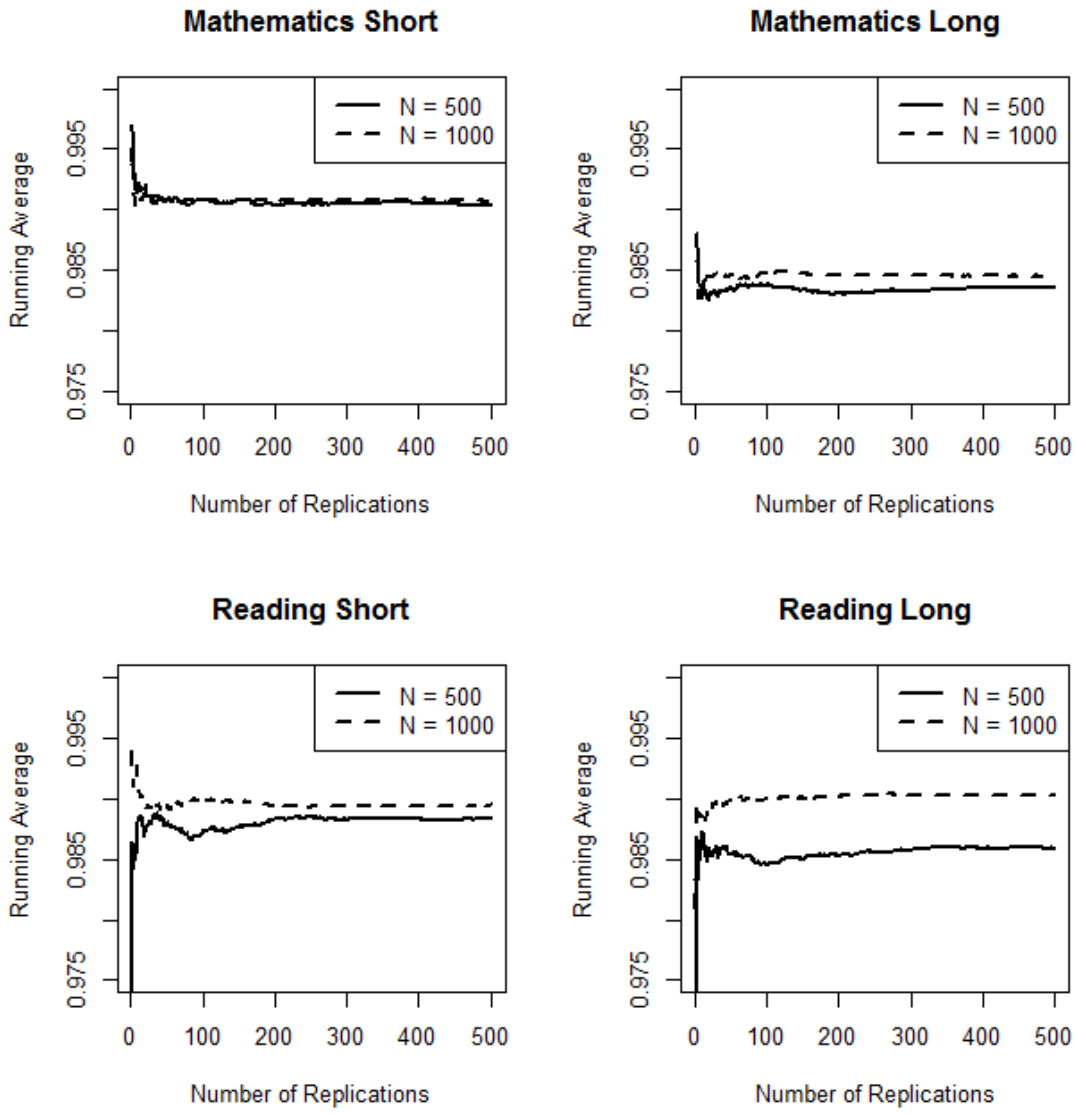


Figure 8A. Running Average CFI (WLSM) for Two-dimensional Model across Four Subtests

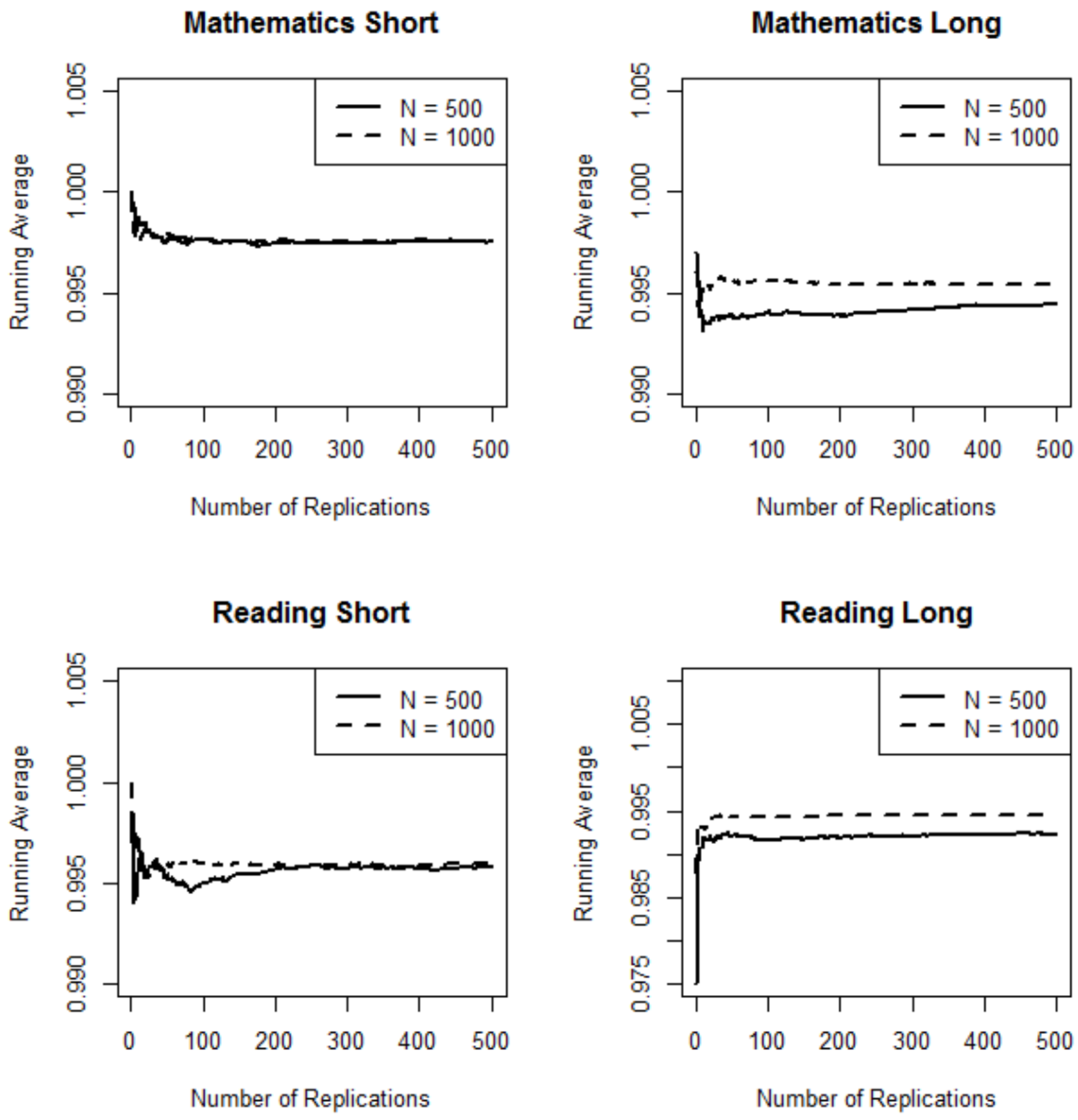


Figure 9A. Running Average CFI (WLSM) for Three-dimensional Model across Four Subtests

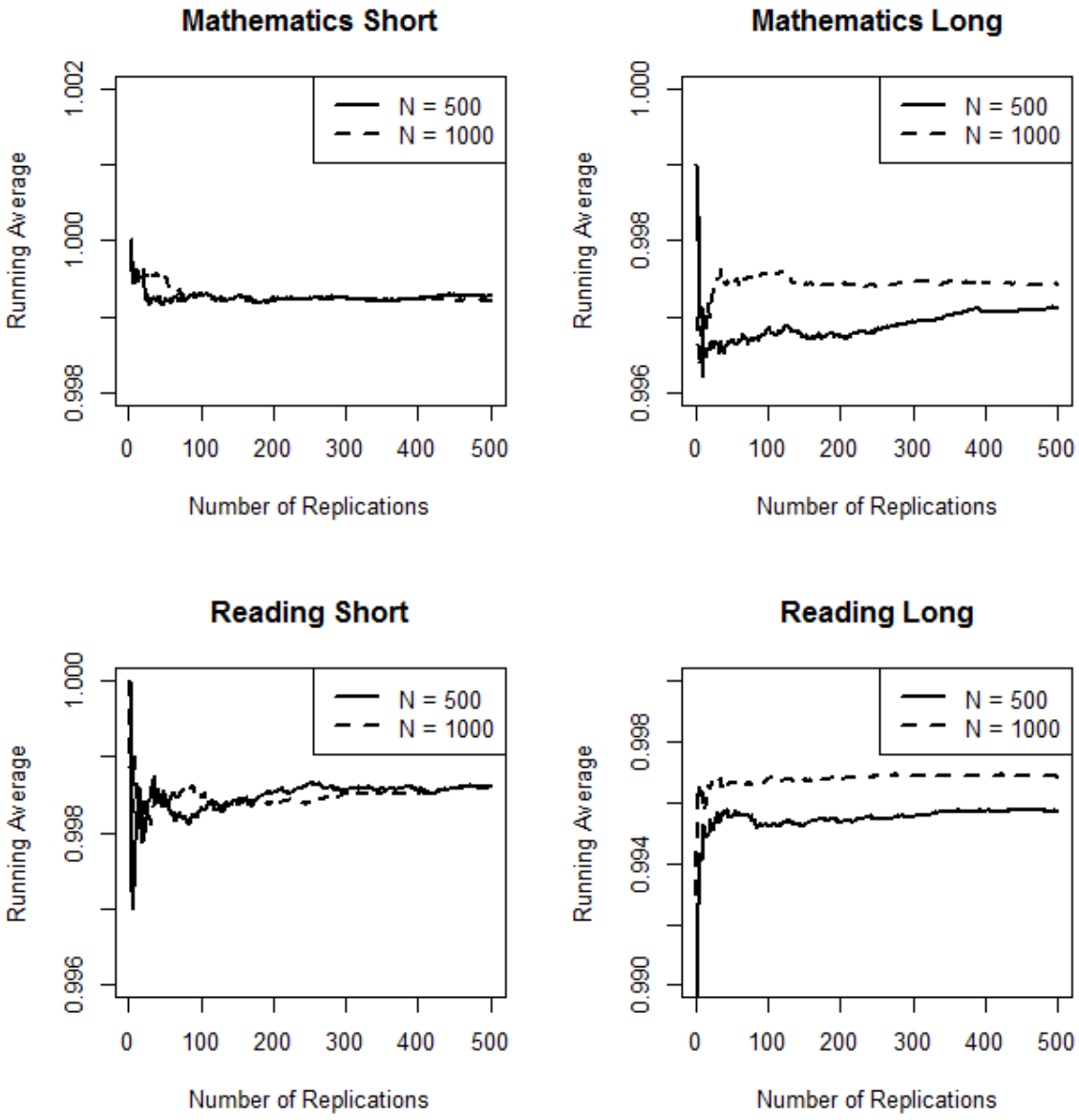


Figure 10A. Running Average SRMR (WLSM) for One-dimensional Model across Four Subtests

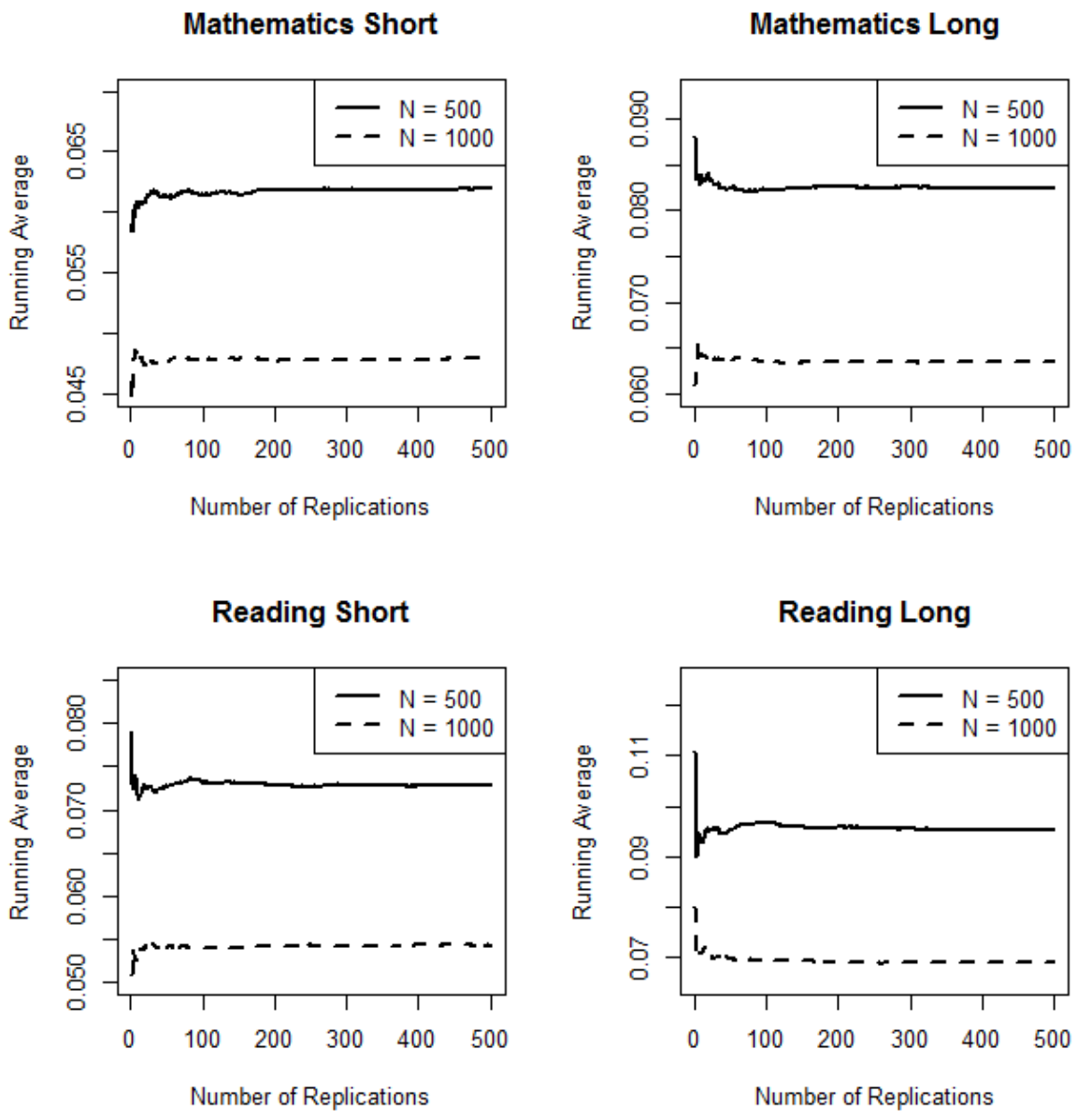


Figure 11A. Running Average SRMR (WLSM) for Two-dimensional Model across Four Subtests

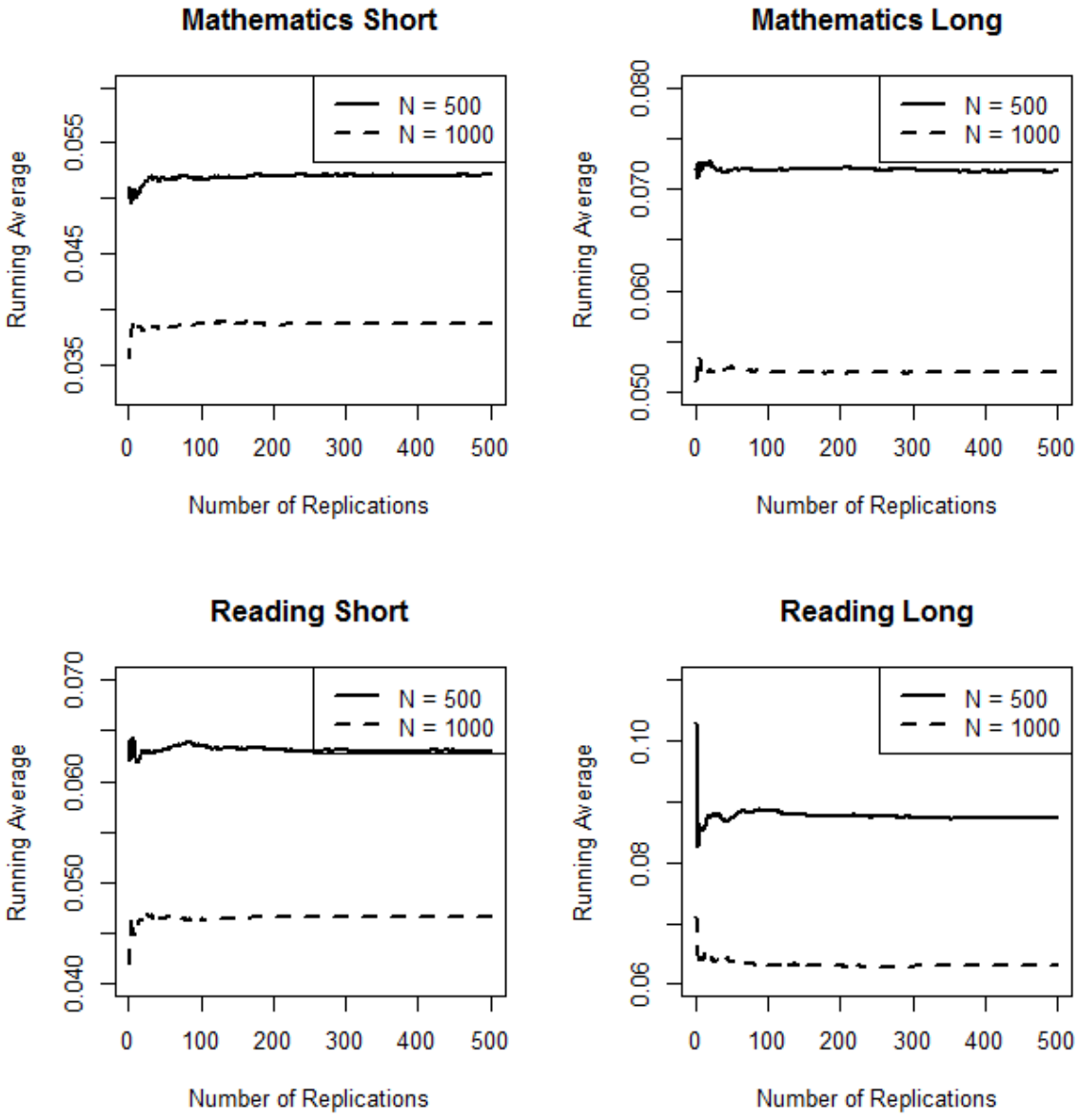


Figure 12A. Running Average SRMR (WLSM) for Three-dimensional Model across Four Subtests

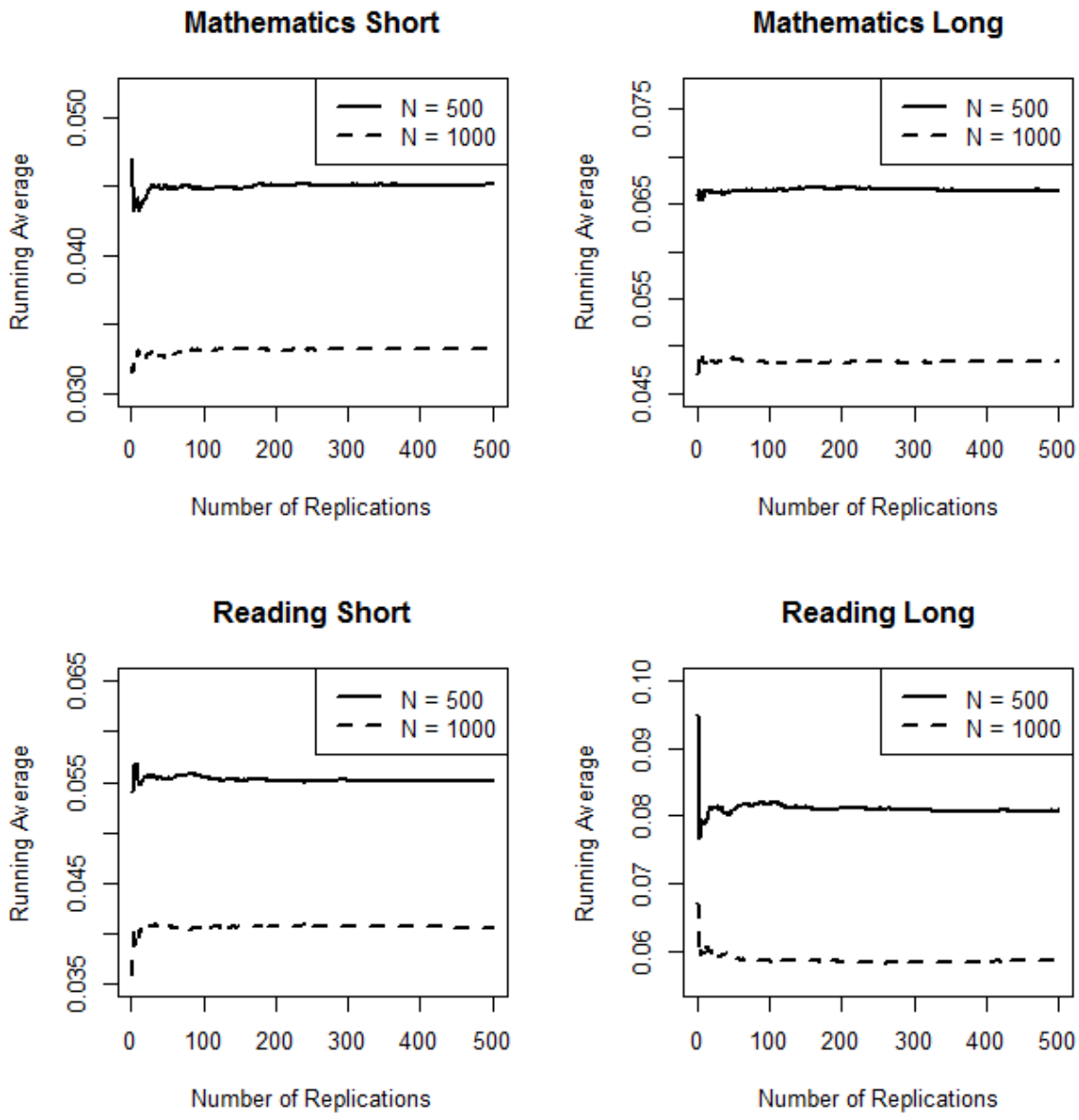


Figure 13A. Running Average Mean-Adjusted MPLUS Chi-Square (WLSM) for One-dimensional Model across Four Subtests

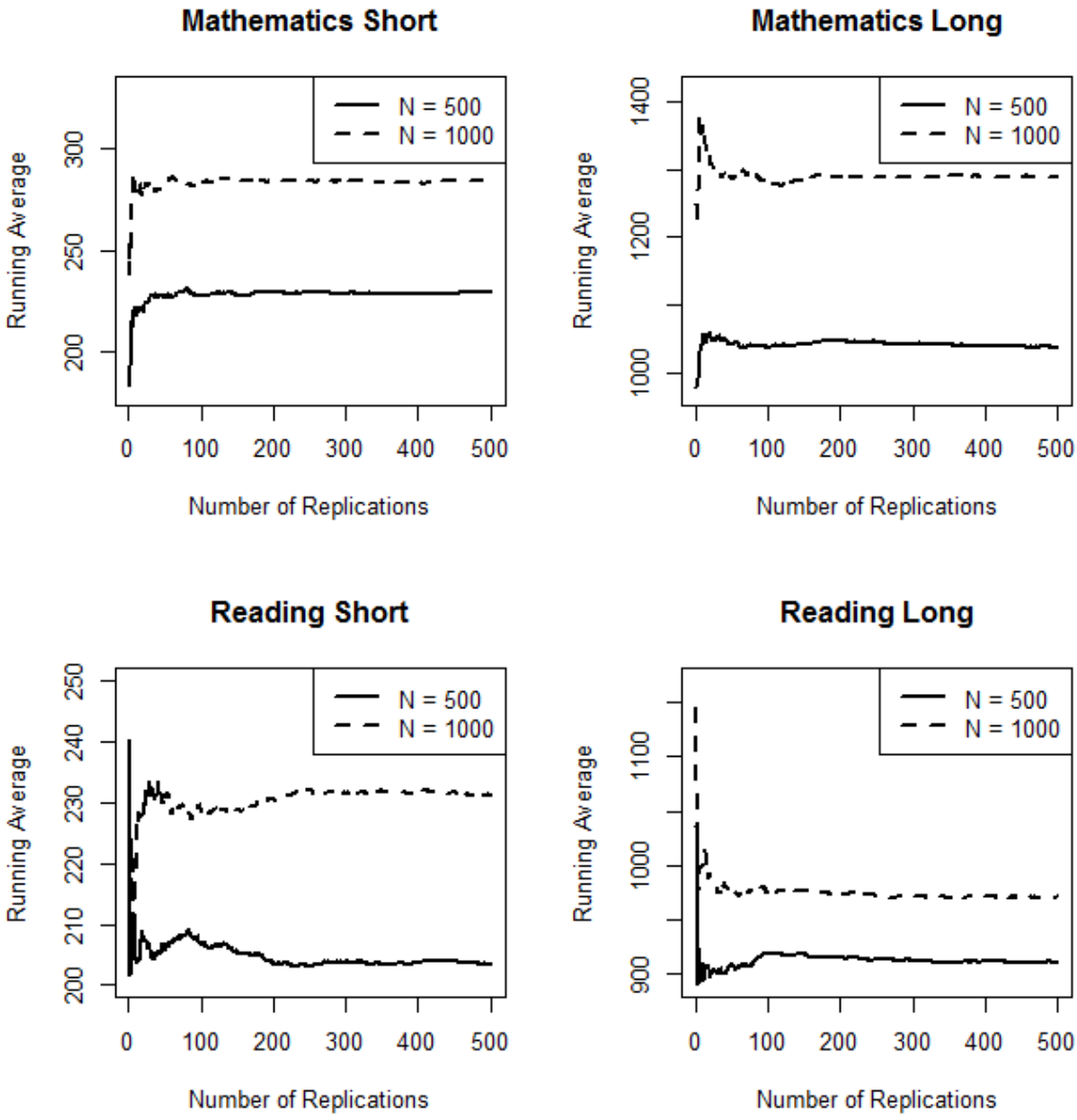


Figure 14A. Running Average Mean-Adjusted MPLUS Chi-Square (WLSM) for Two-dimensional Model across Four Subtests

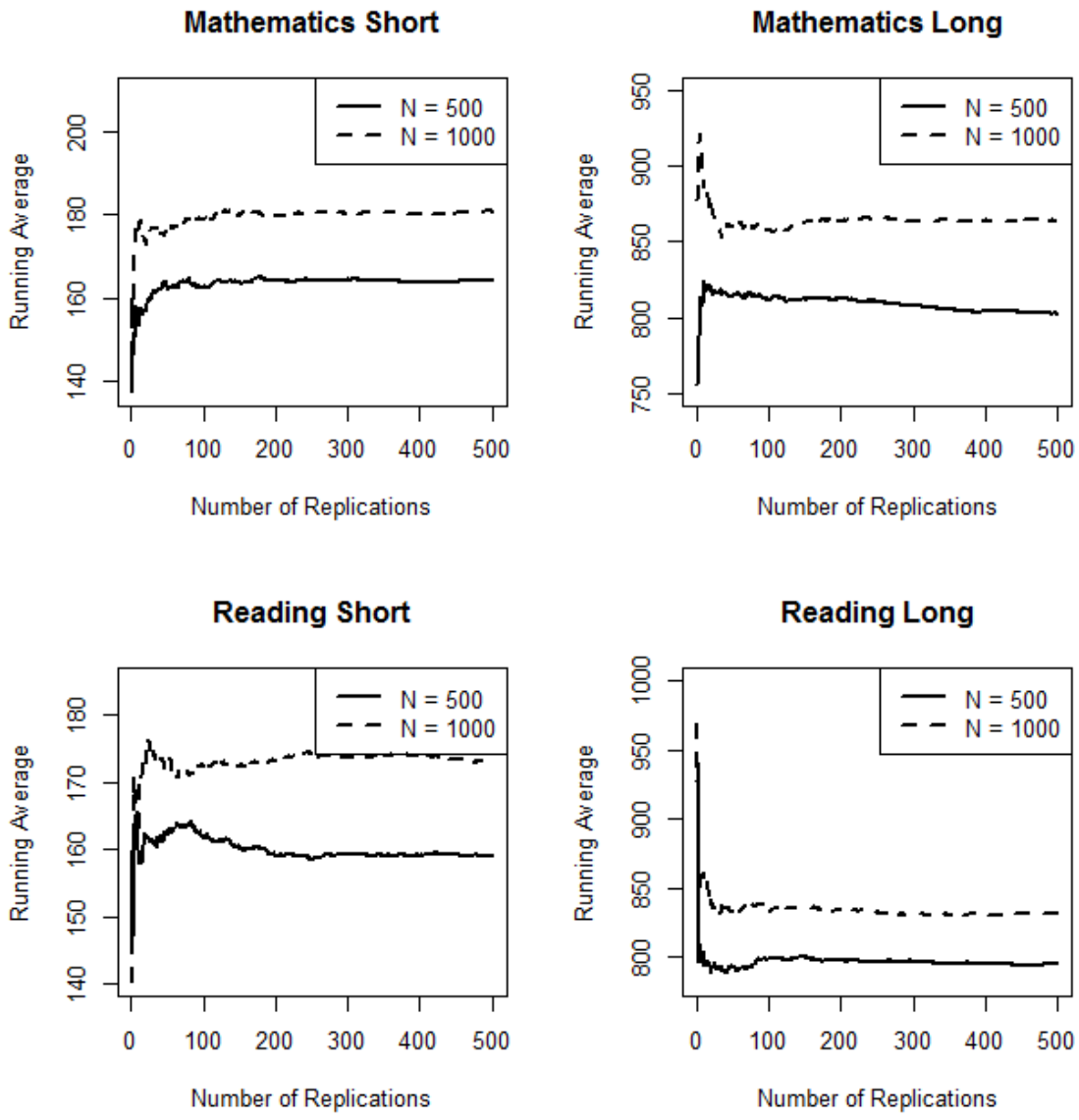


Figure 15A. Running Average Mean-Adjusted MPLUS Chi-Square (WLSM) for Three-dimensional Model across Four Subtests

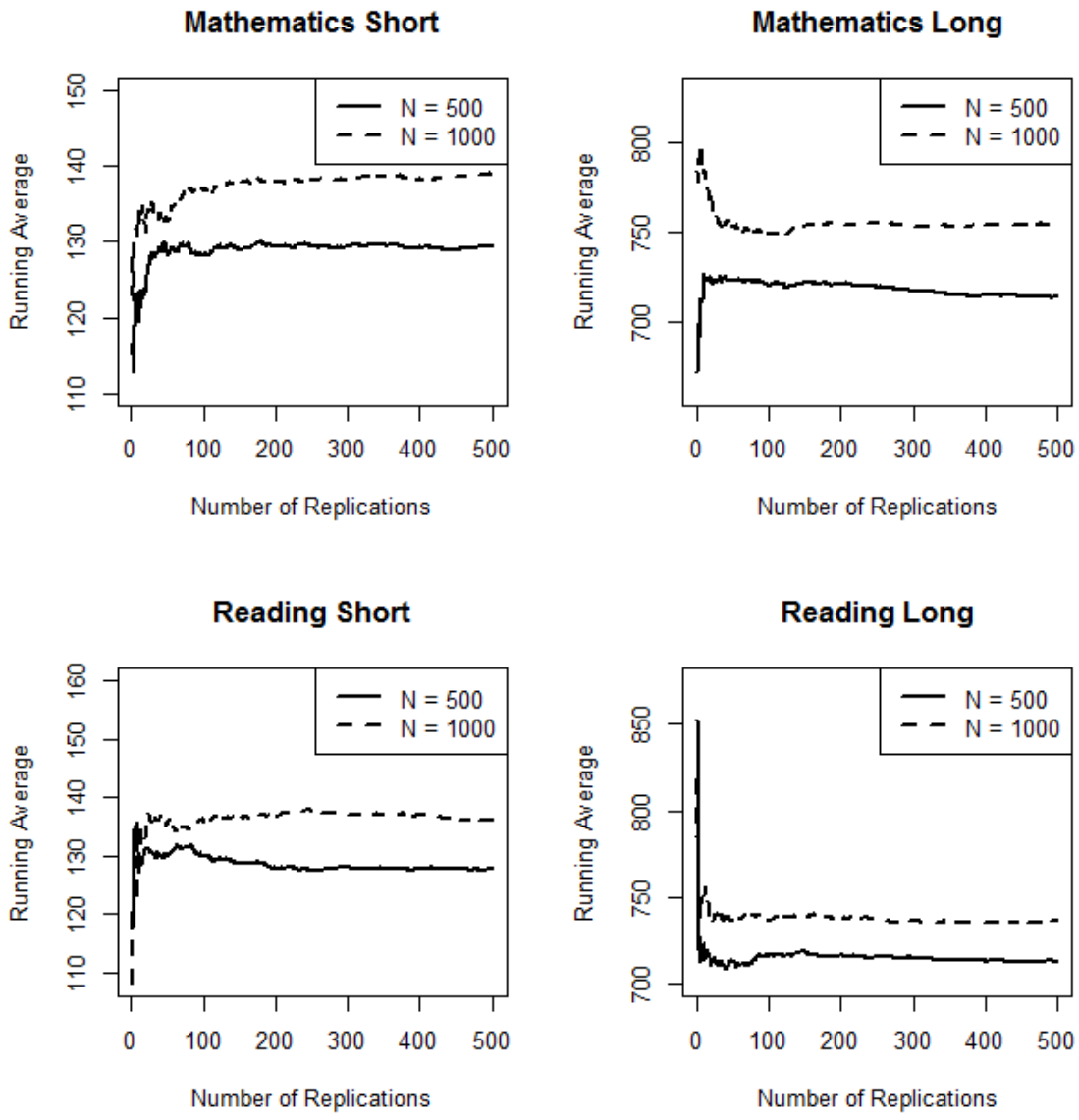


Figure 16A. Running Average Mean-and-Variance Adjusted MPLUS Chi-Square (WLSMV) for One-dimensional Model across Four Subtests

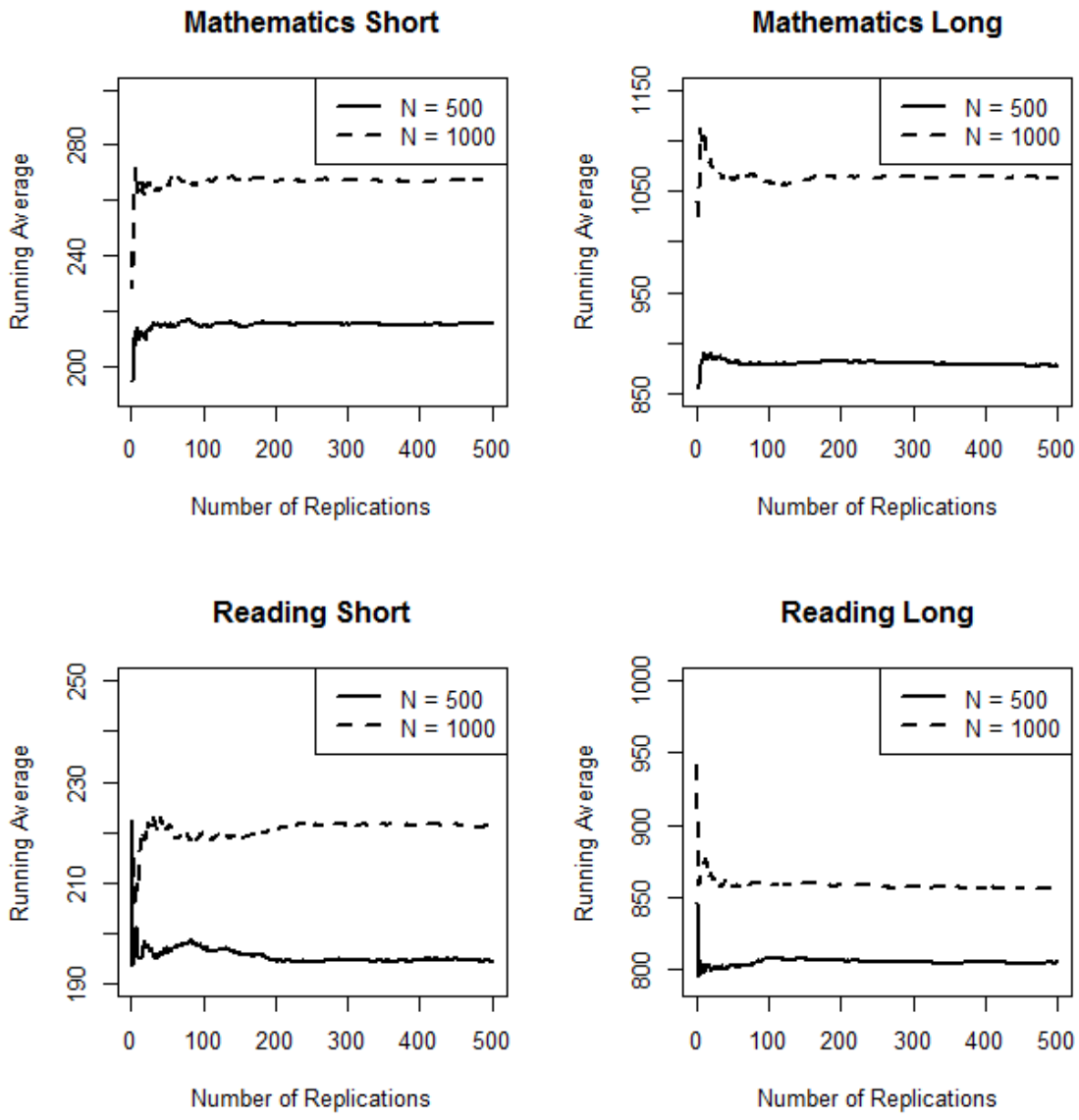


Figure 17A. Running Average Mean-and-Variance Adjusted MPLUS Chi-Square (WLSMV) for Two-dimensional Model across Four Subtests

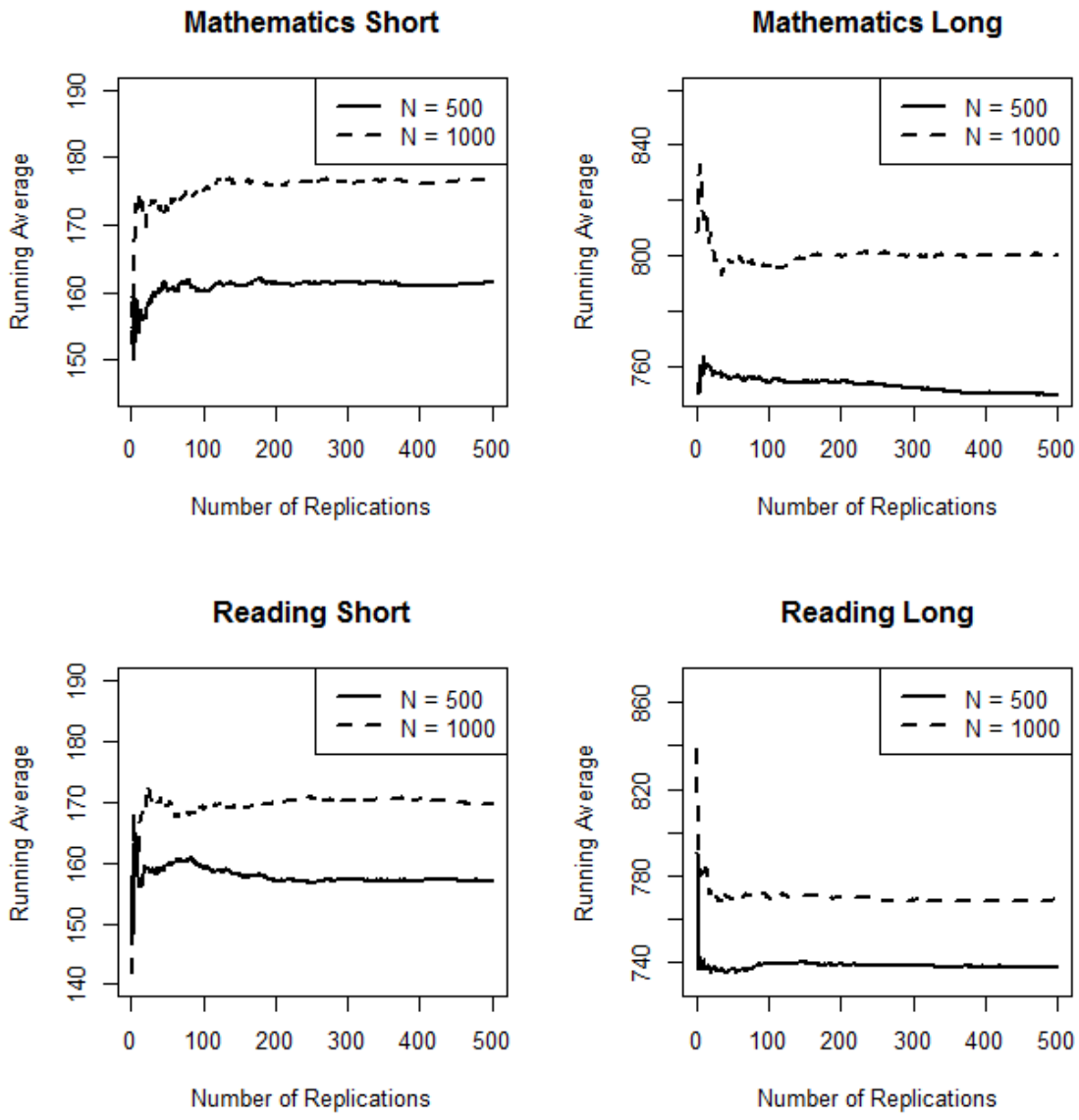


Figure 18A. Running Average Mean-and-Variance Adjusted MPLUS Chi-Square (WLSMV) for Three-dimensional Model across Four Subtests

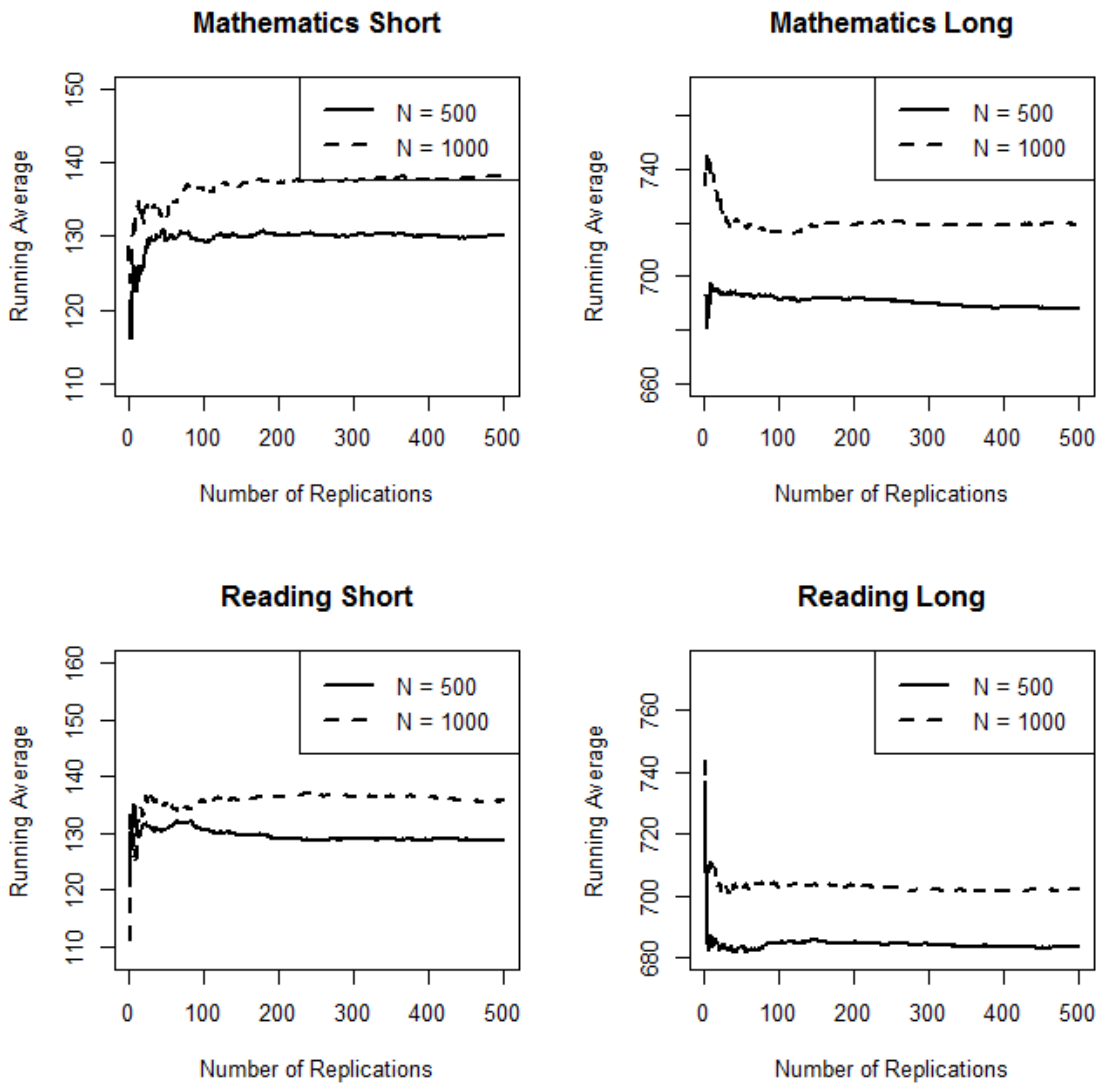


Figure 19A. Running Average Approximate Chi-Square in NOHARM for One-dimensional Model across Four Subtests

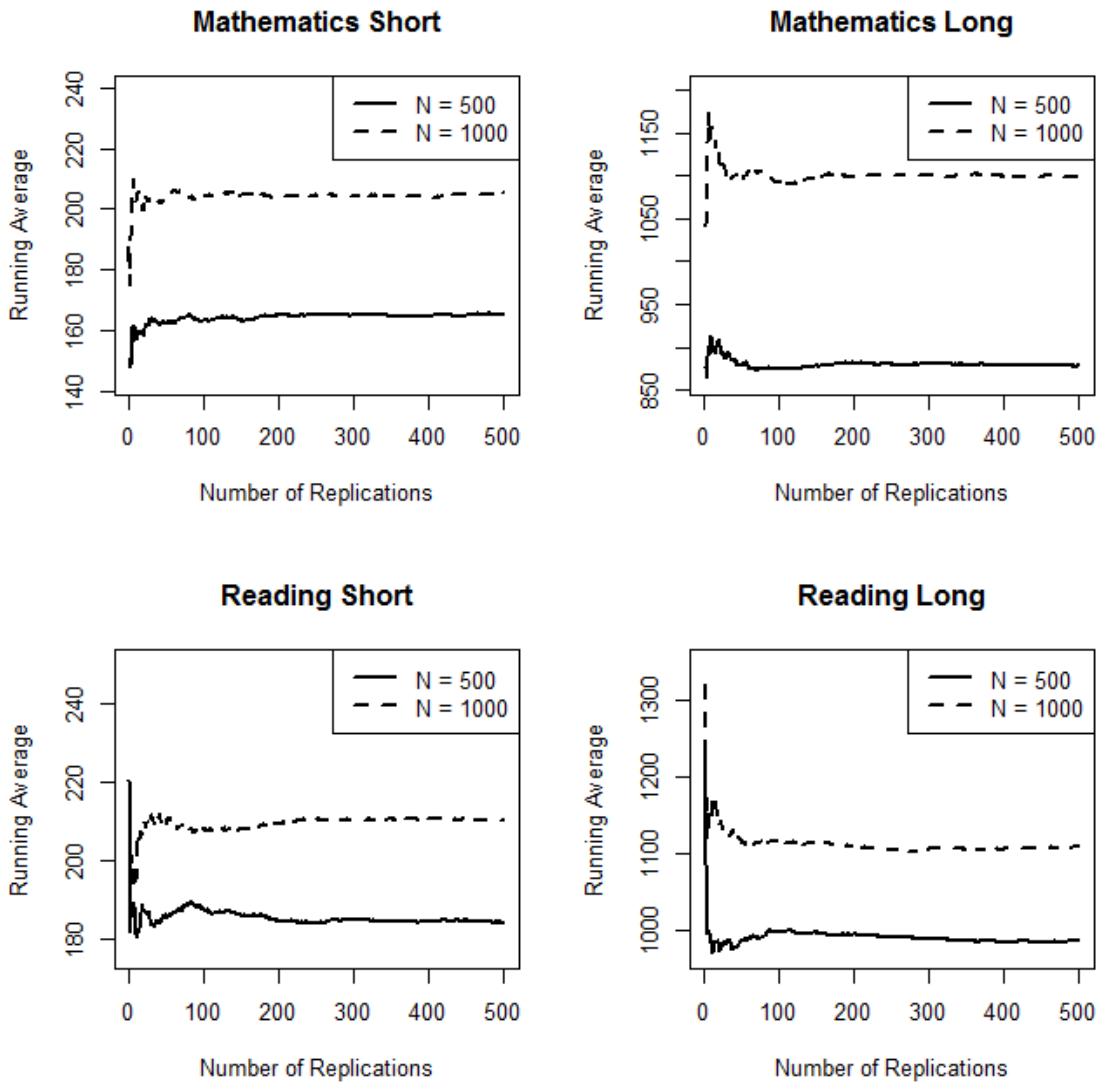


Figure 20A. Running Average Approximate Chi-Square in NOHARM for Two-dimensional Model across Four Subtests

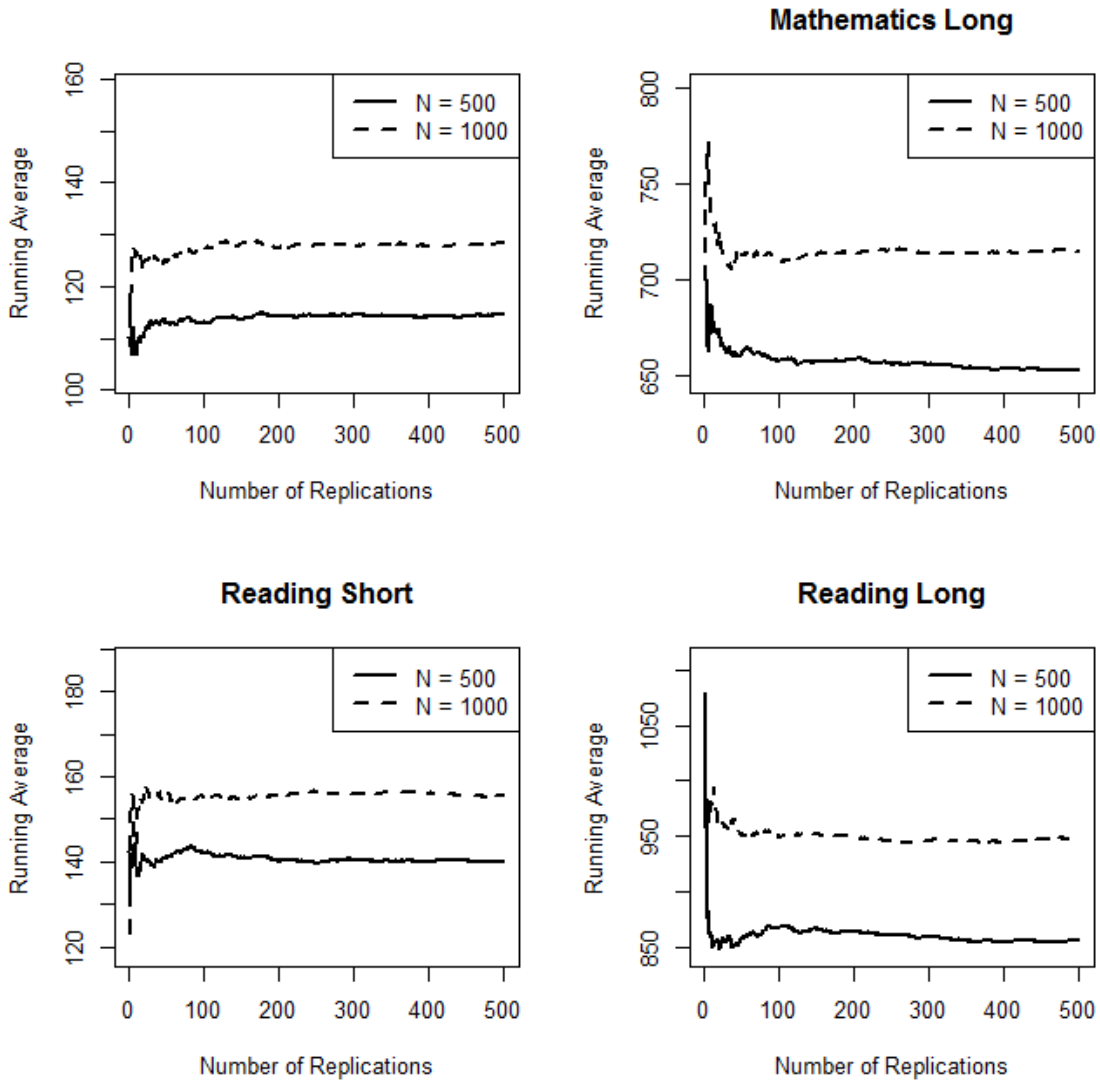


Figure 21A. Running Average Approximate Chi-Square in NOHARM for Three-dimensional Model across Four Subtests

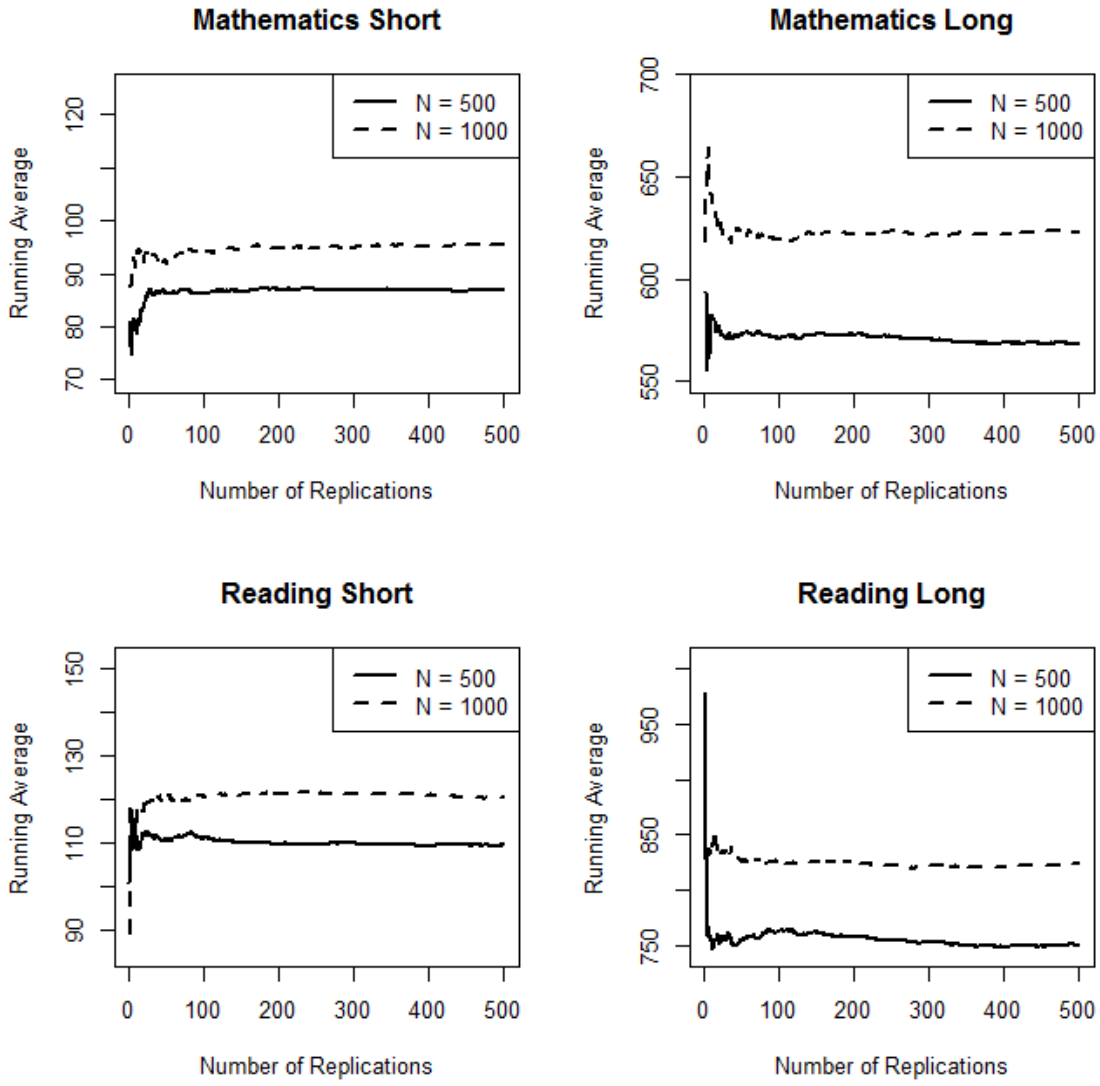


Figure 22A. Running Average Approximate Likelihood Ratio Chi-Square in NOHARM for One-dimensional Model across Four Subtests

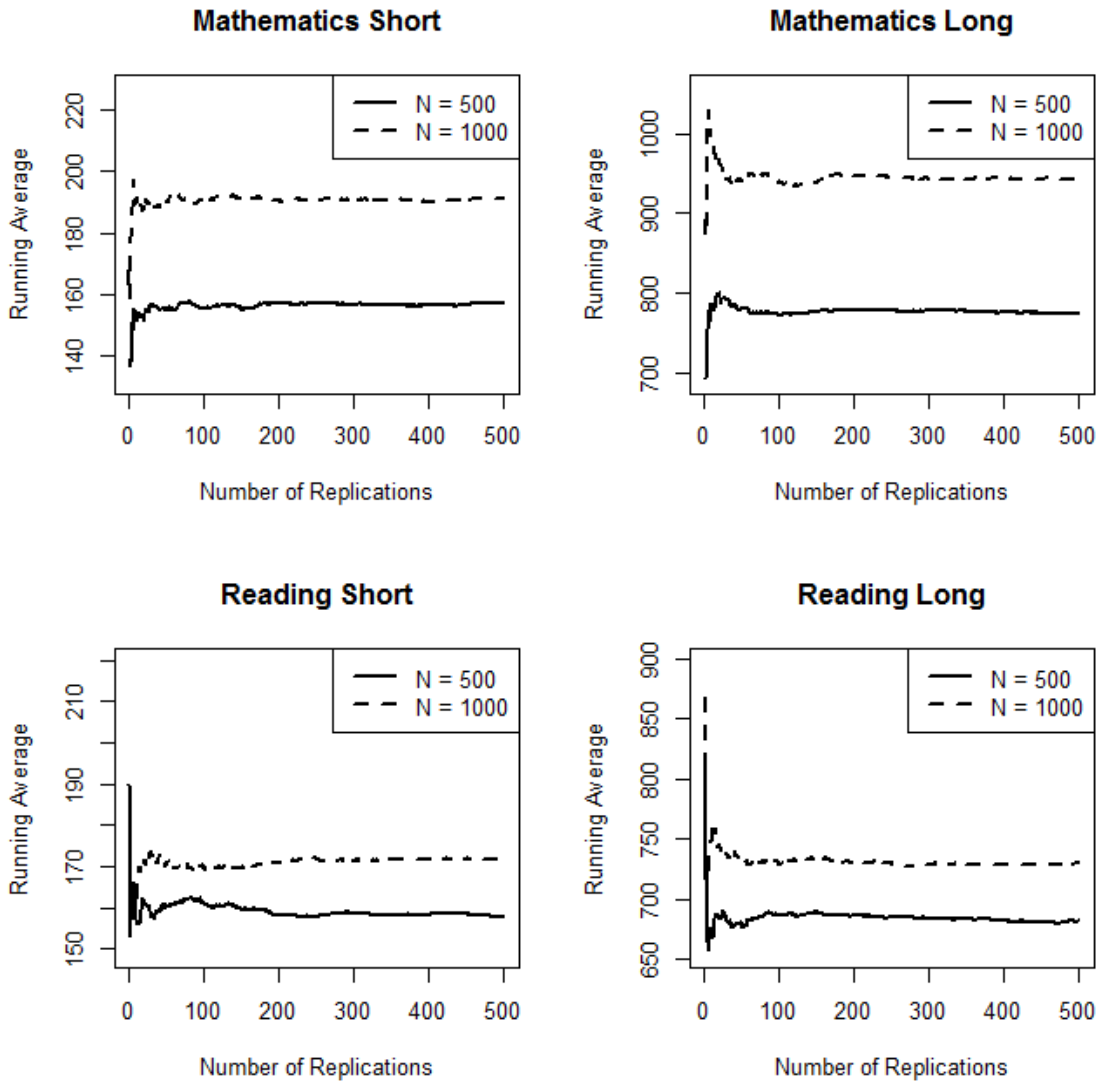


Figure 23A. Running Average Approximate Likelihood Ratio Chi-Square in NOHARM for Two-dimensional Model across Four Subtests

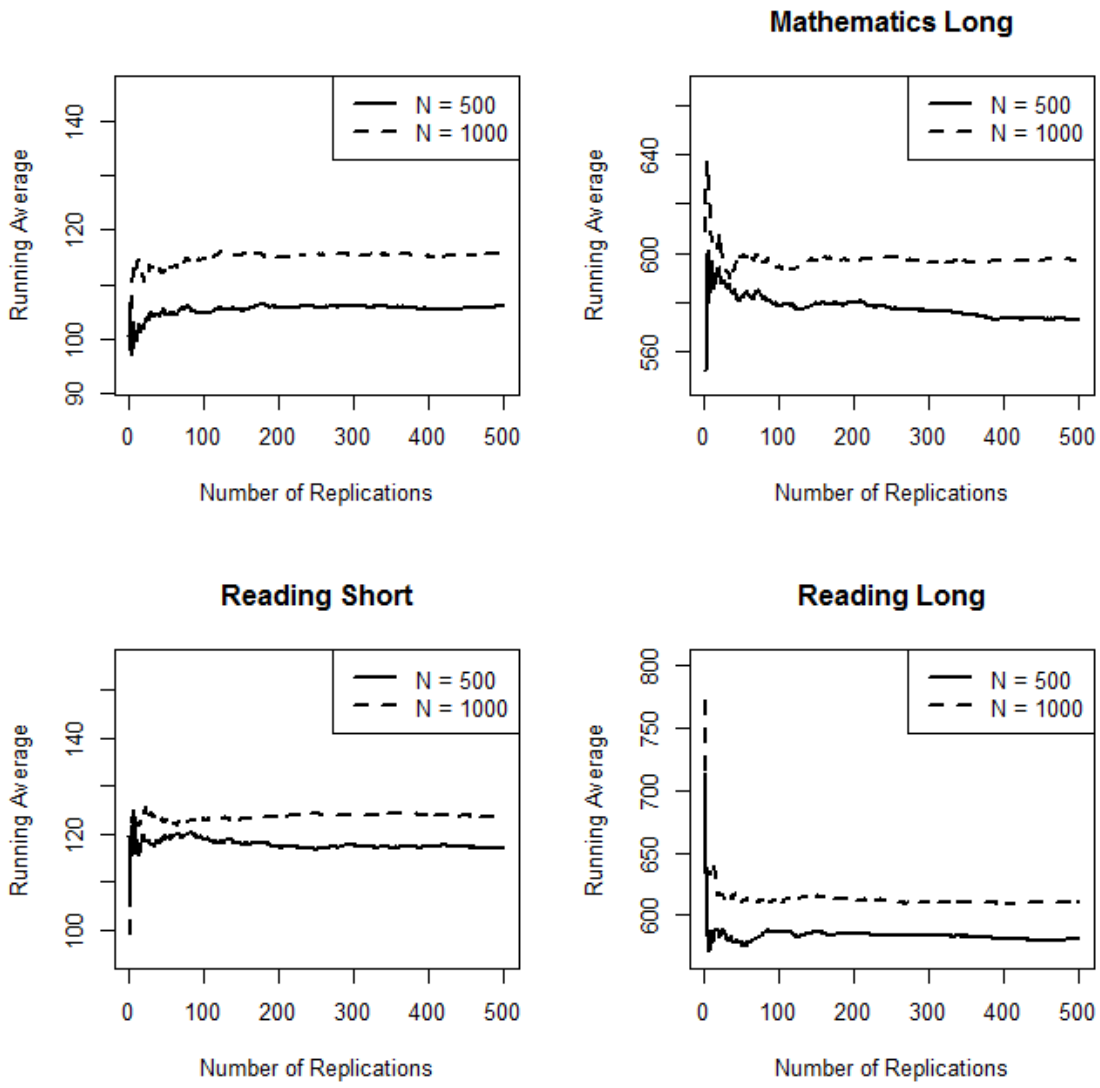


Figure 24A. Running Average Approximate Likelihood Ratio Chi-Square in NOHARM for Three-dimensional Model across Four Subtests

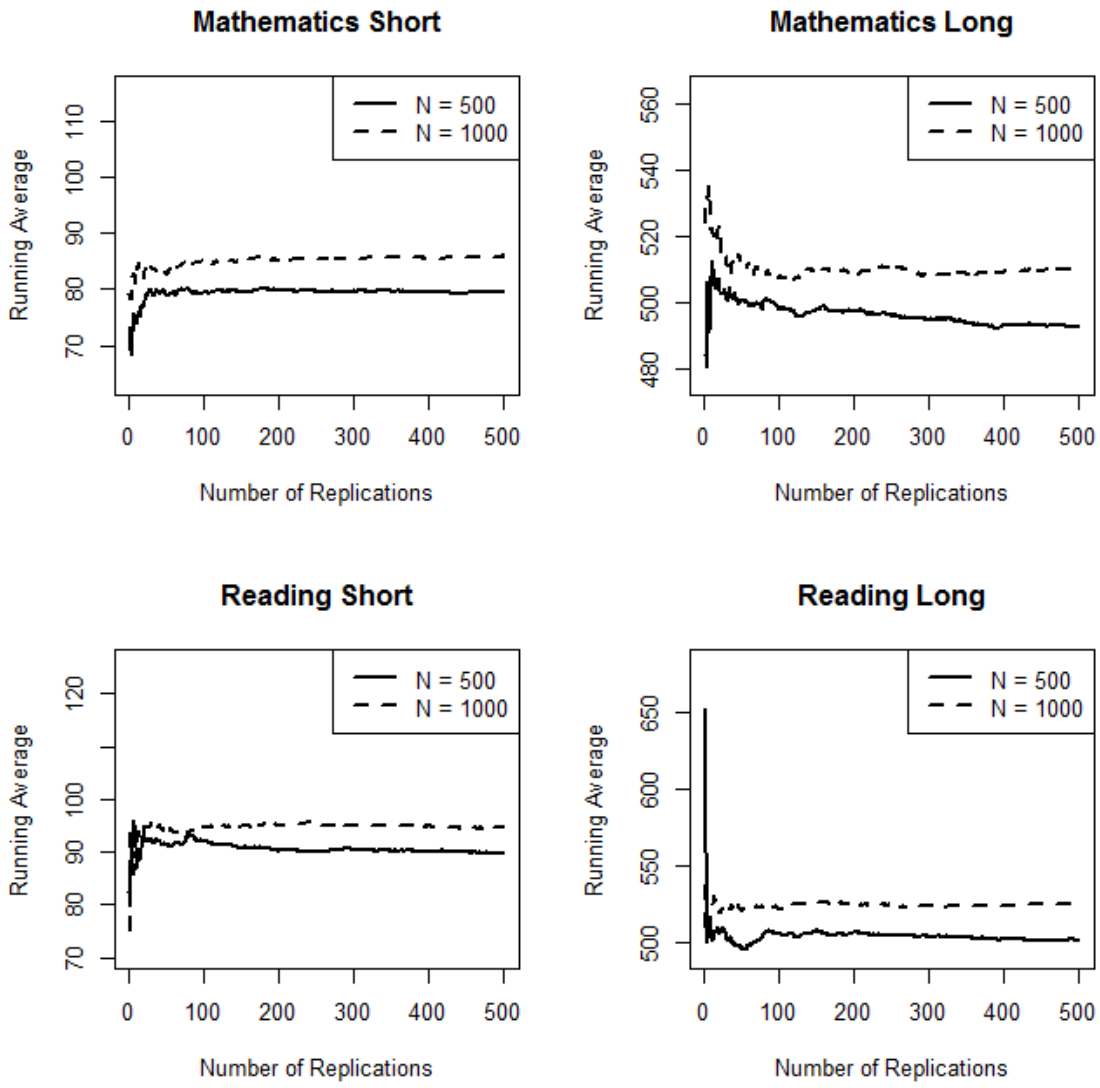


Figure 25A. Running Average Mean-Adjusted NNOHARM Chi-Square for One-dimensional Model across Four Subtests

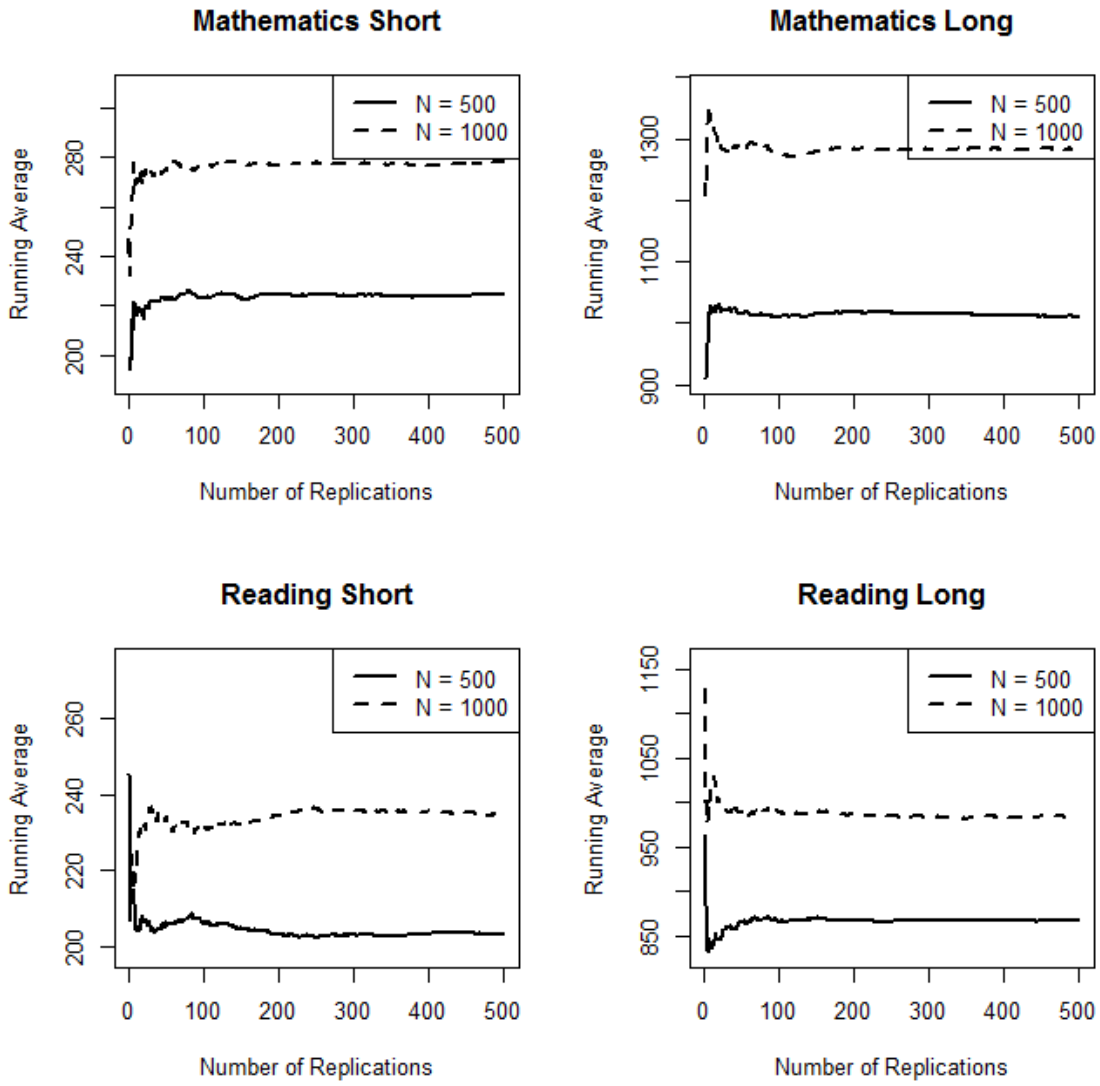


Figure 26A. Running Average Mean-Adjusted NOHARM Chi-Square for Two-dimensional Model across Four Subtests

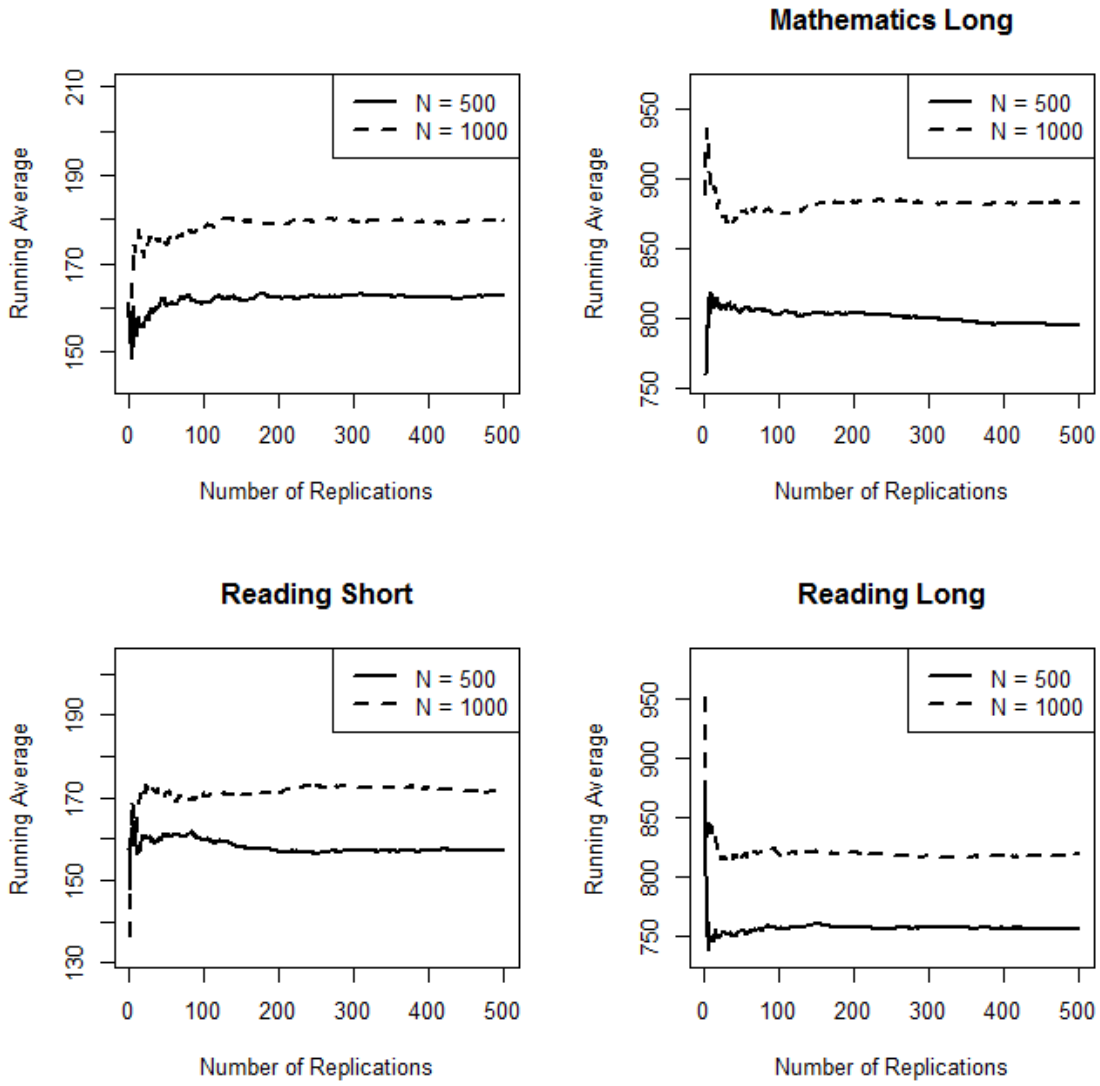


Figure 27A. Running Average Mean-Adjusted NOHARM Chi-Square for Three-dimensional Model across Four Subtests

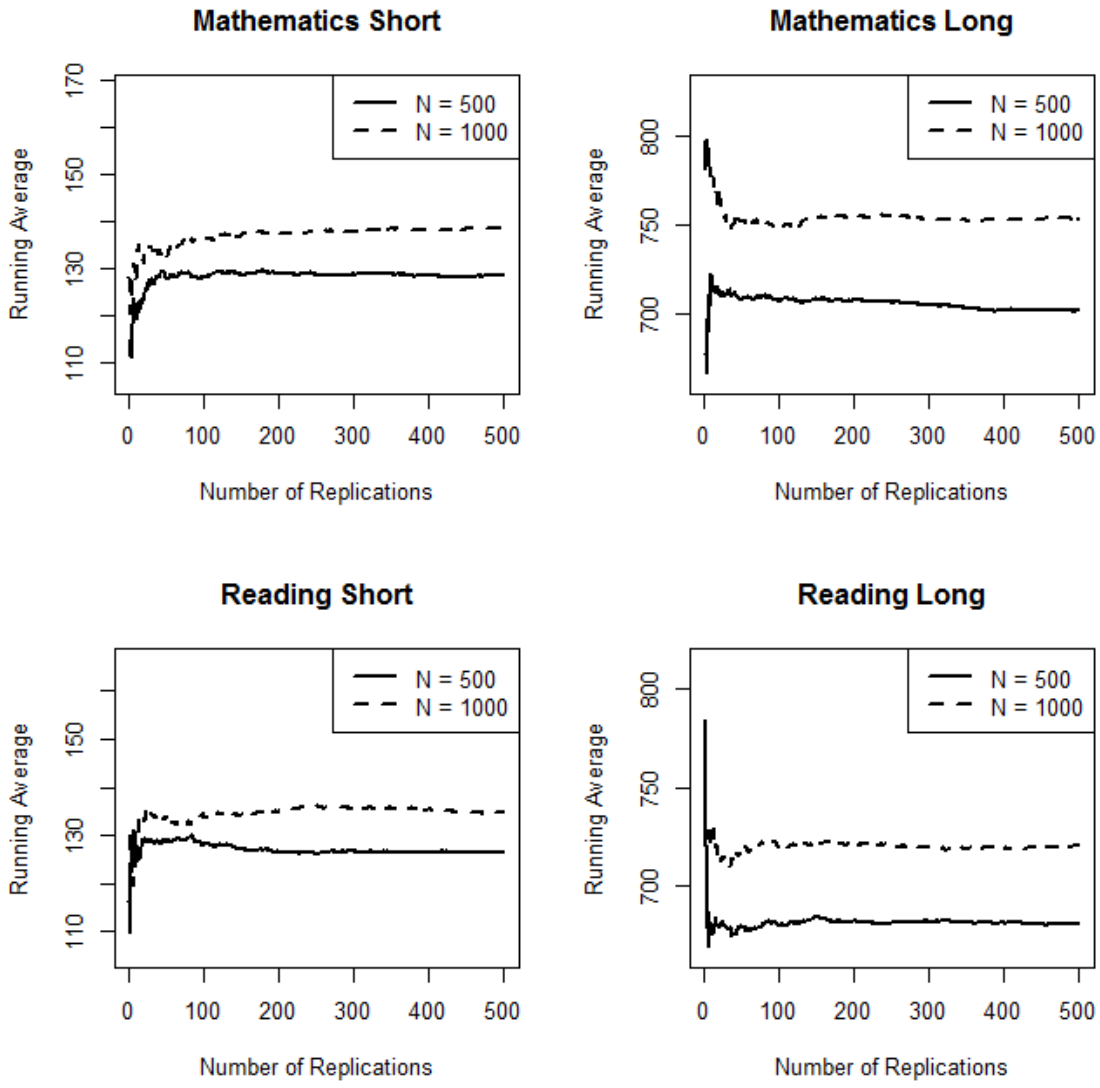


Figure 28A. Running Average Mean-and-Variance Adjusted NOHARM Chi-Square for One-dimensional Model across Four Subtests

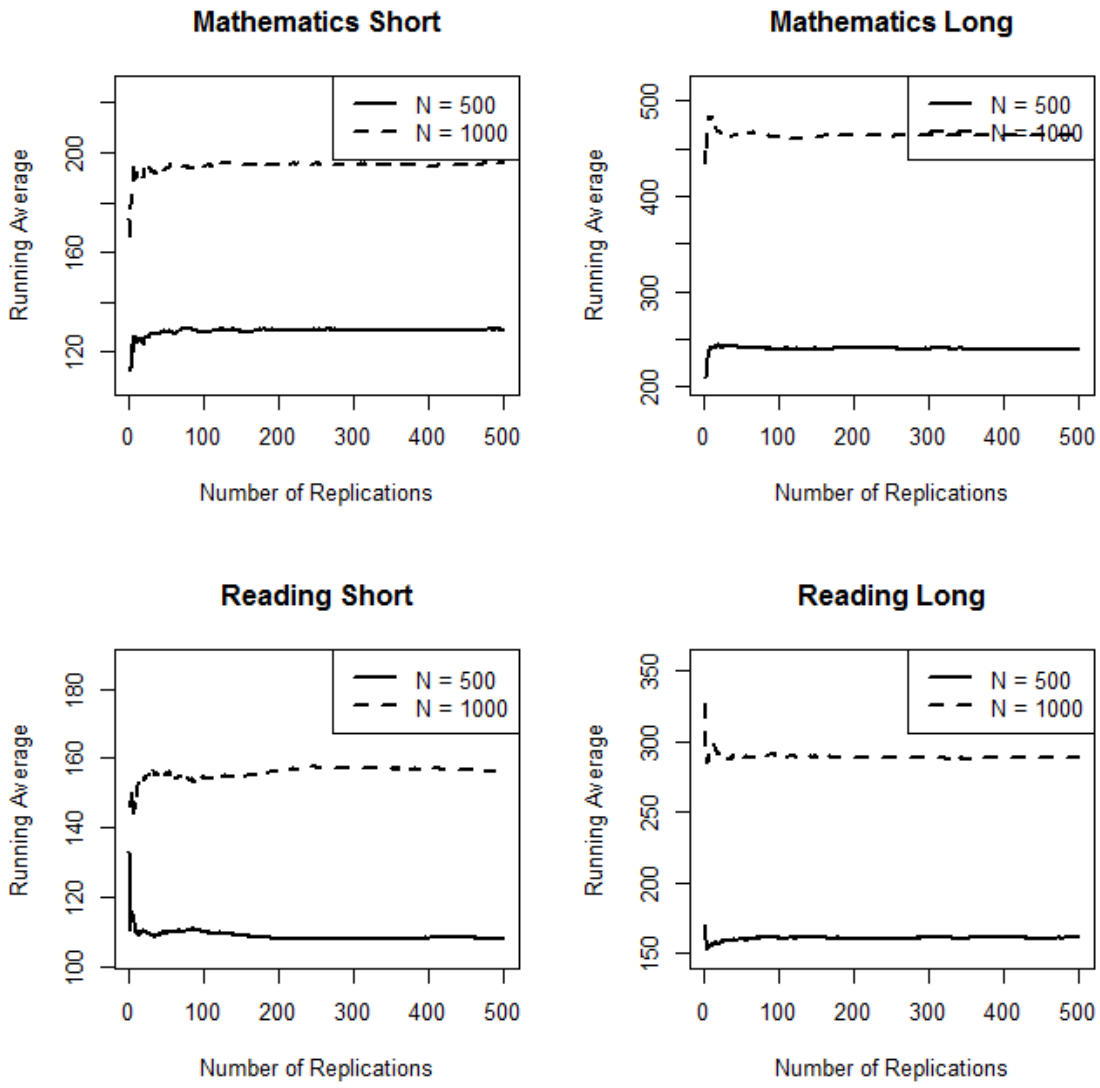


Figure 29A. Running Average Mean-and-Variance Adjusted NOHARM Chi-Square for Two-dimensional Model across Four Subtests

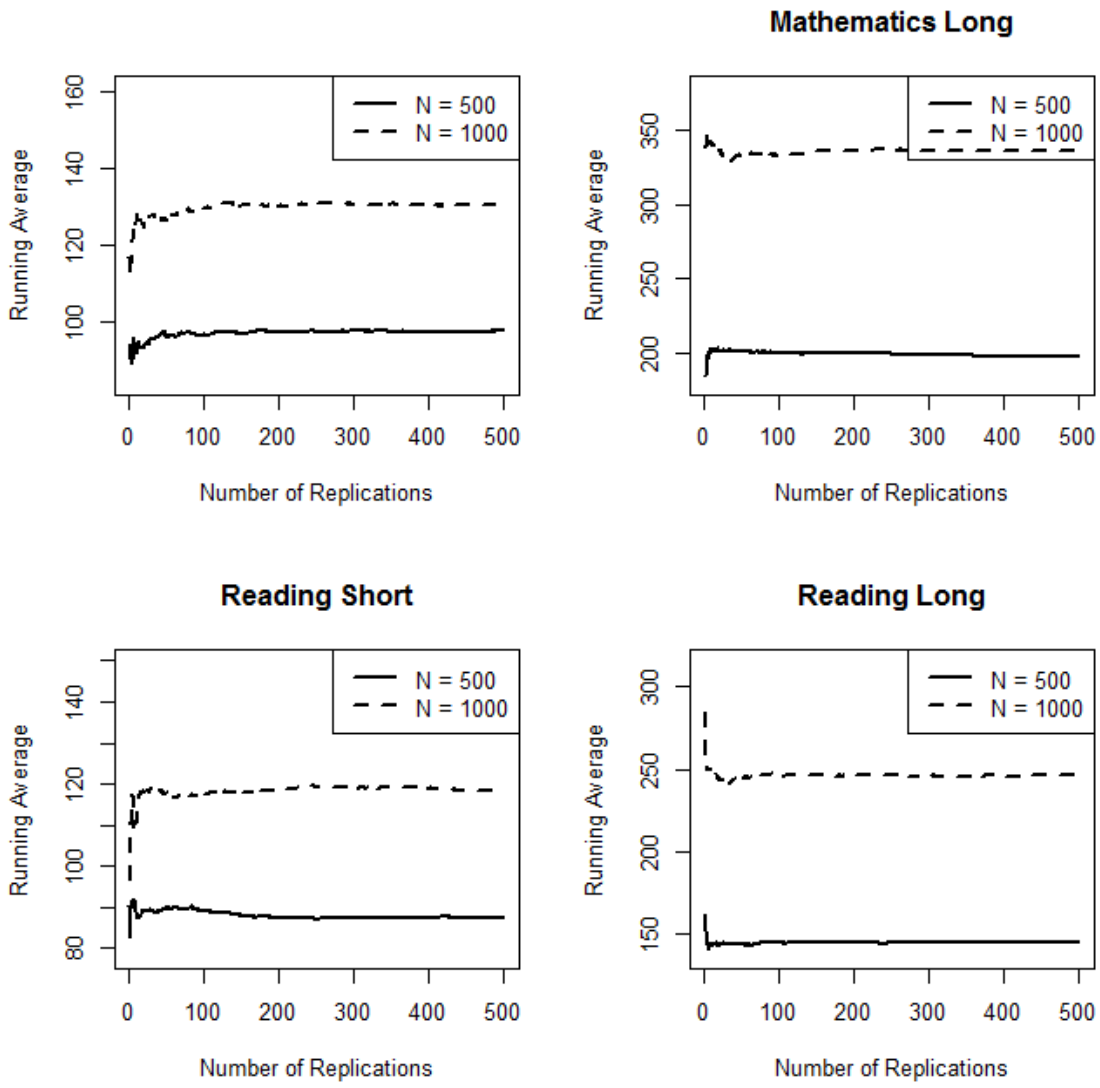


Figure 30A. Running Average Mean-and-Variance Adjusted NOHARM Chi-Square for Three-dimensional Model across Four Subtests

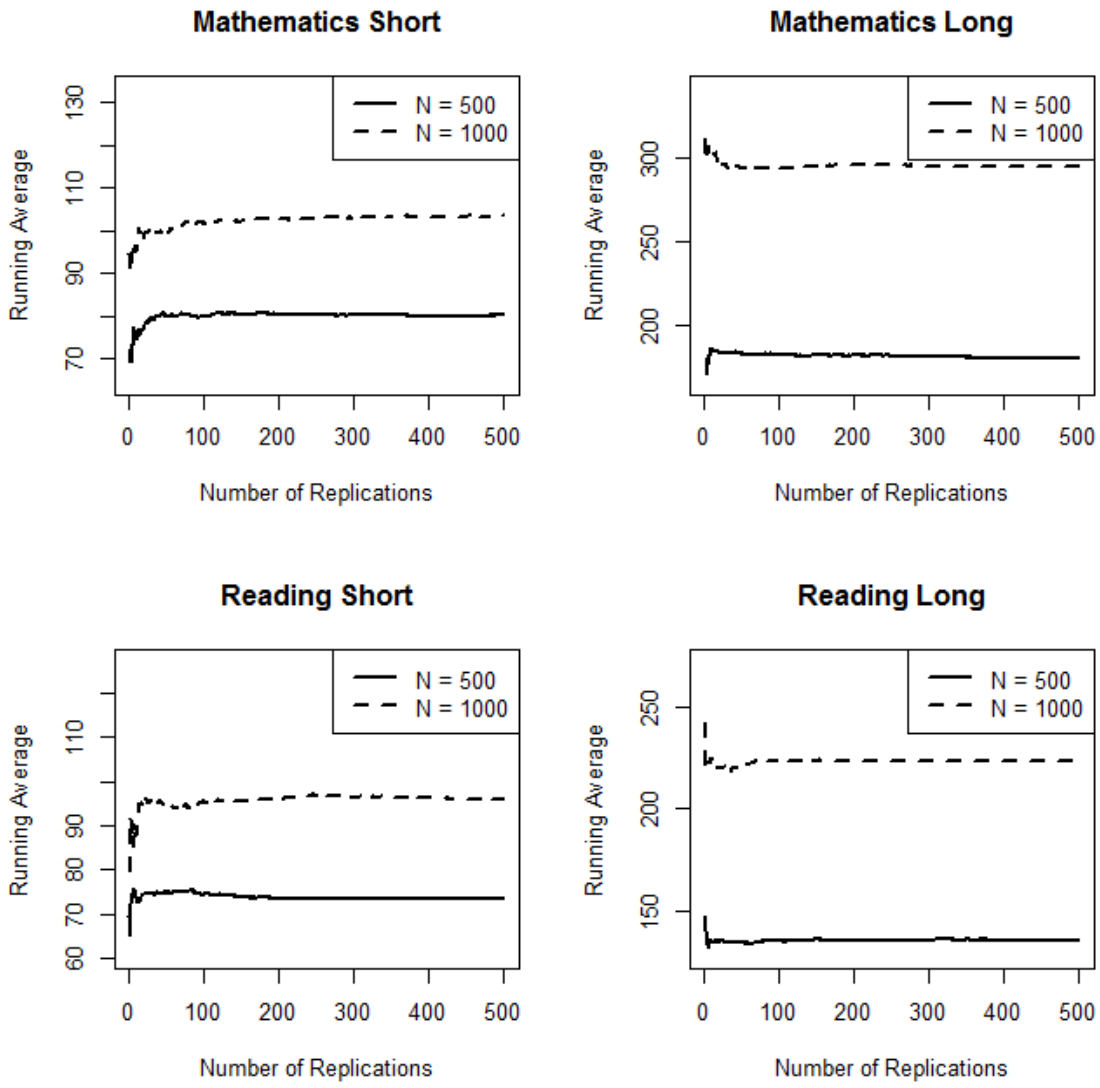


Figure 31A. Running Average Marginal Maximum Likelihood Value (Mplus MLR) for One-, Two-, and Three-dimensional Solutions when Sample Size is 500

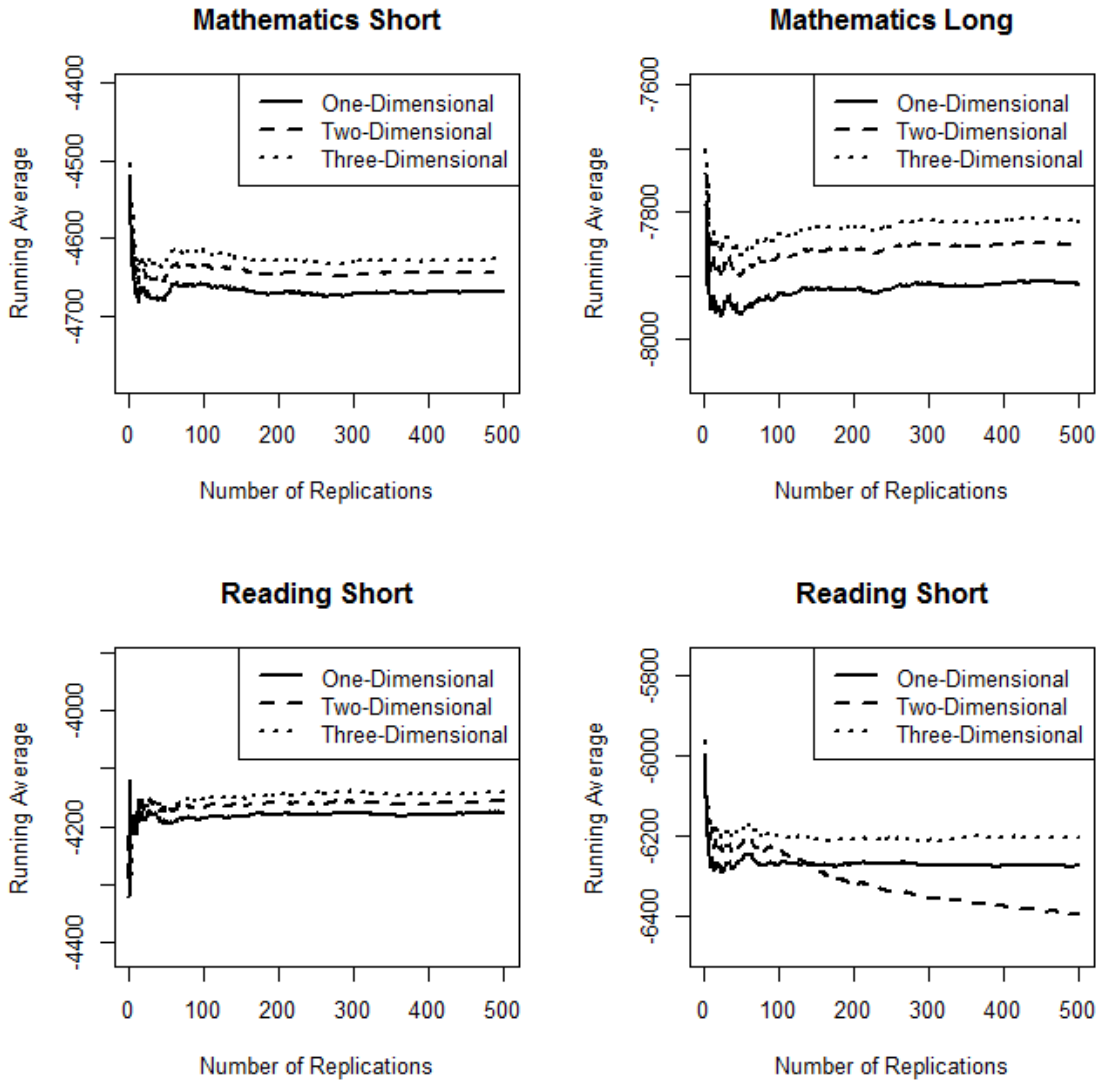
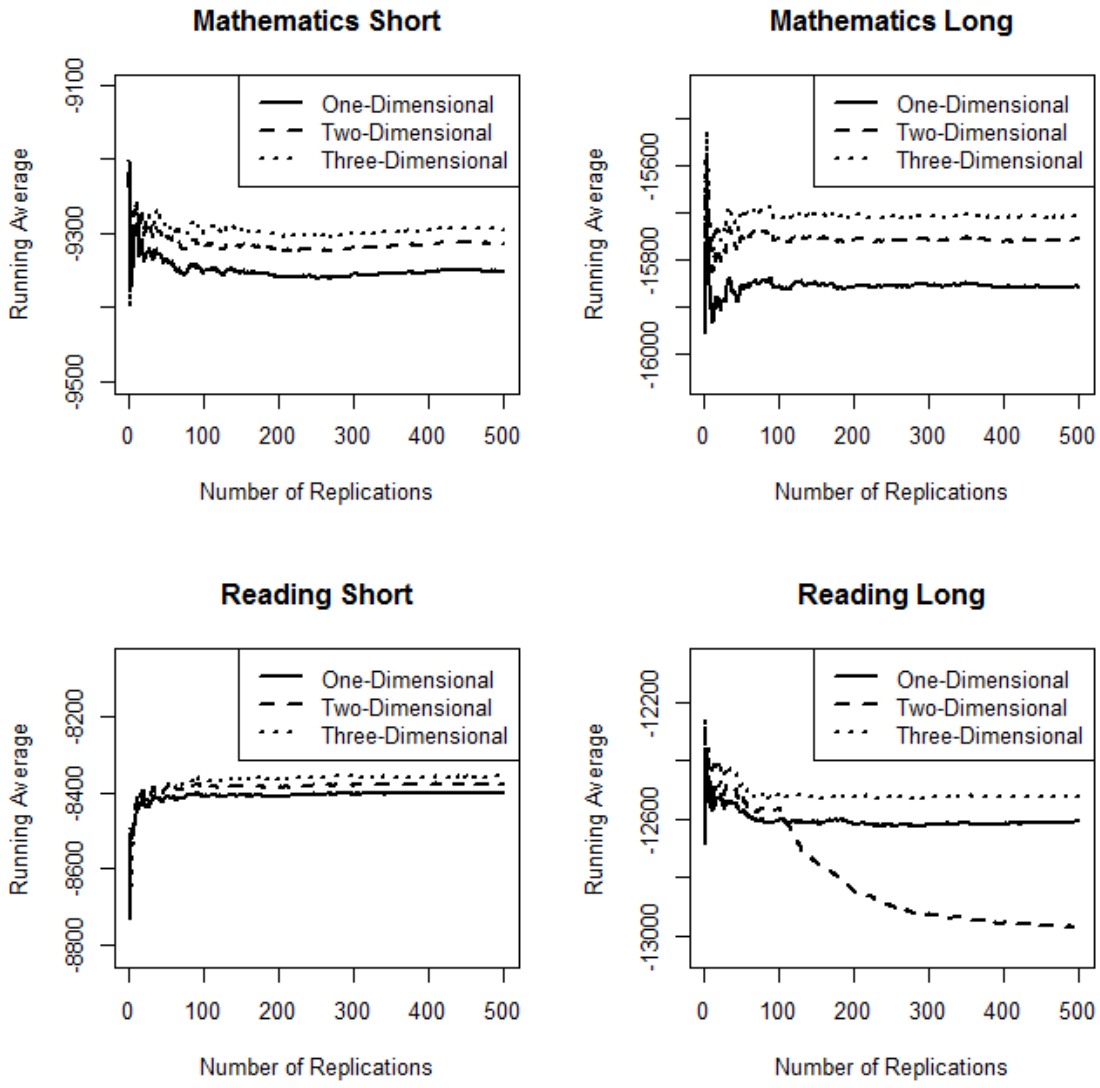


Figure 32A. Running Average Marginal Maximum Likelihood Value (Mplus MLR) for One-, Two-, and Three-dimensional Solutions when Sample Size is 1000



APPENDIX B: R Routines Used in the Study

R Routine to Generate Dichotomous Data with Major and Minor Latent Factors

```
gen.data <- function(N,n,major,cor,nminor,minor.cont) {

  ############################################################################
  # Inputs
  #   N            is sample size
  #   n            is number of items
  #   major        is a vector of length M, where M is the number of major factors
  #                each element indicates the variance accounted by the major factor
  #                (e.g.,0.05,0.20)
  #   cor          correlation among the major factors
  #   nminor       number of minor factors
  #   minor.cont   total variance accounted by minor factors
  ################################################################

  # Generate the factor loading matrix for major factors
  # If the major factor accounts
  # for  5%, generate numbers from a uniform distribution [.14,.30]
  # for 10%, generate numbers from a uniform distribution [.20,.42]
  # for 20%, generate numbers from a uniform distribution [.20,.65]
  # for 30%, generate numbers from a uniform distribution [.32,.75]
  # for 40%, generate numbers from a uniform distribution [.32,.90]

  NM = length(major)  # number of major factors

  major.f <- matrix(nrow=n,ncol=NM)

  for(i in 1:NM) {

    if(major[i]==.05)     { major.f[,i] = runif(n,.14,.30) } else
      if(major[i]==.10)    { major.f[,i] = runif(n,.20,.42) } else
        if(major[i]==.20)   { major.f[,i] = runif(n,.20,.65) } else
          if(major[i]==.30)  { major.f[,i] = runif(n,.32,.75) } else
            if(major[i]==.40) { major.f[,i] = runif(n,.32,.90) }

  }

  # Because of small sample size, the sum of squared loadings
  # do not always equal to the desired level of variance
  # accounted by major factors, they deviate a little bit
  # so we adjust them.

  major.f = sqrt((major.f^2) *
                 matrix((major/(colSums(major.f^2)/n)),nrow=n,ncol=NM,byrow=TRUE))

    # For each item check the variance accounted by major factors
    # It can't be more than (1-minor.cont)
    # Check communality, proportionally reduce the loadings in a
    # row, so the sum of squared loading will be smaller than (1-minor.cont)

  communality <- rowSums(major.f^2)
  limit = 1- minor.cont - .01
  reduce <- which(communality > limit)

  if(length(reduce)!=0) {
    for(i in 1:length(reduce)){
   major.f[reduce[i],]=sqrt((major.f[reduce[i],]^2)*(limit/sum((major.f[reduce[i],]^2))))
    }
  }
```

```
  # Generate the factor loading matrix for minor factors

  W <- matrix(nrow=n,ncol=nminor)

  minor.var <- .9^(0:(nminor-1))
  for(i in 1:nminor) {
    W[,i]=rnorm(nrow(W),0,sqrt(minor.var[i]))
  }

  re.scale <- matrix(minor.cont/rowSums(W^2),nrow=nrow(W),ncol=ncol(W),byrow=FALSE)
  scaled.W <- sqrt(W^2*re.scale)

  # Mega factor loading matrix

  F = cbind(major.f,scaled.W)

  # Uniqueness matrix

  D <- diag(sqrt(1-rowSums(F^2)))

  # Generate Factor Score Matrices

  cor.major = diag(length(major))
  cor.major[lower.tri(cor.major)] = rep(cor,length(major)*(length(major)-1)/2)
  cor.major[upper.tri(cor.major)] = rep(cor,length(major)*(length(major)-1)/2)

  scores.major  <- mvrnorm(N,mu=rep(0,length(major)),Sigma=cor.major)
  scores.minor  <- mvrnorm(N,mu=rep(0,nminor),Sigma=diag(nminor))
  mega.scores   <- cbind(scores.major,scores.minor)

  scores.unique <- mvrnorm(N,mu=rep(0,ncol(D)),Sigma=diag(ncol(D)))

  responses <- as.data.frame(mega.scores%*%t(F)+scores.unique%*%D)

  dif <- runif(n,-1.28,1.28)
  for(j in 1:n) {
    responses[,j]=ifelse(responses[,j]<dif[j],0,1)
  }

  return(list(responses=responses,loadings=major.f,minor=scaled.W))
}
```

R Routine to Compute Tetrachoric Correlation Matrix using TESTFACT

```
tetcor <- function(sample.data,name) {

  # For a given data matrix, this function writes the data in the format
  # "11A1,5X,nA1",creates TESTFACT input syntax for the data file,
  # acccess and runs TESTFACT, and returns unsmoothed tetrachoric
  # correlation matrix

  # INPUT
   # sample.data  ----  data matrix with dichotomous outcomes
   # name         ----  any name for TESTFACT files (e.g.,"sample1")

  #Set the directory you would like to store the results before
  #running the function

  #require "gdata" package

  write.fwf(x=cbind(1:nrow(sample.data)," ",sample.data),
            file=paste(getwd(),"/",name,".TESTFACT.txt",sep=""),
            width=c(11,5,rep(1,ncol(sample.data))),
            rownames=FALSE,colnames=FALSE,sep=""
  )

  ctl <- c(">TITLE")
  ctl <- rbind(ctl,"DATA")
  ctl <- rbind(ctl,"             ITEM and TEST STATISTICS")
  ctl <- rbind(ctl,paste(">PROBLEM","
NITEMS=",ncol(sample.data),",RESPONSE=3;",sep=""))

  nitem <- ncol(sample.data)
  r <- nitem%/%10
  k <- nitem%%10

  if(r==0) {
    a <- c()
    for(j in 1:k) { a <- c(a,paste("I",j,sep="")) }
    ctl <- rbind(ctl,paste(">NAMES       ",paste(a,collapse=","),";",sep=""))
  } else
    if(r==1 & k==0) {
      a <- c()
      for(j in 1:10) { a <- c(a,paste("I",j,sep="")) }
      ctl <- rbind(ctl,paste(">NAMES       ",paste(a,collapse=","),";",sep=""))
    } else
      if(r==1 & k!=0) {
        a <- c()
        for(j in 1:10) { a <- c(a,paste("I",j,sep="")) }
        ctl <- rbind(ctl,paste(">NAMES       ",paste(a,collapse=","),",",sep=""))
        b <- c()
        for(j in 1:k) { b <- c(b,paste("I",j,sep="")) }
        ctl <- rbind(ctl,paste("             ",paste(b,collapse=","),";",sep=""))
      } else

        if(r>1){
          for(u in 1:(r-1)){
            if(u==1){
              a <- c()
              for(j in 1:10) { a <- c(a,paste("I",j,sep="")) }
              ctl <- rbind(ctl,paste(">NAMES       ",paste(a,collapse=","),",",sep=""))
            } else {
              a <- c()
              for(m in 10:1) { a <- c(a,paste("I",u*10-m+1,sep="")) }
```

```
              ctl <- rbind(ctl,paste("             ",paste(a,collapse=","),",",sep=""))
            }
          }


          if(k==0) {
            a <- c()
            for(j in 10:1) { a <- c(a,paste("I",r*10-j+1,sep="")) }
            ctl <- rbind(ctl,paste("              ",paste(a,collapse=","),";",sep=""))
          } else {
            a <- c()
            for(j in 10:1) { a <- c(a,paste("I",r*10-j+1,sep="")) }
            ctl <- rbind(ctl,paste("             ",paste(a,collapse=","),",",sep=""))
            b <- c()
            for(j in 1:k) { b <- c(b,paste("I",r*10+j,sep="")) }
            ctl <- rbind(ctl,paste("             ",paste(b,collapse=","),";",sep=""))
          }
        }

  ctl <- rbind(ctl,paste(">RESPONSE    ' ','0','1';",sep=""))
  res <- rep(1,nitem)
  ctl <- rbind(ctl,paste(">KEY         ",paste(res,collapse=""),";",sep=""))
  ctl <- rbind(ctl,paste(">TETRACHORIC PAIRWISE, TIME, LIST, NDEC=8;",sep=""))
  ctl <- rbind(ctl,paste(">SAVE        CORRELAT;",sep=""))
  ctl <- rbind(ctl,paste(">INPUT
NIDCHAR=11,SCORES,FILE='",paste(name,".TESTFACT.txt",sep=""),"';",sep=""))
  ctl <- rbind(ctl,paste("(11A1,5X,",nitem,"A1)",sep=""))
  ctl <- rbind(ctl,paste(">STOP;",sep=""))
  ctl <- noquote(ctl)

  write(ctl,paste(name,".TSF",sep=""))
  system(paste('"C:/Program Files/TESTFACT4/TSF.exe"',paste(getwd(),"/",name,sep="")))

  cor <- scan(paste(name,".COR",sep=""))

  correlations <- matrix(nrow=ncol(sample.data),ncol=ncol(sample.data))

  for(i in 1:ncol(sample.data)){
    correlations[i,1:length((((i*(i-1)/2)+1)):((i*(i+1))/2))]=cor[(((i*(i-
1)/2)+1)):((i*(i+1))/2)]
  }

  for(i in 1:ncol(sample.data)) { correlations[i,]=correlations[,i]}

  return(correlations)
}
```

R Routine to do MINRES Factor Analysis using TESTFACT

```
minres <- function(sample.data,name,nfac) {

  # For a given data matrix, this function writes the data in the format
  # "11A1,5X,nA1",creates TESTFACT input syntax for the data file,
  # acccess and runs TESTFACT, and returns unrotated MINRES factor loadings

  # INPUT
   # sample.data  ----  data matrix with dichotomous outcomes
   # name         ----  any name for TESTFACT files (e.g.,"sample1")
   # nfac         ----  number of factors to be extracted

  #Set the directory you would like to store the results before
  #running the function

  #require "gdata" package

  write.fwf(x=cbind(1:nrow(sample.data)," ",sample.data),
            file=paste(getwd(),"/",name,".TESTFACT.txt",sep=""),
            width=c(11,5,rep(1,ncol(sample.data))),
            rownames=FALSE,colnames=FALSE,sep=""
  )

  ctl <- c(">TITLE")
  ctl <- rbind(ctl,"DATA")
  ctl <- rbind(ctl,"             ITEM and TEST STATISTICS")
  ctl <- rbind(ctl,paste(">PROBLEM","
NITEMS=",ncol(sample.data),",RESPONSE=3;",sep=""))

  nitem <- ncol(sample.data)
  r <- nitem%/%10
  k <- nitem%%10

  if(r==0) {
    a <- c()
    for(j in 1:k) { a <- c(a,paste("I",j,sep="")) }
    ctl <- rbind(ctl,paste(">NAMES       ",paste(a,collapse=","),";",sep=""))
  } else

    if(r==1 & k==0) {
      a <- c()
      for(j in 1:10) { a <- c(a,paste("I",j,sep="")) }
      ctl <- rbind(ctl,paste(">NAMES       ",paste(a,collapse=","),";",sep=""))
    } else

      if(r==1 & k!=0) {
        a <- c()
        for(j in 1:10) { a <- c(a,paste("I",j,sep="")) }
        ctl <- rbind(ctl,paste(">NAMES       ",paste(a,collapse=","),",",sep=""))
        b <- c()
        for(j in 1:k) { b <- c(b,paste("I",j,sep="")) }
        ctl <- rbind(ctl,paste("             ",paste(b,collapse=","),";",sep=""))
      } else

        if(r>1){
          for(u in 1:(r-1)){
            if(u==1){
              a <- c()
              for(j in 1:10) { a <- c(a,paste("I",j,sep="")) }
              ctl <- rbind(ctl,paste(">NAMES       ",paste(a,collapse=","),",",sep=""))
            } else {
```

```
              a <- c()
              for(m in 10:1) { a <- c(a,paste("I",u*10-m+1,sep="")) }
              ctl <- rbind(ctl,paste("             ",paste(a,collapse=","),",",sep=""))
            }
          }

          if(k==0) {
            a <- c()
            for(j in 10:1) { a <- c(a,paste("I",r*10-j+1,sep="")) }
            ctl <- rbind(ctl,paste("              ",paste(a,collapse=","),";",sep=""))
          } else {
            a <- c()
            for(j in 10:1) { a <- c(a,paste("I",r*10-j+1,sep="")) }
            ctl <- rbind(ctl,paste("             ",paste(a,collapse=","),",",sep=""))
            b <- c()
            for(j in 1:k) { b <- c(b,paste("I",r*10+j,sep="")) }
            ctl <- rbind(ctl,paste("             ",paste(b,collapse=","),";",sep=""))
          }
        }

  ctl <- rbind(ctl,paste(">RESPONSE    ' ','0','1';",sep=""))
  res <- rep(1,nitem)
  ctl <- rbind(ctl,paste(">KEY         ",paste(res,collapse=""),";",sep=""))
  ctl <- rbind(ctl,paste(">FACTOR NFAC=",nfac,", NIT=5;",sep=""))
  ctl <- rbind(ctl,paste(">SAVE        UNROTATED;",sep=""))
  ctl <- rbind(ctl,paste(">INPUT
NIDCHAR=11,SCORES,FILE='",paste(name,".TESTFACT.txt",sep=""),"';",sep=""))
  ctl <- rbind(ctl,paste("(11A1,5X,",nitem,"A1)",sep=""))
  ctl <- rbind(ctl,paste(">STOP;",sep=""))
  ctl <- noquote(ctl)

  write(ctl,paste(name,".TSF",sep=""))
  system(paste('"C:/Program Files/TESTFACT4/TSF.exe"',paste(getwd(),"/",name,sep="")))

  UNR <- read.table(paste(name,".UNR",sep=""),skip=1)

  return(UNR)
}
```

R Routine to Implement Parallel Analysis

```
PA <- function(sample.data,replication,name) {
  ###################################################################
  #Implements Parallel analysis procedure
  #1) simulate multivariate normal data with
  #independent variables
  #2) Dichotomize the simulated data such that the means
  #of variables are equal to the item difficulties
  #in the sample
  #3) Compute the eigenvalues from tetrachoric correlations
  #obtained from multivariate simulated independent binary
  #data
  #4) Repeat from step 1 to step 3 as many as the number
  #of replications
  ###################################################################
  # INPUT
  # sample.data  ----  data matrix with dichotomous outcomes under investigation
  # name         ----  any name for TESTFACT files (e.g.,"sample1")
  # replication  ----  number of random datasets to create empirical
  #                    sampling distribution of random data eigenvalues
  ###################################################################
  cuts <- qnorm(1-colMeans(sample.data,na.rm=TRUE))
  sample.corr <-
cor.smooth(tetcor(sample.data=sample.data,name=paste("pdata_",name,sep="")))
  EVs  <- matrix(nrow=ncol(sample.data),ncol=replication)
  file.remove(paste("pdata_",name,".TSF",sep=""))
  file.remove(paste("pdata_",name,".COR",sep=""))
  file.remove(paste("pdata_",name,".OUT",sep=""))
  file.remove(paste("pdata_",name,".TESTFACT.txt",sep=""))

  for(j in 1:replication) {
    pdata <-
mvrnorm(nrow(sample.data),mu=rep(0,ncol(sample.data)),Sigma=diag(ncol(sample.data)))
    pdic <- pdata
    for(i in 1:ncol(sample.data)){
      pdic[pdata[,i]<cuts[i],i]=0
      pdic[pdata[,i]>cuts[i],i]=1
      pdic <- as.data.frame(as.matrix(pdic))
    }
    corr <- cor.smooth(tetcor(sample.data=pdic,name=paste("pdata_",name,"_",j,sep="")))
    EVs[,j] <- eigen(corr)$values
    file.remove(paste("pdata_",name,"_",j,".TSF",sep=""))
    file.remove(paste("pdata_",name,"_",j,".COR",sep=""))
    file.remove(paste("pdata_",name,"_",j,".OUT",sep=""))
    file.remove(paste("pdata_",name,"_",j,".TESTFACT.txt",sep=""))
  }

  EVs2 <- matrix(nrow=ncol(sample.data),ncol=replication)
  for(u in 1:ncol(sample.data)){ EVs2[u,]=sort(EVs[u,])}
  PA.summary <- as.data.frame(matrix(nrow=ncol(sample.data),ncol=1))
  PA.summary[,1] <- 1:ncol(sample.data)
  PA.summary$Sample.Est <- eigen(sample.corr)$values
  PA.summary$Mean <- rowMeans(EVs2)
  PA.summary$Upper <- apply(EVs2,1,up <- function(d) { d[replication*.95]})
  write.fwf(round(PA.summary,3),paste(name,"_PAsummary.txt",sep=""),
            width=rep(15,4),rownames=FALSE,colnames=TRUE)
  return(sum(((PA.summary$Sample.Est>PA.summary$Upper)*1)[1:20]))
}
```

R Routine to Implement Revised Parallel Analysis

```
RPA <- function(sample.data,name,replication) {

  ############################################################################
  #Green, Levy, Thompson, Lu, & Lo.(2012). A Proposed Solution
  #to the Problem With Using Completely Random Data to Assess
  #the Number of Factors With Parallel Analysis.Educational and
  #Psychological Measurement, 72, 357.
  #
  #RPA requires a sequential simulations to test each eigenvalue separately
  #by conditioning on the size of previous eigenvalue
  ############################################################################
  # INPUT
  # sample.data  ----  data matrix with dichotomous outcomes under investigation
  # name         ----  any name for TESTFACT files (e.g.,"sample1")
  # replication  ----  number of random datasets to create empirical
  #                    sampling distribution of random data eigenvalues
  ####################################################################

  cuts <- qnorm(1-colMeans(sample.data,na.rm=TRUE))
  sample.corr <-
cor.smooth(tetcor(sample.data=sample.data,name=paste("rpdata0_",name,sep="")))
  sampleEV <- eigen(sample.corr)$values
  file.remove(paste("rpdata0_",name,".TSF",sep=""))
  file.remove(paste("rpdata0_",name,".COR",sep=""))
  file.remove(paste("rpdata0_",name,".OUT",sep=""))
  file.remove(paste("rpdata0_",name,".TESTFACT.txt",sep=""))

  RPA.summary <- as.data.frame(matrix(nrow=ncol(sample.data),ncol=1))
  RPA.summary[,1] <- 1:ncol(sample.data)
  RPA.summary$Sample.Est <- sampleEV

  #Test the first eigenvalue

  EVs  <- matrix(nrow=ncol(sample.data),ncol=replication)

  for(j in 1:replication) {
    pdata <-
mvrnorm(nrow(sample.data),mu=rep(0,ncol(sample.data)),Sigma=diag(ncol(sample.data)))
    pdic <- pdata
    for(i in 1:ncol(sample.data)){
      pdic[pdata[,i]<cuts[i],i]=0
      pdic[pdata[,i]>cuts[i],i]=1
      pdic <- as.data.frame(as.matrix(pdic))
    }
    corr <- cor.smooth(tetcor(sample.data=pdic,name=paste("rpdata0_",name,"_",j,sep="")))
    EVs[,j] <- eigen(corr)$values
    file.remove(paste("rpdata0_",name,"_",j,".TSF",sep=""))
    file.remove(paste("rpdata0_",name,"_",j,".COR",sep=""))
    file.remove(paste("rpdata0_",name,"_",j,".OUT",sep=""))
    file.remove(paste("rpdata0_",name,"_",j,".TESTFACT.txt",sep=""))
  }

  EVs2 <- matrix(nrow=ncol(sample.data),ncol=replication)
  for(u in 1:ncol(sample.data)){ EVs2[u,]=sort(EVs[u,])}

  RPA.summary$Mean <- rowMeans(EVs2)
  RPA.summary$Upper <- apply(EVs2,1,up <- function(d) { d[replication*.95]})
  RPA.summary <- cbind(1,RPA.summary)
```

```
  if(sampleEV[1]>EVs2[1,ceiling(replication*.95)]) iter=1 else iter=0

#Test later eigenvalues in a step wise approach

  if(iter!=0) {

    while(sampleEV[iter]>EVs2[iter,ceiling(replication*.95)]) {

      load   <- minres(sample.data,paste(name,iter,sep=""),nfac=iter)
      file.remove(paste(name,iter,".TSF",sep=""))
      file.remove(paste(name,iter,".UNR",sep=""))
      file.remove(paste(name,iter,".OUT",sep=""))
      file.remove(paste(name,iter,".TESTFACT.txt",sep=""))
      problem <- which(rowSums(load^2)>1)
      while(length(problem)>0) {
        for(i in 1:length(problem)) {
          load[problem[i],which.max(load[problem[i],])]=
            load[problem[i],which.max(load[problem[i],])]-.02
        }
        problem <- which(rowSums(load^2)>1)
      }
      uniq  <- diag(sqrt(1-rowSums(load^2)))

      REV.summary <- as.data.frame(matrix(nrow=ncol(sample.data),ncol=1))
      REV.summary[,1] <- 1:ncol(sample.data)
      REV.summary$Sample.Est <- sampleEV
      REVs <- matrix(nrow=ncol(sample.data),ncol=replication)

      for(u in 1:replication) {
        score <- mvrnorm(nrow(sample.data),mu=rep(0,iter),Sigma=diag(iter))
        error <-
mvrnorm(nrow(sample.data),mu=rep(0,ncol(sample.data)),Sigma=diag(ncol(sample.data)))
        pdata <- score%*%t(load)+error%*%uniq
        pdic <- pdata

        for(k in 1:ncol(sample.data)){
          pdic[pdata[,k]<cuts[k],k]=0
          pdic[pdata[,k]>cuts[k],k]=1
          pdic <- as.data.frame(as.matrix(pdic))
        }

        corr <-
cor.smooth(tetcor(sample.data=pdic,name=paste("rpdata",iter,"_",name,"_",u,sep="")))
        REVs[,u] <- eigen(corr)$values
        file.remove(paste("rpdata",iter,"_",name,"_",u,".TSF",sep=""))
        file.remove(paste("rpdata",iter,"_",name,"_",u,".COR",sep=""))
        file.remove(paste("rpdata",iter,"_",name,"_",u,".OUT",sep=""))
        file.remove(paste("rpdata",iter,"_",name,"_",u,".TESTFACT.txt",sep=""))

      }

      EVs2 <- matrix(nrow=ncol(sample.data),ncol=replication)
      for(m in 1:ncol(sample.data)){ EVs2[m,]=sort(REVs[m,])}

      REV.summary$Mean <- rowMeans(EVs2)
      REV.summary$Upper <- apply(EVs2,1,up <- function(d) { d[replication*.95]})
      REV.summary <- cbind(iter+1,REV.summary)
      colnames(REV.summary) <- colnames(RPA.summary)
      RPA.summary <- rbind(RPA.summary,REV.summary)
      iter <- iter+1
      dim <- iter-1
    }
    write.fwf(round(RPA.summary,3),paste(name,"_RPAsummary.txt",sep=""),
```

```
              width=rep(15,5),rownames=FALSE,colnames=TRUE)
    return(dim)
  } else {
    write.fwf(round(RPA.summary,3),paste(name,"_RPAsummary.txt",sep=""),
              width=rep(15,5),rownames=FALSE,colnames=TRUE)
    return(iter)
  }
}
```

R Routine to Implement the DETECT Procedure

```
DETECT <- function(sample.data,name) {

  # INPUT
   # sample.data  ----  data matrix with dichotomous outcomes under investigation
   # name         ----  any name for DETECT files (e.g.,"sample1")

  # Writes the input file for DETECT program, runs DETECT through R,
  # returns the output file

  write.fwf(sample.data,paste(name,".detect.dat",sep=""),
            width=rep(1,ncol(sample.data)),na=" ",sep="",
            rownames=FALSE,colnames=FALSE)


  ctl <- c("name of data file")
  ctl <- rbind(ctl,paste(getwd(),"/",name,".detect.dat",sep=""))
  ctl <- rbind(ctl,c("no.of items"))
  ctl <- rbind(ctl,ncol(sample.data))
  ctl <- rbind(ctl,c("no.of examinees"))
  ctl <- rbind(ctl,nrow(sample.data))
  ctl <- rbind(ctl,c("mincell"))
  ctl <- rbind(ctl,2)
  ctl <- rbind(ctl,c("mutations"))
  ctl <- rbind(ctl,ncol(sample.data)/5)
  ctl <- rbind(ctl,c("max dimensions"))
  ctl <- rbind(ctl,12)
  ctl <- rbind(ctl,c("dropflag"))
  ctl <- rbind(ctl,0)
  ctl <- rbind(ctl,c("no.of items to drop from the analysis"))
  ctl <- rbind(ctl,0)
  ctl <- rbind(ctl,c("items to be dropped"))
  ctl <- rbind(ctl,0)
  ctl <- rbind(ctl,c("confimatory flag"))
  ctl <- rbind(ctl,0)
  ctl <- rbind(ctl,c("crosflag"))
  ctl <- rbind(ctl,0)
  ctl <- rbind(ctl,c("no.of examinees to set aside for cross validation"))
  ctl <- rbind(ctl,0)
  ctl <- rbind(ctl,c("seed"))
  ctl <- rbind(ctl,99991 )
  ctl <- rbind(ctl,c("name of detect summary output file"))
  ctl <- rbind(ctl,paste(getwd(),"/",name,".det",sep=""))
  ctl <- rbind(ctl,c("cluster output flag"))
  ctl <- rbind(ctl,0)
  ctl <- rbind(ctl,c("covariance output flag"))
  ctl <- rbind(ctl,0)
  ctl <- noquote(ctl)

  write(ctl,paste(getwd(),"/detect.in",sep=""))
  system(paste('"C:/Program Files
(x86)/Dimpack1.0/detect4.exe"',paste(getwd(),"/detect",sep="")))

  outputdetect <- scan(paste(name,".det",sep=""),what=c("raw"))
  return(as.numeric
         (outputdetect[which(outputdetect=="MAXIMIZE")+2])
  )
}
```

R Routine to Create Noharm Input Files

```
NOHARM.input <-  function(sample.data,name,nfac) {

#####################################################################
# INPUT
   # sample.data  ----  data matrix with dichotomous outcomes under investigation
   # name         ----  any name for Noharm files (e.g.,"sample1")
   # nfac         ----  maximum number of factors to be extracted
#####################################################################
# Creates and writes necessary NOHARM input files for a given data matrix
# to fit up to the m-dimensional model (m=nfac),
#####################################################################

  prod <- round((t(as.matrix(sample.data))%*%as.matrix(sample.data))/nrow(sample.data),4)
  guessing=0

  for(u in 1:nfac){

    ctl <- c("NOHARM ANALYSIS")
    ctl <- rbind(ctl,paste(ncol(sample.data),u,nrow(sample.data),1,1,0,1,0,sep=" "))
    ctl <- rbind(ctl,paste(rep(guessing,ncol(sample.data)),collapse=" "))
    for(i in 1:ncol(sample.data)){
      ctl <- rbind(ctl,paste(prod[i,1:i],collapse=" "))
    }
    write(ctl,paste(name,"_",u,"FAC",".inp",sep=""))
  }
}
```

R Routine to Create Noharm Batch File

```
# nfac ---- maximum number of factors to be extracted for each dataset
# rep  ---- number of replications (simulated datasets)

batch <- paste(paste("data1_1FAC",".inp",sep=""),
               paste("data1_1FAC.out",sep=""),collapse=" ")

for(i in 2:nfac){
  batch <- rbind(batch,
                 paste(paste("data1_",i,"FAC",".inp",sep=""),
                       paste("data1_",i,"FAC.out",sep=""),collapse=" "))

}

for(j in 2:rep){

  for(i in 1:nfac){
    batch <- rbind(batch,
                   paste(paste("data",j,"_",i,"FAC",".inp",sep=""),
                         paste("data",j,"_",i,"FAC.out",sep=""),collapse=" "))

  }
}

write(batch,"batch.inp")
```

R Routine to Compute Noharm Approximate Chi-square Statistic

```
NOHARM.achi <- function(sample.data,name,nfac) {

#####################################################################
# INPUT
   # sample.data  ----  data matrix with dichotomous outcomes under investigation
   # name         ----  any name for Noharm files (e.g.,"sample1")
   # nfac         ----  maximum number of factors to be extracted
#####################################################################
# Reads Noharm output files, extracts necessary information, and
# computes Noharm Achi statistic, creates a summary output table for the
# multidimensional models up to "nfac" dimensions
#####################################################################

  rescor <- vector("list",nfac)

  for(u in 1:nfac) {
    res <- scan(paste("RES_",name,"_",u,"FAC.out",sep=""))
    res2 <- matrix(nrow=ncol(sample.data),ncol=ncol(sample.data))
    for(i in 1:(ncol(sample.data)-1)){
      res2[i+1,1:length((((i*(i-1)/2)+1)):((i*(i+1))/2))]=
        res[(((i*(i-1)/2)+1)):((i*(i+1))/2)]
    }
    propcor <- colMeans(sample.data,na.rm=TRUE)
    variances <- propcor*(1-propcor)
    varprod <- matrix(nrow=ncol(sample.data),ncol=ncol(sample.data))
    for(i in 1:(ncol(sample.data)-1)){
      for(j in (i+1):ncol(sample.data)) {
        varprod[j,i]=sqrt(variances[i]*variances[j])
      }}
    rescor[[u]] <- matrix(nrow=ncol(sample.data),ncol=ncol(sample.data))
    for(i in 1:(ncol(sample.data)-1)){
      for(j in (i+1):ncol(sample.data)){
        rescor[[u]][j,i]=res2[j,i]/varprod[j,i]
      }}
  }
  Achi <- c()
  for(u in 1:nfac){
    rescor2 <- rescor[[u]][lower.tri(rescor[[u]])]
    Achi[u]=(nrow(sample.data)-3)*(sum(((log((1+rescor2)/(1-rescor2)))/2)^2))
  }
  df <- .5*ncol(sample.data)*(ncol(sample.data)-1)-(ncol(sample.data)*(1:nfac)-
    ((0:(nfac-1))*(1:nfac)/2))
  Chisq <- c()
  for(i in 1:nfac) {
    Chisq[i]=round(pchisq(Achi[i],df[i],lower.tail=FALSE),3)
  }

  if(length(which(Chisq>.05))!=0) { CHI1=which(Chisq>.05)[1] } else CHI1=NA
  output <- as.data.frame(matrix(nrow=length(na.omit(Achi)),ncol=1))
  output[,1] <- 1:length(na.omit(Achi))
  output$Chi.Sq <- Achi
  output$df <- df
  output$p <- Chisq

  write.fwf(round(output,3),paste(name,"_NOHARM_Achi.txt",sep=""),
            width=rep(15,4),rownames=FALSE,colnames=TRUE)
  return(list(Chi.square=CHI1))
}
```

R Routine to Compute Noharm Approximate Likelihood Ratio Chi-Square Statistic

```
NOHARM.alr <- function(sample.data,name,nfac) {
  ######################################################################
  # INPUT
  # sample.data  ----  data matrix with dichotomous outcomes under investigation
  # name         ----  any name for Noharm files (e.g.,"sample1")
  # nfac         ----  maximum number of factors to be extracted
  ######################################################################
  # Reads Noharm output files, extracts necessary information, and
  # computes Noharm ALR statistic, creates a summary output table for the
  # multidimensional models up to "nfac" dimensions
  ######################################################################

  obsprop  <- (t(as.matrix(sample.data))%*%as.matrix(sample.data))/nrow(sample.data)
  resprop  <- vector("list",15)
  predprop <- vector("list",15)
  Gratio   <- vector("list",15)
  Gsquare  <- c()

  for(u in 1:nfac) {
    res <- scan(paste("RES_",name,"_",u,"FAC.out",sep=""))
    res2 <- matrix(nrow=ncol(sample.data),ncol=ncol(sample.data))
    for(i in 1:(ncol(sample.data)-1)){
      res2[i+1,1:length((((i*(i-1)/2)+1)):((i*(i+1))/2))]=
        res[(((i*(i-1)/2)+1)):((i*(i+1))/2)]
    }
    resprop[[u]] <- res2
    predprop[[u]]<- obsprop-resprop[[u]]

    temp <- readLines(paste(name,"_",u,"FAC.out",sep=""))
    start.row <- which(temp=="Final Coefficients of Theta")+5

    betas <- vector("list",(((u*ncol(sample.data)-1)%/%(9*ncol(sample.data)))+1))
    for(i in 1:(((u*ncol(sample.data)-1)%/%(9*ncol(sample.data)))+1)){
      betas[[i]] <- scan(paste(name,"_",u,"FAC.out",sep=""),
                         skip=(start.row-1)+((i-1)*ncol(sample.data))+((i-1)*3),
                         nlines=ncol(sample.data))
      betas[[i]] <- matrix(betas[[i]],nrow=ncol(sample.data),byrow=TRUE)
      betas[[i]] <- betas[[i]][,2:ncol(betas[[i]])]
    }

    beta1 <- c()
    for(i in 1:(length(betas))){
      beta1 <- cbind(beta1,betas[[i]])
    }

    start.row <- which(temp=="Final Constants")+5
    temp2 <- vector("list",ncol(sample.data)/10)

    for(i in 1:(ncol(sample.data)/10)){
      temp2[[i]] <- scan(paste(name,"_",u,"FAC.out",sep=""),
                         skip=start.row-1+4*(i-1), nlines=1)
    }

    beta.0 <- as.matrix(temp2[[1]])

    for(i in 1:((ncol(sample.data)/10)-1)){
      beta.0 <- cbind(beta.0,temp2[[i+1]])
    }
```

```
    beta.0 <- matrix(beta.0,nrow=ncol(sample.data))
    sigma <- diag(u)
    diag(predprop[[u]]) <- pnorm(beta.0/sqrt(1+diag(beta1%*%sigma%*%t(beta1))))
    Gratio[[u]] <- matrix(nrow=ncol(sample.data),ncol=ncol(sample.data))

    for(i in 1:(ncol(sample.data)-1)){
      for(j in (i+1):ncol(sample.data)){
        table1 <- matrix(nrow=2,ncol=2)
        table1[2,2]<- predprop[[u]][j,i]
        table1[2,1]<- predprop[[u]][i,i]-table1[2,2]
        table1[1,2]<- predprop[[u]][j,j]-table1[2,2]
        table1[1,1]<- (1-predprop[[u]][j,j])-table1[2,1]
        table2 <- matrix(nrow=2,ncol=2)
        table2[2,2]<- obsprop[j,i]
        table2[2,1]<- obsprop[i,i]-table2[2,2]
        table2[1,2]<- obsprop[j,j]-table2[2,2]
        table2[1,1]<- (1-obsprop[i,i])-table2[1,2]
        Gratio[[u]][[j,i]]=-2*sum(table2*log(table1/table2))
      }
    }
    Gsquare[u] <- nrow(sample.data)*sum(Gratio[[u]],na.rm=TRUE)
  }

  df <- .5*ncol(sample.data)*(ncol(sample.data)-1)-(ncol(sample.data)*(1:nfac)
                                                    -((0:(nfac-1))*(1:nfac)/2))

  Chisq <- c()
  for(i in 1:nfac) {
    Chisq[i]=round(pchisq(Gsquare[i],df[i],lower.tail=FALSE),3)
  }

  if(length(which(Chisq>.05))!=0) { CHI1=which(Chisq>.05)[1] } else CHI1=NA

  output <- as.data.frame(matrix(nrow=length(na.omit(Gsquare)),ncol=1))
  output[,1] <- 1:length(na.omit(Gsquare))
  output$Chi.Sq <- Gsquare
  output$df <- df
  output$p <- Chisq

  write.fwf(round(output,3),paste(name,"_NOHARM_ALR.txt",sep=""),
            width=rep(15,4),rownames=FALSE,colnames=TRUE)

  return(list(Chi.square=CHI1))
}
```

R Routine to Compute Noharm Scaled Chi-Square Statistics (Mean adjusted & Mean-and-Variance Adjusted)

```
Noharm.T <- function(sample.data,name,nfac){

  ######################################################################
  # INPUT
  # sample.data  ----  data matrix with dichotomous outcomes under investigation
  # name         ----  any name for Noharm files (e.g.,"sample1")
  # nfac         ----  maximum number of factors to be extracted
  ######################################################################
  # Reads Noharm output files, extracts necessary information, and
  # computes mean adjusted and mean-and-variance adjusted chi-square statistics, creates
  # a summary output table for the multidimensional models up to "nfac" dimensions
  ######################################################################
  #Albert Maydeu-Olivares(2001).Multidimensional Item Response Theory Modeling of
  #         Binary Data: Large Sample Properties of NOHARM Estimates.Journal of
  #         Educational and Behavioral Statistics,26,pp. 51-71
  ######################################################################

  require(fMultivar)

  TM <- c()
  TMV <- c()
  dfm <- c()
  dfmv <- c()

  yyprime <- t(as.matrix(sample.data[1,]))%*%as.matrix(sample.data[1,])
  off <- yyprime[2,1]
  for(j in 3:nrow(yyprime)){
    off <- c(off,as.vector(yyprime[j,1:(j-1)]))
  }
  di <- c(diag(yyprime),off)
  D <- as.matrix(di)%*%t(as.matrix(di))

  for(i in 2:nrow(sample.data)){
    yyprime <- t(as.matrix(sample.data[i,]))%*%as.matrix(sample.data[i,])
    off <- yyprime[2,1]
    for(j in 3:nrow(yyprime)){
      off <- c(off,as.vector(yyprime[j,1:(j-1)]))
    }
    di <- c(diag(yyprime),off)
    D <- D+as.matrix(di)%*%t(as.matrix(di))
  }

  thresholds <- qnorm(1-colMeans(sample.data))

  for(u in 1:nfac) {
    out <- readLines(paste(name,"_",u,"FAC.out",sep=""))

    #Read Loadings from output

    if(u<=9) {
      loadings <- as.matrix(matrix(scan(paste(name,"_",u,"FAC.out",sep=""),
                                        skip=which(out=="Factor Loadings")+4,
                                        nlines=ncol(sample.data)),
                                   nrow=ncol(sample.data),byrow=TRUE)[,2:(u+1)])
    }
```

```
    if(u>9 & u<=18) {

      loadings1 <- as.matrix(matrix(scan(paste(name,"_",u,"FAC.out",sep=""),
                                         skip=which(out=="Factor Loadings")+4,
                                         nlines=ncol(sample.data)),
                                         nrow=ncol(sample.data),byrow=TRUE)[,2:10])

      loadings2 <- matrix(scan(paste(name,"_",u,"FAC.out",sep=""),
                               skip=which(out=="Factor Loadings")+47,
                               nlines=ncol(sample.data)),
                               nrow=ncol(sample.data),byrow=TRUE)

      loadings <- cbind(loadings1,loadings2[,2:ncol(loadings2)])
    }

    if(u>18 & u<=27) {

      loadings1 <- as.matrix(matrix(scan(paste(name,"_",u,"FAC.out",sep=""),
                                         skip=which(out=="Factor Loadings")+4,
                                         nlines=ncol(sample.data)),
                                         nrow=ncol(sample.data),byrow=TRUE)[,2:10])

      loadings2 <- matrix(scan(paste(name,"_",u,"FAC.out",sep=""),
                               skip=which(out=="Factor Loadings")+47,
                               nlines=ncol(sample.data)),
                               nrow=ncol(sample.data),byrow=TRUE)

      loadings3 <- matrix(scan(paste(name,"_",u,"FAC.out",sep=""),
                               skip=which(out=="Factor Loadings")+90,
                               nlines=ncol(sample.data)),
                               nrow=ncol(sample.data),byrow=TRUE)
      loadings <-
cbind(loadings1,loadings2[,2:ncol(loadings2)],loadings3[,2:ncol(loadings3)])

    }

    # Predicted correlations

    pijs <- loadings%*%t(loadings)

    # Delta 11 matrix

    delta11 <- diag(-dnorm(thresholds))

    # Delta 21 matrix, Equation 25

    delta21 <- as.data.frame(t(combn(1:(ncol(sample.data)),2))
                             [order(t(combn(1:(ncol(sample.data)),2))[,2]),])

    delta21 <- as.data.frame(cbind(delta21[,2],delta21[,1],
                                   matrix(NA,nrow=ncol(combn(1:(ncol(sample.data)),2)),
                                          ncol=ncol(sample.data))))

    for(i in 1:ncol(combn(1:(ncol(sample.data)),2))){
      d <- delta21[i,1:2]
      ii <- delta21[i,1]
      jj <- delta21[i,2]
      for(rr in d){
        if(rr==ii) { delta21[i,delta21[i,1]+2]=
          delta11[ii,ii]*pnorm((-thresholds[jj]+(pijs[ii,jj]*
          thresholds[ii]))/sqrt(1-pijs[ii,jj]^2))
```

```
      } else
        if(rr==jj) { delta21[i,delta21[i,2]+2]=
          delta11[jj,jj]*pnorm((-thresholds[ii]+(pijs[ii,jj]*
          thresholds[jj]))/sqrt(1-pijs[ii,jj]^2))}
    }
    delta21[i,is.na(delta21[i,])]=0
  }

  delta21 <- delta21[,3:ncol(delta21)]

  # Delta 22 matrix

  delta.p <- matrix(NA,nrow=nrow(loadings),ncol=nrow(loadings))

  for(i in 1:(nrow(loadings)-1)){
    for(j in (i+1):nrow(loadings)){
      delta.p[i,j]=dnorm2d(thresholds[i],thresholds[j],pijs[i,j])
    }}

  delta.p <- diag(delta.p[upper.tri(delta.p)])

  l <- matrix(nrow=nrow(loadings),ncol=ncol(loadings))
  for(i in 1:nrow(loadings)){
    for(j in 1:ncol(loadings)){
      l[i,j]=paste("l",i,j,sep="")
    }}

  for(i in 1:nrow(loadings)){
    for(j in 1:ncol(loadings)){
      assign(paste("l",i,j,sep=""),loadings[i,j])
    }}

  delta.t <- as.data.frame(t(combn(1:(nrow(loadings)),2))
                           [order(t(combn(1:(nrow(loadings)),2))[,2]),])
  delta.t <- as.data.frame(cbind(delta.t[,2],delta.t[,1],
                                 matrix(NA,nrow=ncol(combn(1:nrow(loadings),2)),
                                        ncol=length(loadings))))

  for(i in 1:nrow(delta.t)){
    term1 <- vector("list",ncol(loadings))
    for(r in 1:length(term1)){
      term1[[r]] <- paste(l[delta.t[i,2],r],l[delta.t[i,1],r],sep="*")
    }

    delta.t[i,3:ncol(delta.t)] <- attr(numericDeriv
                                       (parse(text=paste0(term1,collapse="+"))[[1]],
                                        c(t(l),recursive=TRUE)),"gradient")
  }

  delta.t <- as.matrix(delta.t[,3:ncol(delta.t)])
  delta22 <- delta.p%*%delta.t

  C <- (t(as.matrix(sample.data))%*%as.matrix(sample.data))/nrow(sample.data)

  p <- C[2,1]

  for(i in 3:nrow(C)){ p <- c(p,as.vector(C[i,1:(i-1)])) }

  p <- c(diag(C),p)

  pprime <- p%*%t(p)

  Gamma <- (D/nrow(sample.data))-pprime
```

```
    if(length(which(is.na(delta11)==TRUE))==0 &
      length(which(is.na(delta21)==TRUE))==0 &
      length(which(is.na(delta22)==TRUE))==0 &
      length(which(is.na(Gamma)==TRUE))==0 ) {

      omega <- cbind((-as.matrix(delta21)%*%ginverse(delta11)),
                     diag(nrow(delta21)))%*%Gamma%*%t(cbind((-as.matrix(delta21)%*%
                       ginverse(delta11)),diag(nrow(delta21))))

      H <- diag(nrow(delta22))-(delta22%*%(ginverse(t(delta22)%*%delta22))%*%t(delta22))

      res <- scan(paste("RES_",name,"_",u,"FAC.out",sep=""))

      T <- nrow(sample.data)*sum(res^2)

      dfm[u] <- (ncol(sample.data)*(ncol(sample.data)-1)/2)-
        (length(which(loadings!=0)))
      TM[u] <- T*(dfm[u]/sum(diag(H%*%omega))) #mean adjusted
      TMV[u] <- T*sum(diag(H%*%omega))/sum(diag(H%*%omega%*%H%*%omega))
      dfmv[u] <- sum(diag(H%*%omega))^2/sum(diag(H%*%omega%*%H%*%omega))

    } else {

      dfm[u] <- NA
      TM[u] <- NA
      TMV[u] <- NA
      dfmv[u] <- NA
    }
  }

  ChisqM <- c()
  for(i in 1:nfac) {
    ChisqM[i]=round(pchisq(TM[i],dfm[i],lower.tail=FALSE),3)
  }

  ChisqMV <- c()
  for(i in 1:nfac) {
    ChisqMV[i]=round(pchisq(TMV[i],dfmv[i],lower.tail=FALSE),3)
  }

  output <- as.data.frame(matrix(nrow=length(ChisqMV),ncol=1))
  output[,1] <- 1:length(ChisqMV)
  output$TM <- TM
  output$dfm <- dfm
  output$ChisqM <- ChisqM
  output$TMV <- TMV
  output$dfmV <- dfmv
  output$ChisqMV <- ChisqMV

  write.fwf(round(output,3),paste(name,"_NOHARM_T.txt",sep=""),
            width=rep(15,7),rownames=FALSE,colnames=TRUE)

  if(length(which(ChisqM>.05))!=0) { CHIM=which(ChisqM>.05)[1] } else CHIM=NA
  if(length(which(ChisqMV>.05))!=0) { CHIMV=which(ChisqMV>.05)[1] } else CHIMV=NA
  return(list(ChiM=CHIM,ChiMV=CHIMV))
}
```

R Routine to Analyze a Dataset with the Mplus WLSM Estimator

```
WLSM <- function(sample.data,name,nfac) {
  ######################################################################
  # INPUT
  # sample.data  ----  data matrix with dichotomous outcomes under investigation
  # name         ----  any name for Mplus files (e.g.,"sample1")
  # nfac         ----  maximum number of factors to be extracted
  ######################################################################
  # Creates and writes Mplus input file to extract "m" factors with the WLSM estimator,
  # runs Mplus through R and analyze the given dataset, reads Mplus output files,
  # extracts necessary information for several fit indices, creates and writes a summary
  # output table
  ######################################################################
  write.fwf(sample.data,paste(name,".mplus.txt",sep=""),
            width=rep(2,ncol(sample.data)),
            na=" ",sep=".",rownames=FALSE,colnames=FALSE)

  ctl <- c("TITLE: EFA; ")
  ctl <- rbind(ctl,paste("DATA: FILE IS ",name,".mplus.txt",";",sep=""))
  ctl <- rbind(ctl,paste("VARIABLE: NAMES ARE y1-y",ncol(sample.data),
                         "; USEV=y1-y",ncol(sample.data),
                         ";CATEGORICAL=y1-y",ncol(sample.data),";",
                         sep=""))

  ctl <- rbind(ctl,paste("ANALYSIS: TYPE = EFA 1 ",nfac,";
                         ESTIMATOR=WLSM",";",sep=" "))       # "ESTIMATOR=WLSMV"
                                                             # for the WLSMV estimator

  write(ctl,paste(name,"WLSM.mplus",sep=""))
  system(paste('"C:/Program Files/Mplus/Mplus.exe"',
               paste(getwd(),"/",name,"WLSM.mplus",sep="")))

  output <- scan(paste(name,"WLSM.OUT",sep=""),what=c("raw"),fill=TRUE)
  no.converge <- as.numeric(output[which(output=="OCCURRED")+6])
  ts <- (1:nfac)[-no.converge]
  chi.num <- which(output=="Chi-Square")+6
  chi <- strsplit(output[which(output=="Chi-Square")+6]
                  [seq(from=1,to=length(chi.num),by=2)],"")
  chi.sq <- c()
  for(i in 1:length(chi)){
    chi.sq[i]=as.numeric(paste(chi[[i]][1:(length(chi[[i]])-1)],collapse=""))
  }

  dof <- as.numeric(output[which(output=="Chi-Square")+10]
                    [seq(from=1,to=length(chi.num),by=2)])
  sf <- as.numeric(output[which(output=="Chi-Square")+16]
                   [seq(from=1,to=length(chi.num),by=2)])
  Chisq <- c()
  for(i in 1:length(chi.sq)) {
    Chisq[i]=round(pchisq(chi.sq[i],dof[i],lower.tail=FALSE),3)
  }
  RMSEAs <- as.numeric(output[which(output=="C.I.")+1])
  RMSEAsII <- as.numeric(output[which(output=="Estimate")+1])
  diff <- (length(RMSEAsII)-length(RMSEAs))
  kl <- length(RMSEAs)
  if(length(RMSEAsII)!=length(RMSEAs)){
    for(i in 1:diff){
      RMSEAs[kl+i]=RMSEAsII[kl+i]
    }
```

```
  }
  SRMRs <- as.numeric(output[which(output=="SRMR")+7])
  CFIs <- as.numeric(output[which(output=="CFI")+1])

  if(length(no.converge)!=0){
    chi.sq2 <- c();for(h in 1:length(ts)){chi.sq2[ts[h]] <-
chi.sq[h]};chi.sq2[no.converge] <- NA
    dof2 <- c();for(h in 1:length(ts)){dof2[ts[h]] <- dof[h]};dof2[no.converge] <- NA
    sf2 <- c();for(h in 1:length(ts)){sf2[ts[h]] <- sf[h]};sf2[no.converge] <- NA
    Chisq2 <- c();for(h in 1:length(ts)){Chisq2[ts[h]] <- Chisq[h]};Chisq2[no.converge]
<- NA
    RMSEAs2 <- c();for(h in 1:length(ts)){ RMSEAs2[ts[h]] <-  RMSEAs[h]};
RMSEAs2[no.converge] <- NA
    SRMRs2 <- c();for(h in 1:length(ts)){ SRMRs2[ts[h]] <- SRMRs[h]}; SRMRs2[no.converge]
<- NA
    CFIs2 <- c();for(h in 1:length(ts)){ CFIs2[ts[h]] <- CFIs[h]}; CFIs2[no.converge] <-
NA
  } else {
    chi.sq2 <- chi.sq;dof2 <- dof ; sf2 <- sf; Chisq2 <- Chisq
    RMSEAs2 <- RMSEAs; SRMRs2 <- SRMRs; CFIs2 <- CFIs
  }

  if(length(which(Chisq2>.05))!=0) {
    CHI1=which(Chisq2>.05)[1]
  } else CHI1=NA
  RMSEA <- which(RMSEAs2 < .05)[1]
  SRMR <- which(SRMRs2 < .05)[1]
  CFI <- which(CFIs2 > .95)[1]

  output <- as.data.frame(matrix(nrow=length(chi.sq2),ncol=1))
  output[,1] <- 1:length(chi.sq2)
  output$Chi.Sq <- chi.sq2
  output$df <- dof2
  output$sf <- sf2
  output$p <- Chisq2
  output$RMSEA <- RMSEAs2
  output$SRMR  <- SRMRs2
  output$CFI  <- CFIs2

  write.fwf(round(output,3),paste(name,"_WLSM_mplus.summary.txt",sep=""),
            width=rep(15,8),rownames=FALSE,colnames=TRUE)

  return(list(Chi.square=CHI1,rmsea=RMSEA,srmr=SRMR,cfi=CFI))
}
```

R Routine to Analyze a Dataset with the Mplus MLR Estimator

```
MLR <- function(sample.data,name,nfac) {

  ######################################################################
  # INPUT
  # sample.data  ----  data matrix with dichotomous outcomes under investigation
  # name         ----  any name for Mplus files (e.g.,"sample1")
  # nfac         ----  maximum number of factors to be extracted
  ######################################################################
  # Creates and writes Mplus input file to fir m-dimensional model with the MLR
  # estimator, runs Mplus through R and analyze the given dataset, reads Mplus output
  # files, extracts necessary information for several fit indices, creates and writes a
  # summary output table
  ######################################################################
  write.fwf(sample.data,paste(name,".mplus.txt",sep=""),
            width=rep(2,ncol(sample.data)),
            na=" ",sep=".",rownames=FALSE,colnames=FALSE)


  dof <- c()
  sf <- c()
  p.val <- c()
  p.scaled <- c()
  AICs <- c()
  AICcs <- c()
  BICs <- c()
  LLs  <- c()
  ABICs <- c()

#####################################################
#Fit the first factor
  ctl <- c("TITLE: EFA; ")
  ctl <- rbind(ctl,paste("DATA: FILE IS ",name,".mplus.txt",";",sep=""))
  ctl <- rbind(ctl,paste("VARIABLE: NAMES ARE y1-y",ncol(sample.data),
                         "; USEV=y1-y",ncol(sample.data),
                         ";CATEGORICAL=y1-y",ncol(sample.data),";",
                         sep=""))
  ctl <- rbind(ctl,paste("ANALYSIS: TYPE = EFA 1 1; ESTIMATOR=MLR;
                         INTEGRATION=GAUSSHERMITE(7);LOGCRITERION =.01;
                         MITERATIONS =250 ",
                         ";",sep=" "))
  write(ctl,paste(name,"MLR1.mplus",sep=""))
  system(paste('"C:/Program Files/Mplus/Mplus.exe"',
               paste(getwd(),"/",name,"MLR1.mplus",sep=""),
               paste(getwd(),"/",name,"MLR1.out",sep="")
               )
         )
  output <- scan(paste(name,"MLR1.OUT",sep=""),what=c("raw"),fill=TRUE)
  LLs[1] <- as.numeric(output[which(output=="Loglikelihood")
                              [2:length(which(output=="Loglikelihood"))]+3])
  dof[1] <- as.numeric(output[which(output=="Free")+2])
  sf[1] <- as.numeric(output[which(output=="Scaling")+3])
  AICs[1] <- as.numeric(output[which(output=="Akaike")+2])
  nparameters = round(AICs[1]-(-2*LLs[1]),0)/2
  sample.size = nrow(sample.data)
  AICcs[1] <- AICs[1] + ((2*nparameters*(nparameters+1))/(sample.size-nparameters-1))
  BICs[1] <- as.numeric(output[which(output=="Bayesian")+2])
  ABICs[1] <- as.numeric(output[which(output=="Sample-Size")+3])
  p.val[1] <- 0
  p.scaled[1] <- 0
  iter=1
```

```
while(((
    (is.na(p.val[iter])!=TRUE & p.val[iter]<.05)*1==1  |
    (is.na(p.scaled[iter])!=TRUE & p.scaled[iter]<.05)*1==1  |
    which.min(AICs)==length(AICs) | which.min(BICs)==length(BICs) |
    which.min(AICCs)==length(AICCs)|
    which.min(ABICs)==length(ABICs))*1)==1 & iter<nfac) {

    if(iter+1<=4) {
      ctl <- c("TITLE: EFA; ")
      ctl <- rbind(ctl,paste("DATA: FILE IS ",name,".mplus.txt",";",sep=""))
      ctl <- rbind(ctl,paste("VARIABLE: NAMES ARE y1-y",ncol(sample.data),
                          "; USEV=y1-y",ncol(sample.data),
                          ";CATEGORICAL=y1-y",ncol(sample.data),";",
                          sep=""))
      ctl <- rbind(ctl,paste("ANALYSIS: TYPE = EFA",iter+1,iter+1,"; ESTIMATOR=MLR;
                          INTEGRATION=GAUSSHERMITE(5);LOGCRITERION =.01;
                          MITERATIONS =250 ",
                          ";",sep=" "))
                write(ctl,paste(name,"MLR",iter+1,".mplus",sep=""))
    } else

      if(iter+1>4) {
        ctl <- c("TITLE: EFA; ")
        ctl <- rbind(ctl,paste("DATA: FILE IS ",name,".mplus.txt",";",sep=""))
        ctl <- rbind(ctl,paste("VARIABLE: NAMES ARE y1-y",ncol(sample.data),
                            "; USEV=y1-y",ncol(sample.data),
                            ";CATEGORICAL=y1-y",ncol(sample.data),";",
                            sep=""))
        ctl <- rbind(ctl,paste("ANALYSIS: TYPE = EFA",iter+1,iter+1,"; ESTIMATOR=MLR;
                          INTEGRATION=GAUSSHERMITE(3);LOGCRITERION =.01;
                          MITERATIONS =250 ",
                          ";",sep=" "))
        write(ctl,paste(name,"MLR",iter+1,".mplus",sep=""))
      }

        system(paste('"C:/Program Files/Mplus/Mplus.exe"',
               paste(getwd(),"/",name,"MLR",iter+1,".mplus",sep=""),
               paste(getwd(),"/",name,"MLR",iter+1,".out",sep="")
               )
         )

      output <- scan(paste(name,"MLR",iter+1,".OUT",sep=""),what=c("raw"),fill=TRUE)
      conv <- length(which(output=="CONVERGENCE."))
      no.sol <- which(output=="COMPUTED")
     if(length(no.sol)!=0){problem=((output[no.sol-1]=="BE")&(output[no.sol-2]=="NOT"))*1
                           } else problem=0

      if(conv==0 & problem==0) {

        LLs[iter+1] <- as.numeric(output[which(output=="Loglikelihood")
                       [2:length(which(output=="Loglikelihood"))]+3])
        dof[iter+1] <- as.numeric(output[which(output=="Free")+2])
        sf[iter+1] <- as.numeric(output[which(output=="Scaling")+3])
        AICs[iter+1] <- as.numeric(output[which(output=="Akaike")+2])
        AICCs[iter+1] <- AICs[iter+1] + ((2*nparameters*(nparameters+1))/(sample.size-
                                                                    nparameters-1))
        BICs[iter+1] <- as.numeric(output[which(output=="Bayesian")+2])
        ABICs[iter+1] <- as.numeric(output[which(output=="Sample-Size")+3])
        cd <- (dof[iter]*sf[iter]-dof[iter+1]*sf[iter+1])/(dof[iter]-dof[iter+1])
        p.scaled[iter+1]=pchisq((-2*(LLs[iter]-LLs[iter+1]))/cd,(dof[iter+1]-
                           dof[iter]),lower.tail=FALSE)

        p.val[iter+1]= pchisq(-2*(LLs[iter]-LLs[iter+1]),dof[iter+1]-
```

```
                        dof[iter],lower.tail=FALSE)
     } else {

         LLs[iter+1] <- NA
         dof[iter+1] <- NA
         sf[iter+1] <- NA
         AICs[iter+1] <- NA
         AICcs[iter+1] <- NA
         BICs[iter+1] <- NA
         ABICs[iter+1] <- NA
         p.val[iter+1] <- NA
         p.scaled[iter+1] <- NA

       }
              iter=iter+1
  }

  out <- as.data.frame(matrix(nrow=length(LLs),ncol=1))
  out[,1] <- 1:length(LLs)
  out$LL <- LLs
  out$df <- dof
  out$sf <- sf
  out$p.val <- p.val
  out$p.scaled <- p.scaled
  out$AIC <- AICs
  out$AICc <- AICcs
  out$BIC  <- BICs
  out$Adj.BIC  <- ABICs

  write.fwf(round(out,3),paste(name,"MLR_mplus.summary.txt",sep=""),
            width=rep(15,10),rownames=FALSE,colnames=TRUE)

  if(length(which(p.val>.05))!=0) { chifac <- which(p.val>.05)[1]-1
                                    } else chifac <- NA

  if(length(which(p.scaled>.05))!=0) { chifac2 <- which(p.scaled>.05)[1]-1
                                    } else chifac2 <- NA

  if(which.min(AICs)!=length(na.omit(AICs))) {AICfac=which.min(AICs)
            } else AICfac=NA

  if(which.min(BICs)!=length(na.omit(BICs))) {BICfac=which.min(BICs)
            } else BICfac=NA

  if(which.min(ABICs)!=length(na.omit(ABICs))) {ABICfac=which.min(ABICs)
            } else ABICfac=NA

  return(list(Chifac = chifac,
              Chifac2 = chifac2,
              AICfac=AICfac,
              BICfac=BICfac,
              ABICfac=ABICfac
              )
         )
}
```