Assessing Dimensionality in Complex Data Structures:

A Performance Comparison of DETECT and NOHARM Procedures

by

Dubravka Svetina

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2011 by the
Graduate Supervisory Committee:

Roy Levy, Co-Chair
Joanna Gorin, Co-Chair
Roger Millsap

ARIZONA STATE UNIVERSITY

May 2011

ABSTRACT

The purpose of this study was to investigate the effect of complex structure on dimensionality assessment in compensatory and noncompensatory multidimensional item response models (MIRT) of assessment data using dimensionality assessment procedures based on conditional covariances (i.e., DETECT) and a factor analytical approach (i.e., NOHARM).

The DETECT-based methods typically outperformed the NOHARM-based methods in both two- (2D) and three-dimensional (3D) compensatory MIRT conditions. The DETECT-based methods yielded high proportion correct, especially when correlations were .60 or smaller, data exhibited 30% or less complexity, and larger sample size. As the complexity increased and the sample size decreased, the performance typically diminished. As the complexity increased, it also became more difficult to label the resulting sets of items from DETECT in terms of the dimensions. DETECT was consistent in classification of simple items, but less consistent in classification of complex items. Out of the three NOHARM-based methods, $\chi^2_{G/D}$ and *ALR* generally outperformed RMSR. $\chi^2_{G/D}$ was more accurate when N = 500 and complexity levels were 30% or lower. As the number of items increased, *ALR* performance improved at correlation of .60 and 30% or less complexity.

When the data followed a noncompensatory MIRT model, the NOHARM-based methods, specifically $\chi^2_{G/D}$ and *ALR*, were the most accurate of all five methods. The marginal proportions for labeling sets of items as dimension-like

were typically low, suggesting that the methods generally failed to label two (three) sets of items as dimension-like in 2D (3D) noncompensatory situations.

The DETECT-based methods were more consistent in classifying simple items across complexity levels, sample sizes, and correlations. However, as complexity and correlation levels increased the classification rates for all methods decreased. In most conditions, the DETECT-based methods classified complex items equally or more consistent than the NOHARM-based methods. In particular, as complexity, the number of items, and the true dimensionality increased, the DETECT-based methods were notably more consistent than any NOHARM-based method. Despite DETECT's consistency, when data follow a noncompensatory MIRT model, the NOHARM-based method should be preferred over the DETECT-based methods to assess dimensionality due to poor performance of DETECT in identifying the true dimensionality.

DEDICATION

This dissertation is dedicated to Arturo and my family.

ACKNOWLEDGMENTS

Although only my name appears on this dissertation, its completion would not be possible without support of many people. Foremost, I am grateful to my mentor and advisor Dr. Roy Levy, who introduced me to the field of multidimensionality in item response theory and dimensionality assessment; area which captured my interest and eventually led me to my dissertation topic. Dr. Levy's guidance, challenge, and support throughout all stages of my dissertation were instrumental in completion of the project. Intellectual conversations, support, and professional training by both of my co-advisors, Dr. Levy and Dr. Joanna S. Gorin contributed greatly to how I think about issues related to measurement. Their guidance and mentorship allowed me to explore and do research in a number of different areas, which allowed me to expand on how I think as a researcher. I would also like to thank Dr. Roger Millsap, whose comments and suggestions helped me think through challenges in operalizations of performance variables, and whose insightful questions make me think hard about a problem at hand.

In addition to my immediate committee, I wish to thank Drs. Green and Thompson, as well as students in the MSMS program who supported me in numerous ways throughout the years of my graduate training. Although there are many students I extend my thanks to, I would like to specifically thank the following students. I am thankful to Lee Scott for always staying positive and encouraging me to move forward. Derek Fay, who assisted me with the initial

DETECT programming and for sharing the passion of dimensionality assessment, I say thank you. AAron V. Crawford and Katie Poole for being open to listen to me go on about my dissertation and for being supportive of my endeavors.

Lastly, I wish to thank my family. Without them, this dissertation would not become a reality. First, I wish to give my heartfelt gratitude and thanks to my fiancé Arturo Valdivia, whose unconditional support has always been present - during long days and late nights of writing this dissertation. I also thank him for having intellectual discussions about my project and related programming. I wish to thank my brother, Marko Svetina, for intellectual conversations and patient listening about technical issues I encountered during this journey. I want to thank my sister-in-law Olena Tsurska for her help in editing of portions of this script. Lastly, I wish to express the deepest gratitude to my parents who showed me the value of education and whose complete support has been invaluable.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Chapter 1

INTRODUCTION

Dimensionality Defined

Dimensionality in assessment concerns the number of abilities or constructs assessed by a test or a set of items. Dimensionality can be viewed in many different ways, such as through the lens of aspects of assessment design in terms of the dimensions intended to be assessed (e.g., Mislevy, Almond, & Lukas, 2003) or the analysis of observed responses to test items. The current work focuses on analyses of the latter type. Within this area, some researchers define dimensionality as the number of traits that underlie a set of test item responses and which account statistically for variances and covariances among the items (e.g., Hattie, Krakowski, Rogers, & Swaminathan, 1996; Stout, Froelich, & Gao, 2001; Stout, 1990; Zhang 2007). Others further characterize dimensionality as being influenced by the interaction between the test items and the examinees, or understand dimensionality in the context tied to the purpose of a test (e.g., Gierl, Leighton, & Tan, 2006; Reckase, 2009). Few scholars extend these definitions to emphasize the patterns of dependency of the items to their respective dimensions, suggesting that the number of underling dimensions or factors may not be sufficient in understanding dimensionality of data (e.g., Levy & Svetina, in press).

Even though test dimensionality is defined and understood in several different contexts, there seems to be an agreement among the contemporary

researchers that investigation of the dimensional structure of a test is a "requisite part of a comprehensive validation process" (Jang & Roussos, 2007, p. 2).

Dimensional structure can be defined as the relationship between the items on the test and the latent proficiencies believed to be measured by the test. In other words, the internal structure of the test indicates which items are associated with what dimensions, where a dimension is defined as a latent proficiency that accounts for performance on the items and therefore the associations among them.

Often, a dimension is substantive in nature. For example, on a science test, several proficiencies might be measured, including proficiency in life, physical, and earth sciences. If the test ought to measure examinee proficiency in these aspects of science, we might seek evidence to support a three-dimensional structure of items responses via dimensionality assessment.

The Importance of Assessing Dimensionality of the Data

Over the last few decades, researchers have provided arguments for supporting dimensionality assessment and understanding the structure of a test, as an important step in testing (Hambleton, Swaminathan, & Rogers, 1991; Jang & Roussos, 2007; Tate, 2003; Zhang, 2007). The process of developing, evaluating, and maintaining of (large-scale) testing program requires dimensionality assessment as it contributes to providing empirical support for the content and cognitive process aspects of test validity (e.g., AERA, APA, & NCME, 1999; Hattie, 1985; Tate, 2003). By examining the dimensionality, researchers are able to link the substantive interpretation with the statistical outcomes for the purpose

of better understanding examinee-by-item interactions (Gierl, et al., 2006). In broad terms, dimensionality assessment contributes to providing evidence for various aspects of the validity argument.

Assessing the internal structure of the item responses on a test is crucial because it forms the basis of statistical analysis of the data (Hambleton, et al., 1991; Zhang, 2007). Through psychometric modeling of the data, researchers gather evidence for making inferences about students. In order to make such inferences, psychometric models used in the analysis ought to be technically sound and aligned with the data from the tests. For example, in traditional psychometric models of item response theory (IRT; e.g., 1-, 2-, or 3-parameter logistic models), the assumptions that a test measures a single ability and that the item responses "obey the principle of local independence" are explicitly made (Jang & Roussos, 2007, p. 2). Within a classical test theory framework, the same can be expressed through the existence of “homogeneous” items on the test (McDonald, 1999).

In educational tests, it is often the case that multiple proficiencies are present, which leads to multidimensionality of the data. Therefore, understanding the structure of the data is paramount if we are to make appropriate inferences about the scores based on a test. In other words, if a researcher is to draw meaningful inferences about examinee’s standing on the construct(s) of interest, it is essential to assess the (uni)dimensionality of data (Stout, 1987; Stout, et al., 1996). Stone and Yeh (2006) summarized it well in saying that the investigation

of the internal structure of a test allows one “to identify what domains are being measured, identify the relationship between those domains, provide support for the hypothesized multidimensionality and test score interpretations, and identify construct-irrelevant variance” (p. 194). Examination of the relationships between the constructs allows us to find support for the alignment with the intended constructs and to control for the unintended constructs (e.g., by using multidimensional IRT; MIRT). Both of these are essential if we are to maintain consistent measurement and score interpretations across tests.

Negligence in dimensionality assessment or misalignment of the psychometric model and the data may lead to severe consequences in various aspects of testing. These consequences include inaccurate and imprecise item and person parameter estimates, issues in test linking and equating of the tests, item bias and test assembly, and score reporting (e.g., Ackerman, 1989, 1994; Chen & Thissen, 1997; Reckase, Carlson, Ackerman, & Spray, 1986; Walker & Beretvas, 2003; Way, Ansley, & Forsyth, 1988; Yen, 1985).

For instance, Reckase and his colleagues (1986) demonstrated that when multidimensionality and difficulty were confounded, a unidimensional scaling produced different meanings at various points on the scale. Way, Ansley, and Forsyth (1988) examined the effects of using the unidimensional model to estimate two-dimensional data. They found that for data generated by compensatory MIRT model the estimated discrimination parameters were best considered as a complex combination of the discrimination parameters along the

two dimensions, while item difficulty parameter estimates and the ability estimates were close to the average of the their respective values on the two dimensions. Similar findings were obtained in Ackerman (1989), where the author found even stronger relationship between the ability estimate under the unidimensional model and the complex combination of the two abilities (or discrimination) that are approximated by the dimension of best measurement.

In addition to the inaccurate estimates as a result of the inappropriate application of the psychometric model to the data, there also exists a potential concern regarding the score comparability. In situations where equating is important (such as for the purpose of providing a developmental scale across grades), the tests' structures ought to be equitable in order to maintain comparable scores. Changes in test structures from grade to grade could threaten validity such that scale changes artificially increase or decrease the within grade variability (e.g., Yen, 1985). In other words, the invariance structure of the test needs to be preserved (Yeh, 2007, p. 2), and utilizing tools for dimensionality assessment may prove helpful in assuring that such needs are met.

Dimensionality assessment may also provide support for meaningful and appropriate score reporting. According to the legislation of No Child Left Behind (NCLB, 2001), states must report both scale and subscale scores (Goodman & Hambleton, 2004). Through understanding the dimensionality, evidence may be gathered for appropriate score reporting. For example, in a mathematics test, five content areas might be evaluated, including number properties and operations,

measurement, geometry, data analysis, and algebra. If information in the data are consistent with the hypothesis of five distinguishable proficiencies corresponding to the five content areas, subscale score reporting, in addition to the overall mathematics score, may indeed be appropriate. However, if the dimensionality assessment supports an alternative interpretation of the number of dimensions underlying the data, such subscale reporting might not be appropriate.

An added motivation leading to dimensionality assessment, related to the issue of fairness, is raised through the potential presence of bias in the items. This can be understood as the result of a multidimensional test structure that could be related to construct-irrelevant factors (Tate, 2002). Examining the dimensionality of the test and understanding why some items are biased may help avoid such bias in the future constructions of the items.

In summary, by assessing dimensionality of the item responses on a test one can examine and deal with potential threats to various aspects of validity, including substantive and structural, as well as other issues related to testing. By examining the (multi)dimensionality of the data, construct-irrelevant proficiencies potentially measured by some of the items on the test can be found, items with differential item functioning can be examined, and potentially improper equating of the new test forms can be avoided. The above scenarios point to some of the main concerns and potential motivations for assessing the dimensionality of a test (e.g., Tate, 2002, 2003).

It is thus argued that given the role of dimensionality assessment in supporting a variety of psychometric endeavors, assessing dimensionality should be a *prerequisite* to applying most commonly used IRT models (Childs & Oppler, 2000; Jang & Roussos, 2007; Nandakumar & Yu, 1996; Nandakumar, Yu, Li, & Stout, 1998; Seraphine, 2000).

A fair number of techniques have been developed across various modeling paradigms to assess dimensionality of the structure of responses (Levy & Svetina, 2010; Tate, 2003). The techniques may be grouped based on a variety of elements, including approaches to analysis (exploratory, confirmatory), the modeling paradigm within which they are commonly applied (e.g., factor analytic, item response, etc.), and distributional assumptions (parametric, nonparametric). The variety of methods commonly used today offer researchers the flexibility to make appropriate choices about how to determine the number of dimensions present in the data.

Previous research has shown that to a large degree, commonly used methods today perform well under certain conditions (Finch & Habing, 2005; Froelich & Habing, 2008; Gierl, et al., 2006; Hattie, et al., 1996; Nandakumar, 1991, 1993; Nandakumar & Stout, 1993; Nandakumar & Yu, 1996; Nandakumar, et al., 1998; Stout, 1987; Stout et al., 1996; van Abswoude, van der Art, & Sijtsma, 2004; Zhang, 2007; Zhang & Stout, 1999b). These conditions are typically those that align well with the principles upon which the tools were built. However, relatively little research has been conducted on the extent to which

these methods are robust to departures from their assumptions. The current study focuses on how well some of the more commonly used methods work under conditions that do not align with the foundational principles of the tools.

For example, DETECT (Dimensionality Evaluation To Enumerate Contributing Traits; Kim, 1994; Stout et al., 1996; Zhang, 2007; Zhang & Stout, 1999b) is a procedure that seeks dimensionally distinct clusters of items based on the conditional covariances among the item pairs. Dimensionality distinct clusters are sought such that approximate *simple* structure is preserved under a generalized compensatory MIRT model (Zhang & Stout, 1999b). A common 3-parameter compensatory normal-ogive MIRT model expresses the probability of a correct response of person *i* to item *j* as:

$$P\left(X_{ij}=1\middle|\boldsymbol{\theta}_{i,}\boldsymbol{a}_j,d_j,c_j\right)=c_j+\ (1-c_j)\Phi\left(\boldsymbol{a}_j'\boldsymbol{\theta}_i+d_j\right), \qquad 1.1$$

where, Φ is a cumulative normal distribution function, $\boldsymbol{\theta_i}=(\theta_1,\theta_2,\ldots,\theta_m)'$ is a vector of *M* latent variables for examinee *i*, $\boldsymbol{a}_j=(a_{j1},a_{j2},\ldots,a_{jM})'$ is a vector of *M* parameters related to discriminating power of the item *j*, c_j is a lower asympotote parameter for item *j*, and d_j is the intercept related to the marginal difficulty for item *j* (e.g., McDonald, 1997). Following McDonald (1999), an item is referred to as *factorially simple* if it has only one nonzero coefficient in its $\boldsymbol{a}_j$ vector. Conversely, and item is *factorially complex* if it has more than one nonzero coefficients in its $\boldsymbol{a}_j$ vector.

A model for a set of items exhibits *simple structure* if, according to the model, all of the items are factorially simple. In other words, in *simple* structure,

each item is associated with only one latent variable. Moving away from simple structure, *approximate simple structure* refers to situations in which any one item is primarily associated with only one dimension, although trivial but nonzero coefficients in the item's $\boldsymbol{a}_j$ vector allow items to be associated with multiple latent examinee variables. *Complex* structure further extends any one item's association with multiple latent examinee variables; however, those associations are now nontrivial.

These concepts are discussed in more detail in Chapter 2. For the present purposes, it is sufficient to note that DETECT is grounded in principles of simple structure. Therefore, the performance of DETECT in situations where complex structure exists might suffer. More generally, there seems to be a lack of research and support for most, if not all, of the commonly used methods for dimensionality assessment in realistic situations where the principles of the methods and conditions of the data are not aligned. It will be argued that while methods for dimensionality assessment have shown great promise, further research, particularly with respect to complex data, is needed.

Purpose of the Study

Popular methods for dimensionality assessment assume that items simple or approximately simple. Furthermore, these methods are typically applied in the context where a compensatory multidimensional model is assumed. This study seeks to go beyond the present practices.

The purpose of this study is to investigate the effect of complex structure on dimensionality assessment data that follow both compensatory and noncompensatory MIRT models using dimensionality assessment procedures based on conditional covariances (i.e., DETECT) and factor analytical approaches (i.e., NOHARM). The procedures of DETECT and NOHARM, discussed in greater detail in Chapter 2, are chosen because these methods embody the two most common and popular approaches to dimensionality assessment (i.e., conditional covariance and factor analytical). Additionally, both of these methods allow for exploratory nature of dimensionality assessment, have been shown to perform rather well under a variety of conditions, and are to some extent flexible in their application.

The following research questions are addressed in this study:

a) How well do methods based on DETECT and NOHARM perform in estimating the dimensional structure of the data that exhibit *complex structure*? This includes their performance in estimating the number of dimensions that underlie the data, and the interpretability of the resulting groupings of items.

b) Do the underlying MIRT models (compensatory and noncompensatory), correlations among latent variables, sample size, and/or the number of items influence the performance of these dimensionality assessment methods?

In order to investigate the effects of the complex data on the performance of these two procedures, this study will be carried out via a simulation study using a Monte Carlo approach. By using Monte Carlo, the "true" dimensionality structure is known and thus can be compared to the estimated dimensional structure.

The motivation for this study stems in part from the fact that the literature on issues related to dimensionality assessment typically focuses on examining the procedures to assess (i.e., detect departures from) unidimensionality (Hattie et al., 1996; Nandakumar, 1993, 1994; Nandakumar & Stout, 1993; Nandakumar & Yu, 1996; Nandakumar, et al., 1998; Roussos, Stout, & Marden, 1993; Stout, 1987; Stout et al., 2001). The evaluation of dimensionality is no less important when it comes to multidimensional models (Levy & Svetina, in press). This is particularly important, given a recent rise of development and applications in MIRT models such as modeling of multidimensional data, applications in adaptive testing, or equating (e.g., Ackerman, 1996, Bolt & Lall, 2003; De Champlain, 1996; Embretson 1997; McDonald, 1997; Walker & Beretvas, 2003; Yao & Boughton, 2007, etc.). These studies recognize and point to the need for supporting data-model fit procedures, including dimensionality analysis.

The literature on multidimensional item response data primarily models situations where (*approximate*) *simple structure* exists (e.g., Finch & Habing, 2007; Gierl, et al., 2006, etc.). Rarely are exploratory methods assessed in the context of complex data, which is partially due to the fact that several of the

commonly used methods, including DETECT, are based on the principles of simple structure. The only study found to date addressing performance of DETECT in the context of multidimensional data with complex structure is Gierl et al.'s (2006) study. The results of the Gierl's study, reviewed in more detail later, provided important evidence of DETECT's performance. Several important issues were left unexamined, which motivated this study.

The current study attempts to examine issues related to dimensionality assessment when a researcher has no *a priori* hypothesis of the structure of the data, when in fact the data exhibit complex structure. In particular, this study focuses on examining the performance of the procedures when multidimensionality is present and where several items on a test are related to a multiple rather than just a single dimension; that is when some items on a test are factorially complex. In addition to the methods based on the popular, conditional covariance based DETECT procedure, the performance of methods based on the output from a factor-analytical procedure, NOHARM, is examined for comparison purposes.

Chapter 2

BACKGROUND LITERATURE

Definition of Dimension

Though there seems to be an agreement of the importance of assessing dimensionality, the definition of dimensionality of data may vary depending on the adopted modeling paradigm. Traditionally, in defining dimensionality of the data, a researcher tries to address the question of *how many latent variables (factors) are thought to underlie data on a set of test items*. Often, the analyst is interested in understanding and (statistically) explaining the variances and covariances among the items on a test. In assessment settings, we might ask how complex is the latent space needed to adequately represent students' performance on a particular test.

Some of the more recent definitions and references to dimensionality include Camilli, Wang, and Fesq (1995), who defined test dimensionality as "the number of latent variables that account for the correlations among item responses in a particular data set" (p. 80). McDonald (1981) echoed Lord and Novick (1968), when suggesting that the proper quantification of dimensionality in the data ought to be based on the *strong local independence* principle. That is, the dimensionality of the data is that which is needed to achieve strong local independence. In this line of reasoning, Hattie, et al. (1996) suggested that, when the dimensionality is correct, then "[o]nce trait values are fixed at a given value (i.e., conditioned on), the responses to items become statistically independent.

Thus, in order to determine the dimensionality of a set of items it is necessary and sufficient to identify the minimal set of traits such that at all fixed levels of these traits the item responses are independent" (p. 1).

Relaxing the assumption of the strong local independence, a number of researchers (e.g., Junker, 1993; Stout, 1990; Stout et al., 2001; Zhang, 2007), operationalized the definition of the dimensionality of data by describing it in terms of a minimum number of (dominant) dimensions necessary to achieve (pair-wise) *local independence* and *monotonicity* (discussed in further detail in the next section).

Others suggested that the issue of dimensionality involves more than (successfully) arriving to a number proficiencies or dimensions that account for the item responses (Levy & Svetina, in press; McDonald, 2000). These authors point that in addition to arriving to the number of dimensions that underlie the item responses, the relationship between the items and dimensions play a crucial role in dimensionality assessment. One could be successful in identifying the number of dimensions that underlie the data, however, if the relationships between the items and dimensions are incorrectly identified, problems in the appropriate estimation and score reporting may occur. Thus, it is important to not only arrive to the correct number of dimensions but to also appropriately account for the patterns of the relationships as well.

A related but slightly different understanding of dimensionality has emerged from the recent growth of cognitive diagnostic models characterized by

their use of discrete latent variables (Rupp & Templin, 2008). In the binary skills model (Haertel, 1989), latent classes are identified with a distinct pattern of dichotomous skills. Rather than thinking about a single (or multiple) continuous dimension(s), one might think about dimensionality in terms of how skill combinations define classes of students and their proficiency within a specified skill space. The multidimensional nature of the models, as suggested by Rupp and Templin (2008), can be described as "the number of latent variables depends on the number of skills that researchers hope to numerically separate in a reliable manner with the assessment" (p. 228).

Unlike in the typical factor analytical or IRT analyses, where multiple dimensions operationalize different constructs (or different aspects of the same construct), in applications of such latent class models Rupp and Templin (2008) suggest that dimensionality be broken down even further to elementary components and their interaction (p. 228). DiBello, Roussos, and Stout (2007) add that it is the purpose of the assessment that will have "significant impact on whether the targeted latent attribute of skill space will be modeled with one or more than one variable…" (p. 981).

Substantively, "a decision about dimensionality…inevitable rests partly on a substantive basis, and should constitute a conclusion about the detailed structure of the relationships – not merely the number of dimensions" (McDonald, 2000, p. 103). In other words, dimensionality assessment should be a process of both statistical and substantive investigations of the relationships between the items

and/or latent, unobservable, traits, and the pattern of relationships between the items and dimensions.

Though substantive considerations are important, this work focuses on the notions of dimensionality that resemble those of Stout (1990) and others. In particular, statistical investigations meant to account for associations among the items are meant to partially provide support for determining a number of dimensions in a set of items. As seen from a few examples above, the term "dimensionality" has been defined and used in multiple ways. Although often referred to dimensionality of the test, one should really discuss dimensionality of the observed item responses that represent the interaction between examinees and items.

The remainder of this chapter is divided in the following sections. First, concepts related to dimensionality assessment are discussed, including the concepts related to local independence. Next, to motivate a discussion about current dimensionality approaches, parametric and nonparametric based approaches to dimensionality assessment are presented. Each of these approaches is followed by a discussion of commonly used procedures and software for dimensionality assessment. The chapter concludes with current research on dimensionality assessment, with a primary focus on the research evaluating the two methods used in this study; NOHARM and DETECT.

Concepts Related to Dimensionality

Stout (1990) defined the dimensionality of a test as the minimal dimensionality required for a possibly vector-valued latent variable,$\boldsymbol{\theta}$, to produce a model that is both locally independent and monotone. The (increasing) monotonicity is achieved when the probability of a correct response increases as the ability increases. Local independence (LI), also known as strong local independence (SLI) states that the joint probability of the responses to the set of items comprising the test is equal to the product over items of conditional probabilities for all the item responses on a test given $\boldsymbol{\theta}$ (Hattie, et al., 1996; Stout, 1990). This is can be formulated as:

$$P(\mathrm{X}_1, \mathrm{X}_{2,\ldots,}\mathrm{X}_J|\boldsymbol{\theta}) = \prod_{j=1}^{J} P(\mathrm{X}_j|\boldsymbol{\theta}), \qquad 2.1$$

where $\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_J$ are scores for items 1, 2,… up to *J*, typically scored as 0 for an incorrect and 1 for a correct response in dichotomously scored items, and *J* is the total number of items on a test. Equation 2.1 states that a joint probability for all item responses on a test given $\boldsymbol{\theta}$ is a product of each conditional probability separately. In other words, if we condition on $\boldsymbol{\theta}$, the response to any item is independent of the response to any other item.

In practice SLI is difficult to investigate. Thus, weak local independence (WLI), which deals with item pairs rather than joint distribution of all items, is typically used in investigating local independence. WLI is the condition that for all unique item pairs and for all $\boldsymbol{\theta}$**,** the covariance between the item pairs, conditional on $\boldsymbol{\theta}$ is zero:

$$cov(\mathrm{X}_j, \mathrm{X}_{j'}|\boldsymbol{\theta}) = 0, \qquad 2.2$$

where X_j is a scored response to item j and $\mathrm{X}_{j'}$ is a scored response to item $j^{'}$; $j \neq j^{'}$, and *cov* stands for covariance between the items in question. From Equation 2.2, we can see that WLI implies that each item pair has zero covariance once the latent trait(s) has been accounted for.

McDonald (1994) and others argue that in cases of real data for which WLI holds, SLI holds approximately (Stout, et al., 1996). Note that higher-order dependencies are allowed among the items, although if WLI holds, it is unlikely that SLI would not (e.g., Zhang, 2007). Thus, if one accepts that in cases where WLI holds, SLI will also hold approximately (and monotonicity is assumed), then evaluating WLI is sufficient for evaluating SLI and dimensionality.

Here, a cautionary note needs to be made; though LI and dimensionality assumption are related, the two are not identical. For example, if the data follow a model with a particular dimensional structure, and we employ that model, LI will hold. If the data follow a multidimensional structure, and we employ a unidimensional model, LI will not hold. However, if the data follow a unidimensional structure, and we employ a multidimensional model, LI will also hold. Nevertheless, evidence that LI does not hold is *prima facie* evidence that the dimensional structure, and possibly the number of dimensions, is incorrectly specified.

Even in cases where tests are designed to measure a single construct (i.e., to be unidimensional), “minor” or “nuisance” proficiencies are likely to account

for some inter-item dependencies, in addition to a single dominant construct (Goldstein, 1980). These minor proficiencies or dimensions may be functions of the testing environment, characteristics of instruments, or instructional effects (Seraphine, 2000). Further, even when we do have the correct number of dimensions, we still might not have LI if the pattern dependencies of the items on the dimensions are incorrectly specified (Levy & Svetina, in press). In order to take into account minor latent nuisance trait(s), Stout (1987, 1990) broadened conceptualization of dimensionality based on *essential independence*.

The responses to items are essentially independent if the average of all inter-item covariances conditioned on correctly specified (multiple) dimensions approaches zero as the number of items approaches infinity (Nandakumar & Stout, 1993; Stout, 1987; Stout et al., 1996),

$$\frac{\sum_{1\leq j<j'\leq J}\left|cov\left(\mathrm{X}_j,\mathrm{X}_{j'}|\boldsymbol{\theta}\right)\right|}{\binom{J}{2}}\rightarrow 0\ as\ J\rightarrow\infty. \qquad 2.3$$

Concepts related to dimensionality can be illustrated graphically. Figure 1 illustrates three data structures that are relevant for discussion of dimensionality. In panel (a) of Figure 1, an *exact simple structure* is shown. All of the items on a test are associated with one dimension only. Some of the items are influenced by θ_1, while others by θ_2. In panel (b), an *approximate simple structure* is shown. We can see that there is a potential influence of θ_2 on some items primarily influenced by θ_1 and vice versa. Dashed lines indicate that such influence is weak in strength and magnitude. A *complex structure*, as presented in panel (c), suggests that some

items are influenced by both θ_1 and θ_2, while others by a single dimension only.

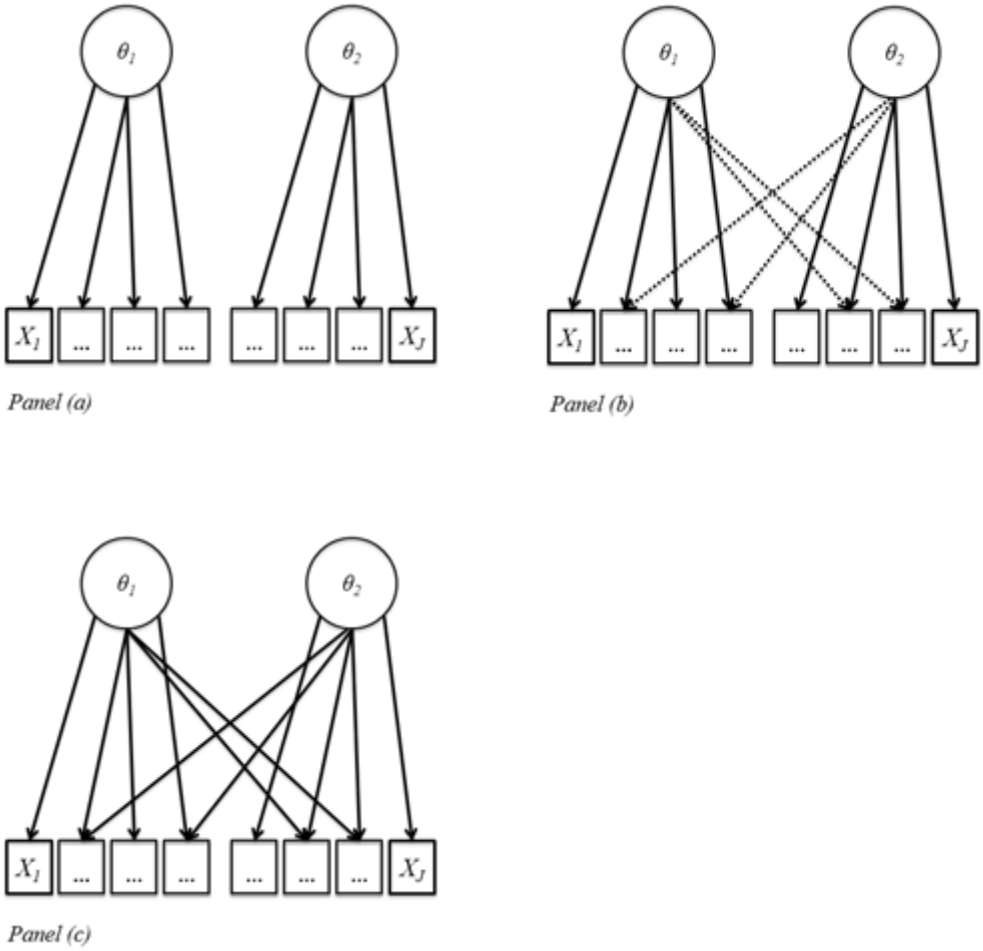

Figure 1. Geometric representation of exact simple (a), approximate simple (b), and complex (c) structure.

An alternative way to represent a two-dimensional latent space and items is given in Figure 2. Such a graphical representation is useful in visualizing structural features, where the coordinates in multidimensional space represent the latent abilities measured by the test (e.g., Ackerman, 1996; Stout et al., 1996).

Analogous to panel (b) of Figure 1, Figure 2 illustrates a two-dimensional test (represented by θ_1 and θ_2), with an *approximate simple structure*. Note that in Figure 2, the two axes (θ_1 and θ_2) are shown to be orthogonal to each other. While this does not have to be the case; for simplicity purposes, the two dimensions pictured here are uncorrelated. The lines coming out from the origin represent *item vectors* – a single line represents an item on this two-dimensional test. The direction/location and the length of the item communicate its characteristics. The direction (angle) of the *item vector* is the direction in multidimensional space that the item provides maximal discrimination, and reflects the relative amount of information that the item provides about the dimensions (i.e., in terms of whether the item vector is closer to θ_1 or θ_2). The length of that *item vector* illustrates multidimensional discrimination of that item (i.e., longer lines indicate higher discrimination values).

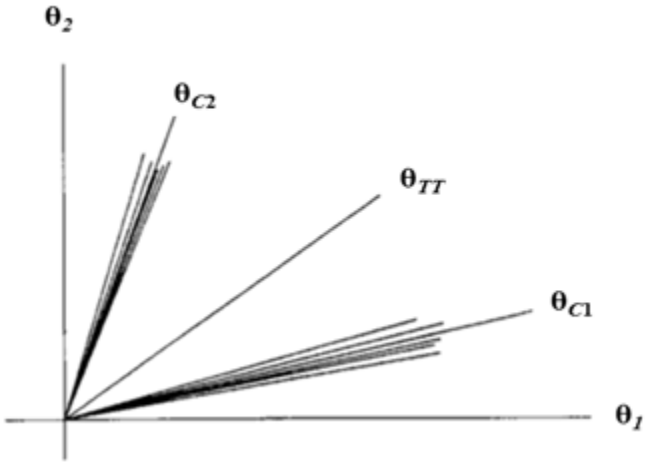

Figure 2. Two-dimensional test displaying approximate simple structure.
Note: Stout at el., 1996.

In Figure 2, the location and direction of the *item vectors* suggest there are two groups (or clusters) of items: one cluster of items that mostly relate to θ_1, and another cluster of items mostly related to θ_2. We can also see that the strength of the relationship between the items and their respective dimensions is relatively comparable for all items (i.e., the lengths of the lines are somewhat similar).

Relating back to the concepts of conditional covariance previously discussed, we describe this set of items using Θ_{C1}, Θ_{C2}, and Θ_{TT}. Here, Θ_{C1} and Θ_{C2} represent the cluster's unidimensional latent variables best measured for that cluster scores, and Θ_{TT} is a unidimensional latent dimension of best measurement for the total test score. Θ_{TT} can be thought of as a dimension in a multidimensional space consisting of $\boldsymbol{\theta}$ along which a set of items maximally discriminates (i.e., rough average of all *item vectors*, Stout et al. 1996). Importantly, Θ_{TT} is analogous to the direction of the latent variable in this multidimensional space that would be obtained by fitting a unidimensional model (Stout, et al., 1996).

Similarly, on a subtest level, there is a unidimensional latent variable best measured for any one subtest (in this case Θ_{C1} and Θ_{C2}). Although not illustrated here, the representation of the *simple* and *complex structures* using example in Figure 2, would be as following. For *simple structure*, all *item vectors* would fall on either θ_1 or θ_2 axis. For *complex structure*, at least some *item vectors* would be closely located around the Θ_{TT} (between 35° and 55° from θ_1, Gierl et al. 2006).

As presented in Figure 2, two clusters are formed, and all *item vectors* lie closely to one of the two axes (i.e., two dominant dimensions exist). Zhang and Stout (1999b) illustrated that within each cluster, items appear relatively homogeneous (i.e., more similar), and their conditional covariance given Θ_{TT} will be positive. For item pairs whose vectors come from different clusters, the conditional covariances given Θ_{TT} will be negative.

Zhang and Stout (1999a) also showed that the angles and lengths of the item vectors project the magnitude of the item's association with dimension, with respect to the direction of best measurement. They showed that as an item "moves" away from the Θ_{TT}, the covariance with items in the cluster (which remain fixed) increases, given Θ_{TT}. For example, consider a different two-dimensional case (Figure 3), where four *items vectors* of equal length are represented by $\boldsymbol{U_1}$ through $\boldsymbol{U_4}$, and where angles and discrimination vectors (lengths) are fixed.

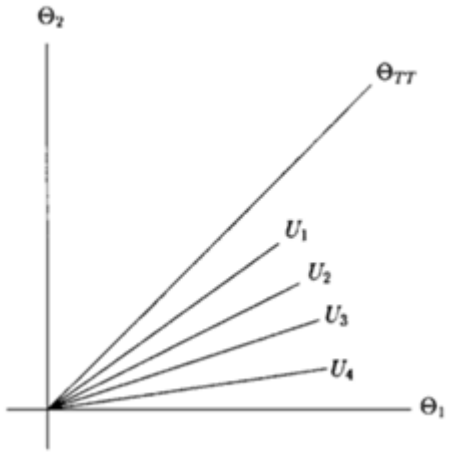

*Figure 3.*Direction of best measurement for four items in two-dimensional space. *Note:* Stout at el., 1996.

Zhang and Stout (1999a) illustrated that $cov(U_1,U_2|\Theta_{TT}) < cov(U_2,U_3|\Theta_{TT}) < cov(U_3,U_4|\Theta_{TT})$. In addition, the authors illustrated that as the angle between the items decreases and as either of the items increase its angle with Θ_{TT}, their conditional covariance, given Θ_{TT}, increases. Similarly, the conditional covariance between the items increases as the lengths of the item vectors increase (i.e., increase in magnitude of item discrimination vectors). These concepts are important in that they provide building blocks for many of the current dimensionality assessment procedures described next.

Dimensionality Assessment Approaches and Methods

There are many ways to organize dimensionality assessment methods. One such way is to think about methods and tools used to examine dimensionality through the lenses of parametric and nonparametric approaches. Another grouping may be based on the methodological nature (e.g., exploratory or confirmatory) or modeling paradigm (e.g., factor analytic or IRT). For the following discussion, a grouping based on parametric and nonparametric approaches will be adopted. Within each of these approaches, various methods have been developed to assess dimensionality. Some methods have factor analytical (FA) roots, while others grew out of IRT traditions. Similarly, some methods are purely exploratory or confirmatory, while others can handle both. Current, commonly used, procedures associated with both parametric and nonparametric approaches are discussed next.

Parametric approach to dimensionality assessment. Within the parametric approach to dimensionality assessment, one of the two frameworks is

typically adopted: FA based methods and MIRT based methods. In this section, the FA and MIRT frameworks are presented fist. Next, the relationship between the two approaches is noted. Lastly, current methods based on parametric approaches to assessing dimensionality are discussed.

Factor analytic (FA) framework. Traditionally, a common approach to testing dimensionality has been through factor analysis methods. In classical linear factor analysis, a researcher seeks to identify a set of factors (dimensions) that can account for the observed pattern of correlations among the scores (Kane, 2006). The relationship between the factor(s) and observed measures is expressed through factor loadings. In the common FA model, each variable is a linear combination of one or more common factors and one unique factor. A unique factor is unobserved and is composed of two parts: the latent factor component that represents unexplained variance and the measurement error due to unreliability of the measured variable. The common factor model for linear factor analysis can be mathematically presented as:

$$X_{ij} = \lambda_{j1}\theta_{i1} + \lambda_{j2}\theta_{i2} + \cdots + \lambda_{jm}\theta_{im} + e_{ij}, \quad 2.4$$

where λ_{jm} is the loading (weight) for item *j* on factor (dimension) *m*, θ_{im} is the factor score for examinee *i* on factor *m*, and e_{ij} is a term that carries a residual (or unique) information for examinee *i* on item *j*.

In a common factor model, variables are assumed to be continuous. In educational data, variables are often scored dichotomously, causing the assumed linear relationship between the items and factors to become nonlinear. This

nonlinear relationship led to occurrences of spurious “difficulty” level factors based on the observed-item correlations (Green, 1983; McDonald & Ahlawat, 1974).

Due to this nonlinear nature of item responses in educational data, researchers developed tools to accommodate the nonlinear relationship between the items and the factors by using tetrachoric (rather than Pearson or phi) correlations in the analysis (Jang & Roussos, 2007). Unfortunately, using tetrachoric correlation matrices could be problematic since they are often not positive definite (Cook, Dorans, & Eignor, 1988; Knol & Berger, 1991; Lord & Novick, 1968). The issue of not positive definite matrices presents the problems in estimation, as typically maximum likelihood (ML) or generalized least square (GLS) estimation procedures are used. Further, tetrachoric matrices may be inappropriate when the distribution of the latent ability is nonnormal (Jang & Roussos, 2007; van Abswoude, et al., 2004) and when a potential for guessing (e.g., in multiple-choice items) is present (Jang & Roussos, 2007; Hattie, et al., 1996; Mislevy, 1986). Therefore, the appropriateness of using linear methods in cases where item responses are nonlinear (as often the case with educational data) may be challenged.

As alternatives, parametric nonlinear factor analytic (NLFA) methods have been proposed. Such methods have been incorporated in procedures including the limited-information, covariance-based method, NOHARM (Fraser & McDonald, 1988) and the full-information based method implemented in

TESTFACT (Wilson, Wood, & Gibbons, 1991). More detail on the NLFA procedure of NOHARM as a dimensionality assessment tool is provided in the subsequent section.

For either linear or nonlinear factor analytic approaches, the dimensionality of item responses can be achieved by appropriate factor identification and examination of the patterns of loadings. Identifying an appropriate number of factors will reflect the dimensionality of the data. It is desirable that the most parsimonious test structure is obtained while at the same time adequate account for the relationships between the items and factors is produced. The issue of proper identification of the number of factors has been debated in the literature. Because the FA approaches were developed originally for continuous data, there has been limited research on their use with dichotomously scored data.

Determining the number of factors. Empirical criteria are frequently used to determine the number of factors that should be extracted, including eigenvalues-greater-than-one criterion (eigenvalues > 1; Guttman, 1954; Kaiser, 1960) based on eigenvalues from a correlation matrix, the scree test (Cattell, 1978), as well as the less commonly applied techniques of the minimum average partial test (MAP; Velicer, 1976) and parallel analysis (PA; Horn, 1965). Other methods could also be applied in determining the number of factors, including decisions based on setting *a priori* desired amount of variance to be accounted for (i.e., selecting the fewest number of factors that reach that amount) and using ML

estimation procedure to estimate the model and compare it to models of higher dimensionality (i.e., nested models comparison via a χ^2 difference test).

The four methods for determining the number of factors (eigenvalues >1, scree test, PA, and MAP) are introduced next. For a review of widely used procedures for determining the number of factors and recommendations for use, see Velicer, Eaton, and Fava (2000).

The *eigenvalues > 1* rule is one of the most commonly used methods in determining the number of factors (Zwick & Velicer, 1986), and it is often a default in common statistical packages (e.g., SPSS, SAS). This rule suggests that the number of factors to be retained from the data should reflect the number of eigenvalues from a correlation matrix that are larger than 1. Research has shown that this rule tends to extract too many factors (i.e., over-extraction), especially in small to moderate sample size in sample data due to capitalization on chance (e.g., Cliff, 1988; Fabrigar, Wegener, MacCallum, & Strahan, 1999; Horn, 1965; Hubbard & Allen, 1987; Preacher & MacCallum, 2003; Revelle & Rocklin, 1979; Zwick & Velicer, 1982, 1986). It is also noteworthy to say that the research has shown that an over-extraction is typically more favorable than under-extraction when it comes to determining the number of factors using any of the extraction methods (Fava & Velicer, 1992, 1996; Wood, Tataryn, & Gorsuch 1996).

The *scree test* (Cattell, 1966) is another criterion that can be utilized in determining the number of extracted factors. The scree plot is a graphical representation of the plotted eigenvalues in descending order. The number of

factors (components) retained using the scree plot is done such that the number of factors above the “elbow” is retained. In other words, graphically, as the eigenvalues tend to level off factors above that leveling point should be retained.

Given its subjective nature, a problem in determining the number of factors via scree plot could arise when there is no clean break between the plotted eigenvalues (i.e., several eigenvalues around the elbow point). The method was also found to be less accurate in smaller sample size and complex patterns (Zwick & Velicer, 1982). Even with its subjective nature, scree plots have shown to yield more accurate results than the eigenvalue > 1 rule, especially with large sample size and strong factors (e.g., Zwick & Velicer, 1982). Further, the scree plot has been recommended to be used in conjunction with other procedures rather than a standalone method (Crawford, et al., 2009, Velicer, et al., 2000).

The *minimum average partial* (MAP; Velicer, 1976) method is based on the matrix of partial correlations. In this approach, after each of the *m* factors (components) is partialed out, the average of the squared correlations of the off-diagonal partial correlation matrix is computed (Velicer, 1976, developed this method for use with principal components analysis, although factors and components can be used interchangeably to represent a dimension). The number of components retained is determined by the point where the average squared partial correlation reaches a minimum. This occurs when the residual matrix most closely resembles the identity matrix.

MAP can be applied to any covariance matrix, and as an exact method, it yields results where at least two variables have high loadings on each retained component; and it directly relates to the concept of factors representing more than one variable (Zwick & Velicer, 1986). MAP method has been shown to be more accurate in determining the number of factors than eigenvalues > 1 rule (e.g., Zwick & Velicer, 1982, 1986). MAP is not a standard procedure in major software packages, although some programs have been written to implement MAP procedure (e.g., Gorsuch, 1991; Reddon, 1985).

Lastly, *parallel analysis* (PA) has shown to be an accurate procedure for determining the optimal number of factors (Fabrigar, et al., 1999; Horn, 1965; Hubbard & Allen, 1987; Preacher & MacCallum, 2003; Zwick & Velicer, 1986). The process of conducting PA begins with generation of a number of correlation matrices of random variables based on the same sample size (*N*) and number of variables (*J*) used in the real dataset. Factor analysis is then performed on the random data and the average (or some percentile of; e.g., 95^{th} or 99^{th}) eigenvalues from the random data (i.e., random eigenvalues) are compared to the eigenvalues from the real data (i.e., observed eigenvalues). The development and implementation of the programs for conducting PA have been traditionally done for the continuous data. This presents problems when dealing with categorical and binary data often find in educational settings.

Recent limited literature on PA for binary data provides inconclusive results and recommendations for conducting PA on binary data. For example,

Cheng and Weng (2005) found that using large sample size and high loadings, using 95^{th} (or 99^{th}) percentile, closer proportions in the categories result in adequate PA performance. The authors examined performance of phi and tetrachoric correlation matrices and found that in the two-dimensional cases, if PA erred it tended to incorrectly extract too many factors (especially with tetrachoric matrices). Further, poor PA performance was noted in small sample sizes (< 200) and extreme distributional proportions, and when low loadings were present regardless of the sample size.

Other research, however, suggests that PA should not be conducted on binary data due to the problematic nature of the method originally developed for continuous data (e.g., difficulty factors, Tran & Forman, 2009). Further, as noted above, indefinite positive correlation matrices often occur in binary data, thus present problems in conducting PA (e.g., Tran & Formann, 2009). Programs currently available to conduct PA are based on the notion of continuous data, and although the observed data can be in a form of phi, tetrachoric, or polychoric matrix, the random datasets generated for comparison are not.

Multidimensional item response theory (MIRT) approach. MIRT has received a lot of attention beginning in the 1970s and 1980s, when traditional IRT models were expanded to realistically represent various educational assessment experiences where any one person's response to an item was assumed to be influenced by multiple latent traits (Yeh, 2007). The link between the factor analysis for dichotomous variables and the normal-ogive model helped further the

development of MIRT. In 1972, Reckase first proposed an extension of the Rasch model to the multidimensional case. A number of general Rasch models with the growth of the logistic form of the MIRT model followed (McKinley & Reckase, 1982), which further led to developments of two- and three-parameter MIRT models. All of these models, both normal-ogive and logistic, could be characterized as *compensatory* (linear) models.

Another set of models developed concurrently in the seventies and eighties was known as *noncompensatory* or (*partially*) *conjunctive* MIRT models (e.g., Sympson, 1978; Whitely, 1980). The key difference between these two sets of models resides in how the latent traits interact with each other to produce the item responses.

Compensatory MIRT model. In compensatory MIRT, if an item on a test requires two different proficiencies (i.e., can be modeled with a two-dimensional space), a person's high proficiency on the first latent trait may compensate the lack of proficiency on the second (or vice versa); thus making it still somewhat probable that a person will respond correctly to the item. For example, two dimensions may underlie a mathematics word problem. The first dimension might reflect mathematics proficiency, while the second dimension could reflect the reading proficiency. If a person has high reading proficiency, he or she may be able to compensate, to some extent, for his or her lower mathematics proficiency. The multidimensional compensatory 3-parameter logistic (MC3PL) model can be

represented by (Reckase, 1985, 1997; McDonald, 1997; Spray, Davey, Reckase, Ackerman, & Carlson, 1990):

$$P\left(\mathrm{X}_{ij}=1\middle|\boldsymbol{\theta}_{\mathbf{i}},\boldsymbol{a_j},d_j,c_j\right)=\ c_j+\left(1-c_j\right)\frac{\exp\left(a_{j1}\theta_1+a_{j2}\theta_2,\ldots,+a_{jm}\theta_{\mathrm{m}}+d_j\right)}{1+\exp\left(a_{j1}\theta_1+a_{j2}\theta_2,\ldots,+a_{jm}\theta_{\mathrm{m}}+d_j\right)}, \qquad 2.5$$

where the only difference between 1.1 and 2.5 is only in the metric of the model; in Equation 1.1, a normal density function is used (i.e., normal-ogive model), while in 2.5, a logistic function is applied to determine the scale; other terms are defined above.

When c_j is set to 0, the MC3PL becomes a multidimensional compensatory 2-parameter logistic (MC2PL) model. Further, when all of the discrimination parameters are set to 1, the model becomes a multidimensional 1-parameter model. The interpretation of the item parameters is similar to interpretation of the unidimensional IRT models. The person parameters in the model are represented as the elements of the $\boldsymbol{\theta}_{\mathbf{i}}$ vector. The number of dimensions that adequately model the data matrix is open to debate and the subject of this research.

Noncompensatory MIRT model. In noncompensatory MIRT, if an item on a test requires two different proficiencies, knowledge or mastery of one may not be able to compensate for the lack of the other. In other words, all underlying proficiencies need to be sufficiently high for an item to be solved correctly. For example, on a verbal analogy item, mastery of two components (proficiencies), rule construction and rule evaluation, may be required for a successful outcome (i.e., correct answer to an item). If a person has high ability on rule construction, but has low ability on rule evaluation, the probability of favorable outcome may

not be high. This kind of relationship is the reason why often these models are referred as *nonadditive* or *multiplicative*.

The noncompensatory multidimensional three-parameter logistic (MNC3PL) model (Sympson, 1978; Whitely, 1980) can be represented by the following function:

$$P\left(\mathrm{X}_{ij}=1\middle|\boldsymbol{\theta}_{\mathbf{i}},\boldsymbol{a_j},\boldsymbol{b_j},c_j\right)=c_j+\left(1-c_j\right)\prod_{m=1}^{M}\frac{1}{1+\exp\left[a_{jm}\left(\theta_{\mathrm{im}}-b_{jm}\right)\right]}\ , \qquad 2.9$$

where, b_{jm} is the location for item *j* along dimension *m*, and other terms were previously defined. The noncompensatory nature of this model is derived from the fact that the probability of the correct response cannot be greater than the minimum value of the product terms (Spray, et al., 1990). As the number of dimensions increase, the probability of the correct response decreases.

The noncompensatory (conjunctive) multidimensional models are less commonly used, possibly due to the increased number of estimated parameters they require when compared to their compensatory counterparts (Knol & Berger, 1991). Furthermore, it may not be always clear which model should be used. Often this is the case when the relationship between the abilities is unclear.

For example, some may suggest that math word problems should be modeled using the compensatory MIRT, suggesting that even if an examinee is not a good reader, his or her high ability level in mathematics could compensate poor reading skills, resulting in a high likelihood of favorable outcome. On the other hand, some might understand (and treat) the relationship between the

abilities of reading comprehension and mathematics to be noncompensatory, suggesting that if an examinee does not know how to read well, a favorable outcome will be very unlikely. This low probability of a correct response may be present despite having a high ability in mathematics, because without being able to read and understand what the question asks, knowledge of mathematics might not be applied. Both of these scenarios are plausible, and it is up to the researcher to decide which model represents the believed hypothesized relationship among the multiple abilities.

Relationship between FA and MIRT. Several researchers have shown formally the mathematical equivalence between the FA and compensatory MIRT models (e.g., Knol & Berger, 1991; Takane & De Leeuw, 1987). A typical FA model presented in Equation 2.6 assumes that the response variable for item *j*, X_j, is governed by a continuous, latent variable Y_j, and threshold $\tau_j$which dichotomizes an item into a "1" for correct and "0" for incorrect response (i.e., if the probability of a correct response is greater than the threshold, $X_j = 1$). Equation 2.4 can be thus rewritten as a normal distribution function Φ for a correct response (McLoad, Swygert, & Thissen, 2001):

$$P(X_j = 1|\boldsymbol{\theta}) = \Phi\left[\frac{\lambda_{j1}\theta_1 + \lambda_{j2}\theta_2 +, \ldots, \lambda_{jm}\theta_m - \gamma_j}{\sigma_j}\right], \qquad 2.10$$

where, θ_m represents the m^{th} latent trait and σ_j^2 is unique variance. If we let

$$a_{jm} = \frac{\lambda_{jm}}{\sigma_j} \text{ and } d_j = \frac{-\gamma_j}{\sigma_j}, \text{ where } \sigma_j = \sqrt{1 - \sum \lambda_{jm}^2}\,, \qquad 2.11$$

then Equation 2.10 can be rewritten as a normal-ogive MIRT:

$$P(\mathrm{X}_j = 1|\boldsymbol{\theta}) = \Phi\left[a_{j1}\theta_1 + a_{j2}\theta_2, \ldots a_{jm}\theta_\mathrm{m} + d_j\right]. \quad 2.12$$

Note that Equations 1.1 and 2.12 are equivalent when c_j is fixed to zero.

Due to equivalency of the Equations 2.10 and 2.12, MIRT parameters can be derived from FA model (see Equation 2.11) and FA parameters can be derived from MIRT parameters as following:

$$\lambda_{\mathrm{jm}} = \frac{a_{jm}}{\sqrt{1+\sum a_{jm}^2}} \text{ and } \gamma_{\mathrm{j}} = \frac{-d_j}{\sqrt{1+\sum a_{jm}^2}}. \quad 2.13$$

It is then at no surprise that the model identification in multidimensional item response model carries directly from the factor theory (e.g., Bollen, 1989); at a minimum, for any model to be estimated, the number of parameters estimated cannot exceed the amount of information contained in the variance/covariance matrix.

Parametric approaches. M*plus* (Muthén & Muthén, 1998-2006) is one of the most diverse and flexible software programs when it comes to modeling and dimensionality assessment as it can handle both continuous and dichotomous data and it supports both exploratory and confirmatory approaches to FA. Furthermore, missing data can be handled within M*plus*. In exploratory FA, M*plus* employs the least-squares based estimators.

Both orthogonal and oblique rotations are permitted and output produced by the program relevant to dimensionality assessment includes eigenvalues for the polychoric correlation matrix, residual correlation matrix, the root mean squared residual (RMSR), χ^2 statistic, and the root mean square error of approximation

(RMSEA). The inclusion of the lower asymptote, however, is not permitted in either the computation of the correlation matrix or the parameter estimation. Similar to exploratory FA, M*plus* in the confirmatory FA may use least squares estimation. Alternatively, it may use full-information maximum likelihood techniques to marginalize over the latent variables.

TESTFACT (Bock, et al., 1999) has capabilities for both exploratory and confirmatory modeling, although confirmatory modeling is limited to bifactor structures, where a single common factor is modeled with one or more orthogonal "group" factors (Version 3.0; Tate, 2003). TESTFACT is considered a full-information based method as it uses full item response vectors in applying the item factor analysis (the program can also apply tetrachoric correlations and use them in a limited-information approach to conduct the analysis).

In TESTFACT, least square or marginal maximum likelihood procedures can be used for parameter estimation. In situations where the tetrachoric correlation matrix is not positive definite, a "smoothing" procedure is applied by using all positive roots of the original tetrachoric matrix in order to arrive at a positive definite matrix (Tate, 2003). Though the program does not estimate the lower asymptote parameters, it does allow for their input by the user (once estimated outside the program using for example BILOG; Mislevy & Bock, 1982). TESTFACT produces a χ^2 statistic for model fit, and in order to assess dimensionality of exploratory solution in TESTFACT, Tate (2003) suggests

conducting the test of the difference of the χ^2 fit statistic by sequential inclusion of additional factors.

Normal Ogive Harmonic Analysis Robust Method (NOHARM; Fraser & McDonald, 1988) is a parametric nonlinear factor analytic (NLFA) method which can be used in either exploratory or confirmatory analysis. Item responses are represented by a nonlinear factor analytic model (i.e., normal-ogive), such as the one represented in Equation 1.1. As a method, NOHARM allows for various rotations (e.g., oblique, orthogonal) in exploratory analysis to provide approximate independent clusters (McDonald, 2000). The estimation procedure employed in NOHARM is unweighted least squares (ULS), which allows for analysis of large number of items and high dimensionality (Fraser & McDonald, 1988; McDonald, 2000; Reckase, Thompson, & Nering, 1997).

Like TESTFACT, NOHARM does not estimate the lower asymptotes; however, it does allow for user input of these values. NOHARM provides covariance residuals and root mean square residuals to summarize the lack of fit. As originally developed, NOHARM does not produce a formal statistic for the model fit. Tate (2003) suggests evaluating model fit by a degree of improvement as dimensionality increases. Specifically, if the higher dimensional model produces 10% or more decrease in RMSR over the preceding model, that dimensional model should be retained.

As is the case with other factor analytic methods, NOHARM produces various fit measures for a given factor model or solution. NOHARM produces a

residual matrix of differences between observed and expected proportions and RMSR (Stone & Yeh, 2006, p. 196). Additionally, factor loadings are provided for each factor solution and they also can be used in evaluating the structure. A formal test the goodness-of-fit of a particular dimensionality solution based on NOHARM output was introduced by Gessaroli and De Champlain (1996) as a X^2 type statistic. This statistic is based on testing the null hypothesis that the off-diagonal elements in the residual correlation matrix produced by the factor analysis equal zero (Finch & Habing, 2005, 2007). If the null hypothesis is not rejected, it can be concluded that the fitted model adequately approximates the observed correlations among the items (Finch & Habing, 2005). The approximate χ^2 statistic can be computed as:

$$\chi^2_{G/D} = (N-3)\sum_{j=2}^{J}\sum_{j'=1}^{j-1} z^{2(r)}_{jj'}, \qquad 2.14$$

where *N* is the number of examinees, *J* is the total number of items, j and j' serve to index the items to define the unique pairings of items, and

$$z^{(r)}_{jj'} = .5\log(1 + r^{(r)}_{jj'}) - .5\log(1 - r^{(r)}_{jj'}) \qquad 2.15$$

is the Fisher's *z* transformation of the residual correlation for a given item pairing, and

$$r^{(r)}_{jj'} = \frac{p^{(r)}_{jj'}}{\sqrt{p^{(0)}_{j}(1-p^{(0)}_{j})p^{(0)}_{j'}(1-p^{(0)}_{j'})}}, \qquad 2.16$$

where $p_j^{(0)}$ is the observed proportion of examinees getting item *j* correct and $p_{jj'}^{(r)}$ is the residual covariance between items *j* and *j*′. The resulting statistic is compared to the reference χ^2 distribution with degrees of freedom, $df = 0.5J(J - 1) - t$, where *t* is the number of independent parameters estimated in fitting the model; in exploratory models, $t = (1+m) \times J$ and *m* is the number of dimensions (Finch & Habing, 2005, 2007).

An alternative to the $\chi^2_{G/D}$ statistic for model fit based on NOHARM output is an approximate likelihood ratio (Gessaroli, De Champlain, & Folske, 1997):

$$ALR = \sum_{j=2}^{J} \sum_{j'=1}^{j-1} G^2_{jj'}, \qquad 2.17$$

where

$$G^2_{jj'} = -2 \sum_{k_j=0}^{1} \sum_{k_{j'}=0}^{1} p_{k_j k_{j'}} \ln\left(\frac{\hat{p}_{k_j k_{j'}}}{p_{k_j k_{j'}}}\right), \qquad 2.18$$

where $p_{k_j k_{j'}}$ and $\hat{p}_{k_j k_{j'}}$ are the observed and expected (model-implied) proportions of examinees with scores of k_j and $k_{j'}$ for items *j* and *j'* (0 or 1), respectively. Given the dichotomous scoring of items, $p_{k_j k_{j'}}$ and $\hat{p}_{k_j k_{j'}}$ yield four combinations: proportion of both items being correctly answered (p_{11}), proportion of both items being incorrectly answered (p_{00}), proportion where item *j* is correctly answered but item *j*′is not (p_{10}), and the proportion where item *j* is incorrectly answered but item *j*′is correctly answered (p_{01}).

NOHARM produces the expected proportions of examinees who receive 1s for both items *j* and *j'* ($\hat{p}_{11}$); the remaining expected proportions need to be determined outside the program using the formula for the expected marginal proportion of examinee answering the item correctly as given by McDonald (1997). McDonald (1997) originally provided formulas for calculations of the expected proportions in unidimensional case, the extension to multidimensional case is straightforward. The *ALR* statistic is compared to the reference χ^2 distribution with the same degrees of freedom as $\chi^2_{G/D}$.

In addition to evaluating dimensionality of item responses on a test level, assessing dimensionality can be conducted at the item-pair level. Methods such as the model-based covariance (*MBC*; Reckase, 1997) and Yen's (1984) Q_3 can be used to assess the assumed dimensionality. To date, most applications and software for assessing LD in item pairs (Chen, 1993) have been confined to assessing the fit of unidimensional models. Research on the performance of many of these indices in unidimensional conditions suggests that the assumed reference distributions (e.g., normal distributions for Fisher's r-to-Z transformation of Q_3) do not hold (Chen & Thissen, 1997). Thus, most applications employ cutoff values; for Q_3, values greater than 0.20 can be interpreted as evidence of sufficient positive LI to warrant concern for the adequacy of the model.

While the form of MBC or Q_3 does not prevent them from being applied to multidimensional models, the problems in defining the appropriate reference distributions are likely present if not exacerbated when fitting multidimensional

models. An alternative approach to constructing reference distributions invokes a Bayesian approach to model-checking. These approaches to using LI indices have been studied in unidimensional modeling contexts by Levy, Mislevy, and Sinharay (2009) and in multidimensional modeling contexts by Levy and Svetina (in press).

Nonparametric approach to dimensionality assessment. Unlike their parametric counterparts, nonparametric approaches do not impose any distributional assumptions. In this section, current commonly used methods and procedures to assess dimensionality based on conditional covariance theory are described.

DETECT (Dimensionality Evaluation to Enumerate Contributing Traits). DETECT (Kim, 1994; Zhang & Stout, 1999a, 1999b) is an estimation procedure typically used as an exploratory tool for dimensionality assessment. The goal of DETECT is to describe the structure of the multidimensional item dispersion relative to the test composite Θ_{TT}. In other words, DETECT partitions the items into clusters such that within a cluster, items are most homogeneous, and clusters themselves are widely separated reflecting an assumption of approximate simple structure. For a given partition of items into clusters, *P*, the theoretical DETECT index is calculated as:

$$D(P,\Theta_{\mathrm{TT}}) = \frac{2}{J(J-1)} \sum_{1 \le j < j' \le J} \delta_{jj'}(P)\, E\left[cov\left(\square_j, \square_{j'} | \Theta_{\mathrm{TT}}\right)\right], \qquad 2.20$$

where,

$$\delta_{jj'}(P) = \begin{cases} 1, \text{if } \theta_j \text{ and } \theta_{j'} \text{are in the same cluter of } P \\ -1, \text{otherwise} \end{cases} . \quad 2.21$$

Given a cluster *P*, the $\delta_{jj'}(P)$ manipulates the expected conditional covariance such that its value is added if items *j* and *j'* are in the same cluster, or subtracted, if items *j* and *j'* belong to different clusters. The nonparametric nature of DETECT is expressed through Θ_{TT}, which represents an estimate of the composite ability best measured by the exam (Finch & Habing, 2005). The advantage of using observed score as conditioning variable is that the composite score does not need to be estimated (this advantage pertains to nonparametric methods in general).

In DETECT, the direction of best measurement is approximated by using the observed (raw) score. DETECT uses two estimators to approximate the conditional covariance. The first estimator uses a total score to approximate the expected covariance among the item pairs. The second estimator uses a rest score (total score minus the two items in question) to approximate the expected covariance. Research has shown that using a total score, the estimator is negatively biased, and that using a rest score, the estimator is positively biased (e.g., Zhang 2007; Zhang & Stout, 1999a). Thus, the final estimator of expected conditional covariance is the average of the two estimators; this average was shown to be optimal in minimizing the bias (Yang & Zhang, 2001).

Thus, if approximate simple structure exists, the theoretical index *D* will be maximized at the correct dimensionality-based cluster partition *D** (i.e. when the partition matches approximate simple structure).

The maximum possible value of *D*, denoted as *D**, indicates the amount of multidimensionality the test displays (i.e. departure from being perfectly fitted by an unidimensional model; Zhang & Stout, 1999b) and is given by:

$$D^*(P, \Theta_{TT}) = \frac{2}{J(J-1)} \sum_{1 \le j < j' \le J} \left| E\left[cov\left(\text{□}_j, \text{□}_{j'} | \Theta_{TT} \right) \right] \right| . \qquad 2.22$$

That means that when the partition matches approximate simple structure the maximum value of DETECT will be obtained because all of the within-cluster conditional covariances will be positive and all between-cluster conditional covariances will be negative (Zhang & Stout, 1999b). The space for all possible partitioning *P* is large, thus in order to search the space intelligently, the DETECT procedure employs a generic algorithm in addition to hierarchical cluster analysis to limit the search (Roussos, et al., 1998; Zhang & Stout, 1999b).

Under the assumptions of unidimensionality, all conditional covariances have an expected value of zero, which is why dimensionality assessment may be thought of as searching for violations of LI in terms of local item dependence (LID; Roussos & Ozbek, 2006). Because the DETECT index estimates the average item-pair conditional covariances, the DETECT value can be thought of as an estimate of the average size of the violation of pairwise LI given a unidimensional model (i.e., an effect size for the amount of multidimensionality or average size of LID).

Research provides some guidelines for the interpretation of the value of DETECT index. Zhang and Stout (1999b) recommended interpreting the DETECT index value of > 1.00 as strong evidence of multidimensionality, values

.40 to 1.00 indicating moderate to large multidimensionality, values between .20 and .40 suggesting weak multidimensionality, and values less than .20 suggesting unidimensionality. Other recommendations are slightly more liberal in interpretation, such that such that values less than .20 indicate weak multidimensionality or approximate unidimensionality, values from .20 to .40 indicate moderate multidimensionality, .41 to 1.00 indicates moderate to large multidimensionality, and > 1.00 values indicate strong multidimensionality (Roussos & Ozbek, 2006).

If the test exhibits the approximate simple structure, the ratio of *D* and *D** will equal 1. Values less than one indicate divergence from the approximate simple structure.

$$r = \frac{D(P, \Theta_{\mathrm{TT}})}{D^{*}(P, \Theta_{\mathrm{TT}})}. \qquad 2.23$$

In practice values of r (sometimes referred to as r_{max}) greater than or equal to 0.8 are interpreted as indicative of approximate simple structure (Jang & Roussos, 2007; Stout et al., 1996). Additionally, if multidimensionality is present, another index produced by DETECT may be considered. *IDN* is the index which reports the percentage of the signs of the conditional covariances that achieve the goal of having all within-cluster conditional covariances be positive and all between-cluster signs be negative. Similar to the *r* ratio, higher values of *IDN* constitute more support for the hypothesis of approximate simple structure.

If the hypothesis of approximate simple structure is supported, the solution may be interpreted in terms of the number of homogeneous item clusters as the

number of dominant dimensions. This is possible because DETECT procedure outputs the number of non-overlapping clusters and items associated with each of the clusters. To the extent where there are clusters with few items or if approximate simple structure does not hold, inferring the number of dominant dimensions should be done with caution (Jang & Roussos, 2007; Zhang & Stout, 1999b).

Although DETECT can be used in a confirmatory mode, where the DETECT index is calculated for a partition pre-specified by a researcher, to date the primary use of DETECT in dimensionality assessment has been in exploratory analyses. Thus exploratory DETECT is utilized in the current study.

Within the exploratory DETECT, both *exploratory* and *cross-validated* DETECT indices can be calculated. The exploratory DETECT index is calculated based on the entire sample. The cross-validated DETECT index can be obtained by partitioning the dataset into two subsets, running the DETECT procedure on one (training) subset, obtaining the optimal partition, and reading in that optimal partition to be imposed on the second subset. If the dataset is not previously subsetted, DETECT can randomly split the data file for training and validation samples (the user can specify the number of examinees for each of the samples).

For example, in Monahan, et al. (2007), the cross-validated DETECT index was calculated such that a 50%/50% split was indicated for each sample. This choice dictates DETECT software to randomly select 50% of the examinees to belong to the training sample, and the remaining 50% to serve as the validation

subsample for each condition. Previous research suggests that cross-validated DETECT index may be useful in overcoming the bias found in the exploratory DETECT index when the number of items on the test is 20 or fewer (Jang & Roussos, 2006).

DIMTEST. DIMTEST (Stout, 1987, 1990) is a nonparametric, confirmatory procedure that detects departures from essential unidimensionality, where the null hypothesis tested states H_o: d_e =1. The first step in applying the DIMTEST procedure is to select a subset of items for the assessment subtest (AT). Items chosen for the AT should be selected based on their substantive analysis of item content, expert opinions or exploratory statistical analyses (e.g., factor analysis, cluster analysis, DETECT). To provide a meaningful test of the null hypothesis assessing essential unidimensionality, AT subtest items should be dimensionally maximally distinct from the direction of best measurement of the remaining items. The remaining items on the test are referred to as the partitioning subtest (PT).For a detailed presentation of earlier and current versions of DIMTEST, see Froelich and Habing (2008), Froelich and Stout (2003), and Stout et al. (2001).

The strength of the DIMTEST procedure lies in its power to detect departures from unidimensionality (Nandakumar & Yu, 1996; Stout et al., 2001). Similarly, DIMTEST is successful in discriminating between essentially unidimensional and multidimensional tests. DIMTEST was found to be robust with respect to minor secondary traits (Nandakumar, 1993), especially in studies

that fitted a compensatory MIRT (Hattie el al., 1996; Nandakumar, 1991; Nandakumar & Stout, 1993; Stout, 1987), where primary and secondary abilities followed standard normal distributions. As with DETECT, the DIMTEST procedure uses raw scores as the conditioning variable, and thus it does not support missing data. Unlike DETECT, however, DIMTEST does allow for the inputting of a single estimate of a guessing parameter applied to all dichotomously scored items.

Although DIMTEST is framed for assessing essential unidimensionality, it can be used to provide dimensionality assessment information in confirmatory multidimensional models with approximate simple structure. Stout et al. (1996) suggested assessing the multidimensional simple structure by using the assumed groupings of items to correspond to hypothesized structure. For example, if we are fitting a two-dimensional model with simple structure, the set of items that are associated with one factor serves as AT while the rest of the items serve as PT.

As discussed above, there are many methods and procedures currently available to assess a set of item responses on an exam. A researcher's choice of some or any of these methods may depend on accessibility, familiarity with the method(s), and the type of data at hand. For this project, DETECT and NOHARM procedures are selected because they are both current and popular methods used in dimensionality assessment. As discussed next, both procedures have been shown to work well at counting the number of dimensions when the underlying model is a compensatory with approximate simple structure.

More importantly, these methods are built on two different building blocks: the DETECT procedure is rooted in conditional covariance theory and the NOHARM method uses a factor analytical approach in assessing dimensionality. The inclusion of both methods will thus enable a comparative investigation of the procedures.

In their study, Finch and Habing (2005) undertook a quest in addressing the challenge set by McDonald (2000), who stated:

> These procedures [including DETECT] might result in useful applications, although a considerable amount of critical theoretical work, simulation, and empirical studies are needed to determine how they compare with the application of the well-known classical strategies [NOHARM] for dealing with these problems, and to establish their suitability for applications. (p. 99)

The current study seeks to extend this quest by paying particular attention to the data structure (exact simple versus complex) and the underlying MIRT model (compensatory versus noncompensatory). Prior to description of the design of the current study, the existing research on DETECT and NOHARM is summarized next.

Research Related to DETECT and NOHARM

Several researchers have investigated the performance of dimensionality assessment procedures. As argued above, many of the methods currently developed for dimensionality assessment perform well under certain conditions.

Research on the performance of DETECT, including addressing the issue of bias in the DETECT index, and NOHARM-based statistics, such as $\chi^2_{G/D}$ and *ALR*, have shed light into the workings of these procedures. Studies relevant to the current project are discussed next.

Research on DETECT. Zhang and Stout (1999b) provided the theoretical foundation for DETECT. In addition to the theoretical underpinnings of the procedure and building on previous work of Kim (1994), the authors demonstrated DETECT's performance via two simulation studies. In the first simulation study, Zhang and Stout (1999b) manipulated the number of dimensions (2, 3, or 4), the number of items (20 or 40), and the number of examinees (400 or 800) to generate the item-response data. Each of the conditions was replicated 100 times. Using a multidimensional compensatory model, data exhibiting approximate simple structure was generated. The authors found that as the number of examinees increased, the performance of DETECT improved. Holding the number of examinees and item constant and increasing the number of dimensions resulted in poorer performance of DETECT, especially with 20 items, 400 examinees, and 4 dimensions.

The second simulation study concerned unidimensional cases, with manipulated factors of test length (20 or 40), sample size (400 or 800), and the value of guessing parameter (.00 or .20). Zhang and Stout (1999b) found DETECT to be successful in verifying that the simulated tests were unidimensional in all cases. In summary, the authors found that, when

approximate simple structure held, DETECT performed well in identifying the dominant latent dimensions and estimating the amount of multidimensionality present in the test. Even when the approximate simple structure failed to hold, they argued that DETECT could still be informative, because it still could locate relatively dimensionally homogenous clusters. There would be no "best" partition among the clusters though, because there would be little separation between some clusters (i.e., an item pair from two clusters that are close to each other could have similar directions of best measurement and hence should be similar substantively, Zhang & Stout, 1999b, p. 215).

Van Abswoude, van der Ark, and Sijtsma (2004) investigated the effectiveness of Mokken Scaling procedure (MSP; Mokken, 1971), DETECT, and HCA/CCPROX (Roussos, et al., 1998) for dimensionality assessment in multidimensional data exhibiting simple structure. In their simulation study, they manipulated the MIRT model (extension of 2-PL model like the one in Equation 2.5 where c_j is fixed to 0, or a five parameter acceleration model), the number of dimensions (2 or 4), the correlations among the traits (.00, .20, .40, .60, .80, or 1.00), the number of items per trait (7, 21, or a combination), and the discrimination levels of the items (high or low). General findings suggested that DETECT and HCA/CCPROX outperformed MSP in retrieving the simulated dimensional structure. This was the case even when the correlation between the traits was high (.80). DETECT performed poorer in situations with low

discriminating items and longer tests, and in conditions where the number of items per trait was unequal.

The efficacy of DETECT depends greatly on the minimally biased estimation of conditional covariances for all item pairs (Roussos & Ozbek, 2006); thus understanding the extent to which the DETECT index is biased, is important. Specifically, bias has implications in describing magnitudes of multidimensionality present in data, as the theoretical DETECT index under unidimensionality equals zero. Thus, empirical bias, defined as the mean of the DETECT index over replications (Monahan, et al., 2007), can have an effect wherein researchers potentially (falsely) conclude the data are multidimensional when in truth they are unidimensional. The effect of bias might not directly impact the number of clusters DETECT finds, however, it certainly plays a role in evaluation of the magnitude of multidimensionality present in the data. Monahan and his colleagues (2007) caution that:

> Bias could lead one to conclude that item responses come from multiple dimensions, when in fact this result is simply due to statistical bias. Likewise, inflated standard error implies instability in the estimate of the DETECT indices. Such instability could lead one to conclude unidimensionality with one sample and multidimensionality with another sample. (p. 496)

The existing research, summarized next, has shed some light on the presence of bias in DETECT. While most research primarily focused on

unidimensional cases, Roussos and Ozbek (2006) generalized further to address multidimensional simulated item response data (see below).

Monahan, et al. (2007) examined the issue of bias in DETECT index with respect to the type of index (exploratory versus cross-validated) under the conditions of unidimensionality. In the simulation study, the authors manipulated the test length (5, 10, 15, 20, 40, and 80 items), the sample size (100, 500, 1000, and 5,000), and the IRT model used (1-PL, 2-PL, and 3-PL). For each of the 500 replications per condition, the authors calculated the exploratory and cross-validated DETECT index. Monahan et al. (2007) found the only significant interaction to be sample size by type of index, resulting in running separate analysis for each of the indices.

The authors found that bias was strongly related to the number of items for both indices. As the number of items decreased, the bias increased, especially in the exploratory index. Similarly, as the sample size decreased, the bias increased. Furthermore, at every combination of the test length and sample size, the exploratory index showed more bias than the cross-validated index. In terms of the IRT model underlying the item responses, the authors found little difference between bias found in the exploratory and cross-validated DETECT indices.

In addition to examining bias of the indices, the authors examined the standard errors and root mean squared errors for both exploratory and cross-validated DETECT indices. With respect to the standard errors, the cross-validated index showed greater amount of errors for all levels; differences in the

standard errors from exploratory and cross-validated approaches increased as the sample size and the number of items decreased. For example, for a sample size of 1000, the standard error of the DETECT index became problematic for 5 or fewer items for the exploratory index, and 10 or fewer items for the cross-validated index. The larger standard errors for the cross-validated index across these conditions are the result of fewer data (items or people). Little difference in the average standard error was found across the IRT models for either index. The results of the RMSE were opposite of those found for the standard errors. The RMSEs were greater for the exploratory than the cross-validated DETECT index for all levels of all factors, particularly in conditions with fewer examinees.

In summary, Monahan et al. (2007) found that bias in exploratory DETECT index appeared to be strongly related to both the sample size and the test length, while bias in the cross-validated index appeared to be influenced largely by the test length. Standard errors in cross-validated DETECT index were affected by both the sample size and the test length, while in exploratory DETECT index, only the test length seemed significant. Overall, Monahan and his colleagues (2007) agreed with previous research by Zhang and Stout (1999b) when suggested that cross-validated index should always be preferred over the exploratory index when DETECT is utilized.

Roussos and Ozbek (2006) evaluated the amount of statistical bias present in the DETECT index using very large sample size (120,000). The authors simulated data to follow a variety of dimensionality structures. The authors

manipulated the following factors: the number of dimensions (1, 2, or 3), correlations among dimensions (.50, .70, or a combination), the number of items per dimension (ranged from 5 to 40 in unidimensional, and 20 or 40 in multidimensional case), and the data structure (simple or approximate simple).

The authors found that the DETECT estimator had some statistical bias in unidimensional cases, particularly in conditions with 10 or fewer items. Based on these results, the authors suggested not to use DETECT with fewer than 20 items. In multidimensional cases, the authors found that the large-sample DETECT index showed "remarkably small bias for all simulated conditions (Roussos & Ozbek, 2006, p. 237). Furthermore, the authors found that DETECT had a high accuracy rate in forming clusters. Only three out of 45 multidimensional cases had less than perfect accuracy rate (i.e., 100%), with the lowest classification rate being 91% for the two-dimensional condition with test length of 20 items, approximate simple structure, and .7 correlations between the traits. Additionally, Roussos and Ozbek (2006) found some bias in the estimator of the conditional covariance (*IDN*). Similar to bias in the DETECT index, bias in the estimator of the conditional covariance decreased as the test length increased.

In an extensive simulation study, Finch and Habing (2005) compared the performance of exploratory DETECT and NOHARM-based statistics, $\chi^2_{G/D}$ (Equation 2.14) and *ALR* (Equation 2.17) where two- and six-dimensional datasets were generated. The authors manipulated the following factors: the type of the MIRT model (2PL or 3PL), the number of items (15, 30, or 60) and

subjects (1000 or 2000), the skewness of the latent traits (-1.5, -.5, 0, .5, 1.5), and correlations among the traits (.00, .30, .80, or .95). It is noteworthy that the authors also included two different sets of item parameters; one set reflecting a rather easy test (basic skill), while the other set of item parameters reflecting a more difficult exam. Each condition was replicated 500 times.

The authors used four criteria to evaluate the performance of the two methods: a) the ability to perfectly recreate the dimensional structure; b) the proportion of items falsely separated; c) the proportion of items that were falsely grouped into the same cluster; and d) the number of dimensions found. While DETECT outputs the number of clusters it finds, making the identification of the number of dimensions straight forward process, NOHARM does not. Finch and Habing (2005) recommend using a sequential procedure in determining the number of factors.

First, for each *K*-dimensional fitted model, $\chi^2_{G/D}$or *ALR* is calculated. The sequential testing begins by subtracting the calculated statistic from the *K*-dimensional model from the statistic from the (*K*-1)-dimensional model. The difference is treated as a χ^2 variate with degrees of freedom equal to the difference in the number of estimated parameters. If this difference is larger than the critical value based on the appropriate χ^2 distribution, it is inferred that the *K*-dimensional model is favored and selection stops. Alternatively if this difference is less than the critical value based on the appropriate χ^2 distribution, then it is inferred that the models fit equally well and the procedure is repeated, comparing the (*K*-1) -

dimensional model to the (*K*-2)-dimensional model. Once the preferred model is selected, NOHARM output for that model (i.e., the estimated factor loadings) is used in reporting of the results.

Finch and Habing (2005) found that in two-dimensional case, the two procedures performed similarly well. In the case where the parameters reflected a basic skills test, the DETECT procedure was more likely to achieve perfect matches (i.e., perfectly recreate dimensionality structure) when the correlation among dimensions was low, and the two procedures performed equally when the correlation was .80 or higher. This difference at lower correlations was less pronounced in the conditions with parameters that reflected the more difficult test, where DETECT and *ALR* performed similarly in selecting the number of dimensions. The number of subjects did not seem to have a great impact on the ability for either approach to identify the number of underlying dimensions and to group the items correctly. The number of items and the skewness, however, seemed to result in the shift of the rates of the perfect matches: *ALR* and DETECT performed similarly under 15 and 60 item conditions, but not for conditions with 30 items. For 30 items, performance of both declined with respect to the perfect match rates in both sets of item parameters and for models with and without guessing.

The results for the six-dimensional conditions suggested that the *ALR* type statistic outperformed the DETECT in the perfect match rates, most notably due to the deterioration of performance of DETECT (as compared to the two-

dimensional cases).[1] As in the two-dimensional conditions, the number of subjects did not seem to have an impact for either *ALR* or DETECT; however, in terms of the items, opposite effects were found. For *ALR*, an increase in the number of items resulted in higher rates of the perfect matches. In DETECT, increase in the number of items resulted in worse performance.

In addition, when errors occurred, *ALR* appeared to group items that should have been kept separate, while DETECT separated items that should have been grouped together. This pattern generally held for both two- and six-dimensional conditions, regardless of the number of items and the number of examinees.

Finch and Habing (2005) suggested that the relative performance of the two methods was dependent on the number of dimensions; where DETECT outperformed *ALR* for two-dimensional case, while the opposite was true for the higher, six-dimensional conditions. Furthermore, regardless of the number of dimensions, when the methods erred, DETECT tended to overestimate the number of clusters and falsely separate the items, while *ALR* tended to falsely combine the items into clusters. Unlike Finch and Habing (2003), Finch and Habing (2005) found that guessing had little effect on either of the methods.

Perhaps the most relevant study involving DETECT for the current project is the Gierl, et al.'s (2006) study. They evaluated the performance of DETECT in

[1] Due to superior performance of *ALR* over $\chi^2_{G/D}$, the authors only reported results for *ALR* in comparison to DETECT.

terms of its classification accuracy and consistency in situations where the data displayed various degrees of complex structure (i.e., item pattern structures differed). In their simulation study, Gierl and his colleagues examined datasets with 40 items that followed two-dimensional structures and manipulated three variables: degree of complexity (0%, 10%, 30%, or 50% of items display complex structure where items have angular direction between 35° and 55° relative to dimension 1), correlations between dimensions (.00, .30, .60, .75, or .90), and the sample size (500, 1000, or 1500). Each condition was replicated 100 times.

The authors were interested in examining the classification accuracy (defined as the number of times that an item was correctly assigned to a cluster by DETECT when compared to its true cluster membership) and the classification consistency (defined as the number of times that an item was classified in the same cluster for two randomly equivalent samples). They considered classification rates to be acceptable when the agreement between true dimension and DETECT classification met or exceeded .90 (90%).

Overall results for classification suggested that DETECT was very successful in accurately recovering the dimensional structures in conditions where the correlation between traits was .60 or lower for all sample sizes and across all structures. An exception was found in a condition with small sample size (i.e., N = 500), correlation of .60 between dimensions, and highly complex data structure (i.e., 50%), where the accuracy rate was .84. As the correlation increased and the

degree of items exhibiting complex structure increased, the performance of DETECT was diminished.

Classification results for the complex structure items alone showed that DETECT was able to successfully classify complex items in conditions with uncorrelated latent traits regardless of the sample size. In the remaining conditions, the correlations between the latent traits and sample size became more noteworthy. For complex structures where the correlation between the dimensions was .30, DETECT obtained high classification rates for both 1000 and 1500 examinees. However, accuracy rates fell below 90% when the sample size dropped to 500. At correlations of .60, a sample size of at least 15000 was required to yield satisfactory classification rates. DETECT failed to recover satisfactory the dimensional structure for any sample size when correlation between the traits was .75 or .90.

With respect to the consistency of the DETECT's performance, the authors found that in conditions of all sample sizes and correlations between the latent traits of .60 or below, high consistency rates were obtained. In only four conditions, the consistency rate was below the desired .90, including the .60 correlation, 30% complex, and N = 500; .30 correlation, 50% complex, and N = 500; and .60 correlation, 50% complex with N = 500 and N = 1000. The consistency rates exceeded .90 for all sample sizes when correlation was .75 and simple structure was present. However, as the amount of complexity increased,

larger sample sizes were required for satisfactory performance. At .90 correlation, none of the conditions produced high consistency rates.

In summary, Gierl et al. (2006) found that DETECT produced high classification and consistency rates for most conditions where the correlation between latent traits was .60 or lower. Further, the authors concluded that DETECT can adequately classify items in two-dimensional space for some complex structures, particularly when 30% or less items are complex, correlation between the traits is $\leq .75$, and $N \geq 1000$. The authors recommend that in cases when large numbers of items are expected to display complex structure, DETECT should be used for dimensionality analysis with large sample size, $N \geq 1500$ and in situations where latent traits are correlated up to .60.

Research on NOHARM-based statistics. Researchers have suggested that NOHARM "model provides a sound theoretical framework on which indices as well as statistics could be developed to determine the number of dimensions which are adequate for item response modeling" (Gessaroli & De Champlain, 1996, p. 157). To that extent, Gessaroli and De Champlain (1996) investigated usefulness of the NOHARM-based $\boldsymbol{\chi^2_{G/D}}$ and Stout's *T* statistic (implemented in DIMTEST, here after referred to as DIMTEST) in identifying unidimensional and two-dimensional structures.

In generating unidimensional simulated data, the authors manipulated the sample size (500 or 1000), the test length (15, 30, or 45 items), and test reliability expressed by using varying means and standard deviations for discrimination

parameter (weak, moderate, or strong). For two-dimensional cases, they added an additional factor: dimension dominance. While each of the multidimensional structures displayed simple structure (i.e., each item only relates to one dimension), the balance of items belonging to a dimension varied (equal or unequal number of items associated with each dimension). Both the empirical Type I error rates (α = .05) based on unidimensional dataset and the rejection rates based on the multidimensional datasets were obtained; each condition was replicated 100 times.

The $\chi^2_{G/D}$ statistic correctly identified unidimensional model in most of the unidimensional conditions, with a maximum number of rejections being four (out of 100) in any one condition. Further, $\chi^2_{G/D}$ correctly rejected unidimensionality in two-dimensional datasets 95 out of 100 times in any one condition. The authors concluded that for the studied conditions, the $\chi^2_{G/D}$ statistic had both good control of the Type I error and good power. Gessaroli and De Champlain (1996) further suggested that the performance of the statistic improves as the test length, sample size, and test reliability increase. Test structures with unequal number of items per dimension resulted in poorer performance, although that performance was still largely satisfactory.

In terms of the DIMTEST, when unidimensional data were simulated, the Type I error rates came very close to the nominal levels in all conditions. In two-dimensional cases, DIMTEST performed well for conditions with larger sample size (N = 1000) and test lengths of at least 30 items.

Comparing the performance of the two statistics, the authors concluded that the major differences in performance were found in conditions with fewer items (i.e., 15) and a sample size of 500, where $\chi^2_{G/D}$ clearly outperformed DIMTEST. With respect to the Type I error rates, $\chi^2_{G/D}$ was more conservative than DIMTEST. The performance of $\chi^2_{G/D}$ in rejecting the false null hypothesis was very comparable to DIMTEST in conditions where DIMTEST was known to perform well, and much higher in other conditions (i.e., smaller sample size and fewer items). Overall, the authors concluded that $\chi^2_{G/D}$performed well under the studied conditions, although the authors recognized that the set of conditions was limited (e.g., uncorrelated factors, no lower asymptote parameter).

In a different study, De Champlain and Gessaroli (1998) examined the usefulness of the $\chi^2_{G/D}$ statistic by comparing it to the performance of two other statistics: likelihood-ratio χ^2 statistic provided in TESTFACT and the χ^2 goodness-of-fit statistic provided in LISREL8. In this simulation study, both unidimensional and two-dimensional structures were examined. In unidimensional cases, the authors generated data employing a 3PL model, by manipulating the number of examinees (250, 500, or 1000) and the number of items (20 or 40). Two-dimensional datasets were generated via compensatory model using the same factors as in unidimensional case, with two added factors; correlation between traits (.00 or .70), and item pattern structure (simple versus complex), where complex datasets included 50% of items to load on both dimensions equally

strong (i.e., same loadings on both dimensions). Each condition was replicated 100 times.

In examining the statistics, De Champlain and Gessaroli (1998) found that in comparison to other indices, $\chi^2_{G/D}$ had desirable characteristics: near the nominal Type I error rates ($\alpha = .05$) in unidimensional cases (largest error rate of .07 was found in condition of 40 items and 1000 examinees) and high power rates to reject the multidimensional models. Further, $\chi^2_{G/D}$ was successful in identifying true multidimensional nature of the simulated datasets, for both correlated and uncorrelated conditions.

Most importantly, initial results from this study suggested that $\chi^2_{G/D}$ was relatively unaffected by the sample size, the number of items, the item parameter structure, and correlations between the traits considered in the study. The authors cautioned that the results while encouraging pertained to only the restricted set of conditions as outlined in the study design and called for further investigations to include more complex, multidimensional, models.

Finch and Habing (2007) further examined the performance of the goodness-of-fit statistics based on NOHARM by comparing them to DIMTEST in detecting the violations of unidimensionality. The three NOHARM-based statistics included in the study were $\chi^2_{G/D}$, *ALR*, and T_s, a goodness-of-fit statistic proposed by Maydeu-Olivares (2001).

Via a simulation study, the authors examined both the Type I error rates and the power of the procedures. The manipulated factors included: the

underlying model (2-PL or 3-PL), the number of items (15, 30, or 60), the sample size (1000 or 2000), skewness (-1.5, -.5, 0, .5, or 1.5), the value of c parameter (constant for all items versus varying), and for two-dimensional sets, the correlation between the traits (.00, .30, .80, or .95). Two-dimensional data were generated following the compensatory MIRT described by Reckase (1997, see Equation 2.5).

In addition, the authors used two sets of item parameters. The first set represented a basic skill test, with the mean (standard deviation) of discrimination and difficulty +.97 (.32) and -.92 (.76), respectively. The second set approximated parameters on a test representing non-basic skills, such that the mean (standard deviation) of discrimination and difficulty were 0.00 (.35) and 0.00 (1.00), respectively. Each of the conditions was replicated 500 times, and Type I error rates and power rates were calculated for each of the procedures.

In models with no guessing, $\chi^2_{G/D}$ seemed to display Type I error rates that were lower than those of other statistics for both sets of item parameters (the only exception was found in the 15-item condition where the *ALR* and T_s had lower Type I error rates based on $\alpha = .05$). *ALR* had lower Type I error rates than DIMTEST for most of the conditions, and also lower Type I error rates than T_s for conditions with 30 and 60 items. In shorter exams and larger sample sizes, both *ALR* and DIMTEST displayed increased Type I error rate, while T_s tended to have elevated rates for larger sample size and more items. Skewness seemed to affect *ALR* more than either $\chi^2_{G/D}$ or DIMTEST, particularly when negative skew

existed in either set of parameters. Skewness also impacted Type I error rates of the T_s, although here both the positive and the negative skew made an impact.

In models with guessing, the NOHARM-based statistics best performed when the actual (varying) *c* parameters were provided, as opposed to situations in which the *c* parameters were constrained to be at a constant value for all items or were not provided at all. The difference in performance, however, was not large (differences in Type I error rates were never > .02). Unlike the conditions 2-PL conditions, in 3-PL conditions, DIMTEST had lower Type I error rates than the NOHARM-based statistics across all other manipulated factors. Out of the three NOHARM-based statistics, T_s, had Type I error rates closest to the nominal value and was most comparable to the DIMTEST results. The T_s statistic had elevated Type I error rates for larger sample sizes and more items than DIMTEST. Neither *ALR* nor $\chi^2_{G/D}$ maintained the error rate at the nominal levels for the 3-PL, with one exception (*ALR* in the condition with 15 items, no skew, and basic skill item parameters).

ALR and $\chi^2_{G/D}$ had slightly higher power rates than DIMTEST across all levels for both sets of parameters in conditions where no guessing was introduced; except when the correlation between the dimensions was .95. T_s had generally lower power rates than the other statistics, although the pattern was not uniform. In conditions with present pseudo-guessing parameter, due to high Type I error

rates, the empirical power for all four statistics was calculated.[2] *ALR* statistic had the highest empirical power in the 3PL conditions using the non-basic item parameters than the other statistics for the most situations. In conditions where the data were generated using the basic skills parameters, in most study conditions, $\chi^2_{G/D}$ had comparable power to *ALR* and DIMTEST, whereas the T_s again showed slightly lower power. Overall, the power for all four statistics was higher for longer tests, especially for DIMTEST with basic skills set of parameters, and no skew in the latent abilities. Further, as the correlation between the traits increased, the power rates decreased in the statistics.

Finch and Habing (2007) concluded that the relative performance of the DIMTEST and NOHARM-based statistics depended on the model underlying the item responses. If the guessing is known not to be present in the data, one of the NOHARM-based statistics should be used; however, if guessing is present, DIMTEST might be more appropriate as it maintains the nominal Type I error rate (and has comparable power to the NOHARM-based statistics). Furthermore, the authors warn that if the data are skewed, power of any of the studied statistics will decrease and the Type I error rate will likely increase.

The recent literature outlined above suggests that performance of DETECT and NOHARM-based procedures show promise in dimensionality

[2] Empirical power was calculated such that first the empirical .05 critical value for all four statistics was determined. Then, based on those values, the power rates were recalculated using the new values of the statistics.

assessment. In particular, the NOHARM-based and DETECT methods generally perform well under conditions with larger sample sizes and lower correlations between dimensions, with simple and approximate simple structures, and when the underlying multidimensional model is compensatory. However, NOHARM-based statistics did not perform well in situations with nonnormal data and higher correlations between dimensions, and DETECT was found to perform poorly with large number of dimensions, low discriminating items, and smaller sample sizes.

Broadly stated, more is to be learned about the efficacy of the procedures, particularly in situations that depart from foundations upon which the procedures (or associated statistics) are built upon, as the performance of either method is limited to the conditions examined in the current studies. Aspects of inclusion of the *c* parameter or complex structure have been investigated in only a few studies, and under a limited set of conditions. To date, compensatory models have been used almost exclusively to generate data that are then used in methodological research on dimensionality assessment. Thus, the performance of these methods when data are generated using noncompensatory model is largely unknown.

In order to provide additional utility and generalizability to the statistics and procedures, conditions that include different models (e.g., noncompensatory), data structures (e.g., complex), or estimation procedures are needed. This work attempts to contribute to the literature on the performance of the procedures from an exploratory perspective, primarily focusing on the issue of complex data

structure in a multidimensional space (> 2 dimensions) using two different underlying models (compensatory and noncompensatory), described next.

Chapter 3

METHODOLOGY

This study is primarily motivated by the general lack of research in the area of dimensionality assessment in complex data structures. The purpose of this study is to investigate the effect of complex structure in dimensionality assessment by using current, easily accessible tools; specifically, NOHARM and DETECT procedures are used in this study. Five methods are considered in the study: DETECT-based exploratory (DETECTe), DETECT-based cross-validated (DETECTcv), NOHARM-based $\chi^2_{G/D}$, NOHARM-based *ALR*, and NOHARM-based RMSR.[3] In the simulation study, manipulated factors are selected such that they address previously established hypotheses that reflect a number of different, yet plausible, testing situations, and build off existing research, including Gierl et al. (2006).

Study Design

The following factors are manipulated in the study: a) number of dimensions, b) structure type of data, c) correlations between dimensions, d) MIRT model type, e) sample size, and f) number of items per dimension. In Table 1, the study design is presented in a tabulated form for a quick review.

[3] Here and throughout the study, when discussing performance of $\chi^2_{G/D}$, *ALR*, and RMSR methods, it is implied that these methods are obtained using NOHARM output and are being evaluated as such. Thus, it is the methods based on the output that are being evaluated, as opposed to the NOHARM procedure itself.

Table 1.

Manipulated Factors For Data Generation

Factors	Levels	Total # Levels
Dimensions	2 or 3	2
Data Structure	0%, 10%, 30%, or 50%	4
Correlations	.00, .30, .60, .75, or .90	5
MIRT Model	Compensatory or noncompensatory	2
Sample Size	500, 1000, or 2000	3
Items/dimension	10 or 20 per dimension	2
	Total # of Conditions	480

Number of dimensions. Two different multidimensional data structures are examined: 2- (2D) and 3-dimensional (3D) structures are considered in the study. Gierl et al. (2006) considered 2D structures; the current includes the 2D structure, and also includes 3D structures. Typically, research in dimensionality assessment includes two to three levels of dimensional space (e.g., two- and six-dimensional spaces were simulated in Finch & Habing, 2005; one- and two-dimensional data were simulated in De Champlain & Gessaroli, 1998).

Structure type of data. In order to investigate the effect of complex data, the percent of items in the data that are factorially complex is manipulated. Following Gierl et al. (2006), the percent of items in the data modeled as complex included: 0%, 10%, 30%, or 50%. The amount of complexity is held constant with respect to dimensionality that is modeled. For example, in a condition with 2D, 10 items per dimension, and 10% of complex items, one item associated with

each dimension is modeled as complex, for a total of 2 complex items on the test.

Correlations between dimensions. Correlations among dimensions in the population include the values of .00, .30, .60, .75, and .90. Within 3D conditions, the three correlations were constant. The aim is to cover a range of possible correlations for generalizability purposes. Similar values of correlations were examined in previous studies (see Literature Review section). Further, correlations such as these are often found in empirical studies of educational tests (Jang & Roussos, 2007).

Model type. Multidimensional data are simulated from either a 2-parameter compensatory or noncompensatory MIRT model (see equations 2.5 and 2.9, respectively). To date, little work has been done utilizing noncompensatory models. Further, both NOHARM and DETECT are grounded on the compensatory models, making the inclusion of noncompensatory important for evaluating the generalizability of these approaches to analyzing data that follow noncompensatory models.

Sample size. Recent studies examining the performance of either DETECT or NOHARM typically investigated a range of sample sizes, including 500 and 1000 (e.g., Finch & Habing, 2005, 2007; Gessaroli & De Champlain, 1996). This study examines sample sizes of 500, 1000, and 2000.

Number of items per dimension. In order to investigate the effect of the number of items on the performance of the two methods, the number of items per dimension is manipulated. The number of items associated with each dimension is

set to be either 10 or 20. The choice of 10 or 20 items per dimension yields different test lengths to be examined: for 2D tests, the test length is 20 or 40 items, and for 3D tests, the test length is 30 or 60 items. The choice of examining these test lengths comes from surveying the current literature, where similar test lengths were employed (e.g., Gierl, et al., 2006; van Abswoude, et al., 2004).

Data Generation

All item responses are generated using R (R Development Core Team, 2010) such that each item response conforms to the conditions outlined above. The above presented study design yields a total of 480 conditions, and each condition is replicated 500 times (Finch & Habing, 2007; Harwell, et al., 1996).

Item parameters used to generate the data are presented in Tables 2 through 7. For both compensatory and noncompensatory models, the literature was surveyed to determine typical parameter values found in realistic testing scenarios (e.g., Bolt & Lall, 2003; Embretson, 1983; Gierl, et al., 2006). The selected item parameters are *fixed* across all conditions and they range in values to approximate a typical educational assessment. For conditions with 20 items per dimension, the item parameters presented in the tables are doubled (tripled) for the 2D (3D) conditions. In order to avoid confounding of difficulty and dimensionality (as shown in Reckase, et al., 1986), item parameters are balanced across dimensions for all conditions. The lower asymptote parameter for all conditions is fixed to 0.

Little is known about the performance of dimensionality assessment methods in cases of complex structure, and therefore other factors that optimize the performance are preserved as much as possible. Person parameters are generated from multivariate normal distributions with an appropriately sized mean vector of $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}$, where the diagonal elements of $\mathbf{\Sigma}$ are all 1 and the off-diagonal elements are given by the correlation for the associated condition.

Estimation Methods

For the purpose of examining (and comparing) their performance in conducting dimensionality assessment, exploratory DETECT and NOHARM methods are utilized with their default options. For DETECT, this means that the MINCELL option is set at its default value of 2, where the value indicates the minimum number of examines required to be present in any one cell when calculating the conditional covariances. The MUTATIONS option allows for specification of the number of vectors that are mutated in the genetic algorithm, and per Monahan, et al. (2007), it is set to equal the recommended value that ranges between one fifth to one tenth of the total number of items (e.g., 2 for 20 item test, 4 for 40 item test). Additionally, the maximum number of extracted clusters is set to 5.

Further, as indicated above, DETECT can be run in exploratory or cross-validated modes. Research showed that bias in the exploratory DETECT index can be substantial in conditions with fewer items and smaller sample size

(Monahan, et al., 2007; Zhang, et al., 2003), thus both exploratory and cross-validated DETECT index are included in this study. For cross-validated DETECT index, the training sample calculation is obtained by setting a 50%/50% split in each sample, dictating DETECT software to randomly select 50% of the examinees to belong to the training sample, and the remaining 50% to serve as the validation subsample for each condition (Monahan, et al., 2007).

Table 2.

Item Parameters for 2D Compensatory MIRT Model for 10 Items per Dimension for all Types of Structures

		Exact Simple Structure		10% Complex Structure		30% Complex Structure		50% Complex Structure	
Item	*d*	a_1	a_2	a_1	a_2	a_1	a_2	a_1	a_2
1	-1.50	0.60	0.00	0.60	0.00	0.60	0.00	0.60	0.80
2	-0.75	0.60	0.00	0.60	0.00	0.60	0.00	0.60	0.00
3	0.00	0.90	0.00	0.90	1.10	0.90	1.10	0.90	1.10
4	0.75	0.90	0.00	0.90	0.00	0.90	0.00	0.90	0.00
5	1.50	1.20	0.00	1.20	0.00	1.20	1.00	1.20	1.00
6	-1.50	1.20	0.00	1.20	0.00	1.20	0.00	1.20	0.00
7	-0.75	1.50	0.00	1.50	0.00	1.50	0.00	1.50	1.30
8	0.00	1.50	0.00	1.50	0.00	1.50	0.00	1.50	0.00
9	0.75	1.80	0.00	1.80	0.00	1.80	1.60	1.80	1.60
10	1.50	1.80	0.00	1.80	0.00	1.80	0.00	1.80	0.00
11	1.50	0.00	0.60	0.00	0.60	0.00	0.60	0.00	0.60
12	0.75	0.00	0.60	0.00	0.60	0.80	0.60	0.80	0.60
13	0.00	0.00	0.90	0.00	0.90	0.00	0.90	0.00	0.90
14	-0.75	0.00	0.90	0.00	0.90	0.00	0.90	1.10	0.90
15	-1.50	0.00	1.20	0.00	1.20	0.00	1.20	0.00	1.20
16	1.50	0.00	1.20	1.00	1.20	1.00	1.20	1.00	1.20
17	0.75	0.00	1.50	0.00	1.50	0.00	1.50	0.00	1.50
18	0.00	0.00	1.50	0.00	1.50	1.30	1.50	1.30	1.50
19	-0.75	0.00	1.80	0.00	1.80	0.00	1.80	0.00	1.80
20	-1.50	0.00	1.80	0.00	1.80	0.00	1.80	1.60	1.80
M	0.00	0.60	0.60	0.65	0.66	0.76	0.79	0.89	0.89
SD	1.09	0.69	0.69	0.68	0.68	0.65	0.67	0.63	0.63

Table 3.

Item Parameters for 3D Compensatory MIRT Model for 10 Items per Dimension for Exact Simple and 10% Complex Structures

		Exact Simple Structure			10% Complex Structure		
Item	*d*	a_1	a_2	a_3	a_1	a_2	a_3
1	-1.50	0.60	0.00	0.00	0.60	0.00	0.00
2	-0.75	0.60	0.00	0.00	0.60	0.00	0.00
3	0.00	0.90	0.00	0.00	0.90	1.10	1.30
4	0.75	0.90	0.00	0.00	0.90	0.00	0.00
5	1.50	1.20	0.00	0.00	1.20	0.00	0.00
6	-1.50	1.20	0.00	0.00	1.20	0.00	0.00
7	-0.75	1.50	0.00	0.00	1.50	0.00	0.00
8	0.00	1.50	0.00	0.00	1.50	0.00	0.00
9	0.75	1.80	0.00	0.00	1.80	0.00	0.00
10	1.50	1.80	0.00	0.00	1.80	0.00	0.00
11	1.50	0.00	0.60	0.00	0.00	0.60	0.00
12	0.75	0.00	0.60	0.00	0.00	0.60	0.00
13	0.00	0.00	0.90	0.00	0.00	0.90	0.00
14	-0.75	0.00	0.90	0.00	0.00	0.90	0.00
15	-1.50	0.00	1.20	0.00	0.00	1.20	0.00
16	1.50	0.00	1.20	0.00	1.00	1.20	1.40
17	0.75	0.00	1.50	0.00	0.00	1.50	0.00
18	0.00	0.00	1.50	0.00	0.00	1.50	0.00
19	-0.75	0.00	1.80	0.00	0.00	1.80	0.00
20	-1.50	0.00	1.80	0.00	0.00	1.80	0.00
21	-1.50	0.00	0.00	0.60	0.00	0.00	0.60
22	-0.75	0.00	0.00	0.60	0.00	0.00	0.60
23	0.00	0.00	0.00	0.90	0.00	0.00	0.90
24	0.75	0.00	0.00	0.90	0.00	0.00	0.90
25	1.50	0.00	0.00	1.20	0.00	0.00	1.20
26	-1.50	0.00	0.00	1.20	0.00	0.00	1.20
27	-0.75	0.00	0.00	1.50	1.10	1.30	1.50
28	0.00	0.00	0.00	1.50	0.00	0.00	1.50
29	0.75	0.00	0.00	1.80	0.00	0.00	1.80
30	1.50	0.00	0.00	1.80	0.00	0.00	1.80
M	0.00	.40	.40	.40	.47	.48	.49
SD	1.08	.63	.63	.63	.64	.65	.66

Table 4.

Item Parameters for 3D Compensatory MIRT Model for 10 Items per Dimension for 30% and 50% Complex Structures

		30% Complex Structure			50% Complex Structure		
Item	*d*	a_1	a_2	a_3	a_1	a_2	a_3
1	-1.50	0.60	0.00	0.00	0.60	0.80	1.00
2	-0.75	0.60	0.00	0.00	0.60	0.00	0.00
3	0.00	0.90	1.10	1.30	0.90	1.10	1.30
4	0.75	0.90	0.00	0.00	0.90	0.00	0.00
5	1.50	1.20	1.00	0.80	1.20	1.00	0.80
6	-1.50	1.20	0.00	0.00	1.20	0.00	0.00
7	-0.75	1.50	0.00	0.00	1.50	1.30	1.10
8	0.00	1.50	0.00	0.00	1.50	0.00	0.00
9	0.75	1.80	1.60	1.40	1.80	1.60	1.40
10	1.50	1.80	0.00	0.00	1.80	0.00	0.00
11	1.50	0.00	0.60	0.00	0.00	0.60	0.00
12	0.75	1.00	0.60	0.80	1.00	0.60	0.80
13	0.00	0.00	0.90	0.00	0.00	0.90	0.00
14	-0.75	0.00	0.90	0.00	0.70	0.90	1.10
15	-1.50	0.00	1.20	0.00	0.00	1.20	0.00
16	1.50	1.00	1.20	1.40	1.00	1.20	1.40
17	0.75	0.00	1.50	0.00	0.00	1.50	0.00
18	0.00	1.30	1.50	1.10	1.30	1.50	1.10
19	-0.75	0.00	1.80	0.00	0.00	1.80	0.00
20	-1.50	0.00	1.80	0.00	1.60	1.80	1.40
21	-1.50	1.00	0.80	0.60	1.00	0.80	0.60
22	-0.75	0.00	0.00	0.60	0.00	0.00	0.60
23	0.00	1.10	1.30	0.90	1.10	1.30	0.90
24	0.75	0.00	0.00	0.90	0.00	0.00	0.90
25	1.50	0.00	0.00	1.20	1.00	0.80	1.20
26	-1.50	0.00	0.00	1.20	0.00	0.00	1.20
27	-0.75	1.10	1.30	1.50	1.10	1.30	1.50
28	0.00	0.00	0.00	1.50	0.00	0.00	1.50
29	0.75	0.00	0.00	1.80	1.40	1.60	1.80
30	1.50	0.00	0.00	1.80	0.00	0.00	1.80
M	0.00	0.67	0.64	0.63	0.77	0.79	0.78
SD	1.08	0.65	0.67	0.66	0.63	0.64	0.63

Table 5.

Item Parameters for 2D Noncompensatory MIRT Model for 10 Items per Dimension for all Types of Structures

			Exact Simple Structure		10% Complex Structure		30% Complex Structure		50% Complex Structure	
Item	b_1	b_2	a_1	a_2	a_1	a_2	a_1	a_2	a_1	a_2
1	-1.50	-1.00	0.60	0.00	0.60	0.00	0.60	0.00	0.60	0.80
2	-1.00	-1.00	0.60	0.00	0.60	0.00	0.60	0.00	0.60	0.00
3	0.00	-0.50	0.90	0.00	0.90	1.10	0.90	1.10	0.90	1.10
4	1.00	-0.50	0.90	0.00	0.90	0.00	0.90	0.00	0.90	0.00
5	1.50	0.00	1.20	0.00	1.20	0.00	1.20	1.00	1.20	1.00
6	-1.50	0.00	1.20	0.00	1.20	0.00	1.20	0.00	1.20	0.00
7	-1.00	0.50	1.50	0.00	1.50	0.00	1.50	0.00	1.50	1.30
8	0.00	0.50	1.50	0.00	1.50	0.00	1.50	0.00	1.50	0.00
9	1.00	1.00	1.80	0.00	1.80	0.00	1.80	1.60	1.80	1.60
10	1.50	1.00	1.80	0.00	1.80	0.00	1.80	0.00	1.80	0.00
11	-1.50	-1.00	0.00	0.60	0.00	0.60	0.00	0.60	0.00	0.60
12	-1.00	-1.00	0.00	0.60	0.00	0.60	0.80	0.60	0.80	0.60
13	0.00	-0.50	0.00	0.90	0.00	0.90	0.00	0.90	0.00	0.90
14	1.00	-0.50	0.00	0.90	0.00	0.90	0.00	0.90	1.10	0.90
15	1.50	0.00	0.00	1.20	0.00	1.20	0.00	1.20	0.00	1.20
16	-1.50	0.00	0.00	1.20	1.00	1.20	1.00	1.20	1.00	1.20
17	-1.00	0.50	0.00	1.50	0.00	1.50	0.00	1.50	0.00	1.50
18	0.00	0.50	0.00	1.50	0.00	1.50	1.30	1.50	1.30	1.50
19	1.00	1.00	0.00	1.80	0.00	1.80	0.00	1.80	0.00	1.80
20	1.50	1.00	0.00	1.80	0.00	1.80	0.00	1.80	1.60	1.80
M	0.00	0.00	.60	.60	.65	.66	0.76	0.79	0.89	0.89
SD	1.17	0.73	.69	.69	.68	.68	0.65	0.67	0.63	0.63

Table 6.

Item Parameters for 3D Noncompensatory MIRT Model for 10 Items per Dimension for Exact Simple and 10% Complex Structures

				Exact Simple Structure			10% Complex Structure		
Item	b_1	b_2	b_3	a_1	a_2	a_3	a_1	a_2	a_3
1	-1.50	-1.00	1.20	0.60	0.00	0.00	0.60	0.00	0.00
2	-1.00	-1.00	0.70	0.60	0.00	0.00	0.60	0.00	0.00
3	0.00	-0.50	0.00	0.90	0.00	0.00	0.90	1.10	1.30
4	1.00	-0.50	-0.70	0.90	0.00	0.00	0.90	0.00	0.00
5	1.50	0.00	-1.20	1.20	0.00	0.00	1.20	0.00	0.00
6	-1.50	0.00	-1.20	1.20	0.00	0.00	1.20	0.00	0.00
7	-1.00	0.50	-0.70	1.50	0.00	0.00	1.50	0.00	0.00
8	0.00	0.50	0.00	1.50	0.00	0.00	1.50	0.00	0.00
9	1.00	1.00	0.70	1.80	0.00	0.00	1.80	0.00	0.00
10	1.50	1.00	1.20	1.80	0.00	0.00	1.80	0.00	0.00
11	-1.50	-1.00	1.20	0.00	0.60	0.00	0.00	0.60	0.00
12	-1.00	-1.00	0.70	0.00	0.60	0.00	0.00	0.60	0.00
13	0.00	-0.50	0.00	0.00	0.90	0.00	0.00	0.90	0.00
14	1.00	-0.50	-0.70	0.00	0.90	0.00	0.00	0.90	0.00
15	1.50	0.00	-1.20	0.00	1.20	0.00	0.00	1.20	0.00
16	-1.50	0.00	-1.20	0.00	1.20	0.00	1.00	1.20	1.40
17	-1.00	0.50	-0.70	0.00	1.50	0.00	0.00	1.50	0.00
18	0.00	0.50	0.00	0.00	1.50	0.00	0.00	1.50	0.00
19	1.00	1.00	0.70	0.00	1.80	0.00	0.00	1.80	0.00
20	1.50	1.00	1.20	0.00	1.80	0.00	0.00	1.80	0.00
21	-1.50	-1.00	1.20	0.00	0.00	0.60	0.00	0.00	0.60
22	-1.00	-1.00	0.70	0.00	0.00	0.60	0.00	0.00	0.60
23	0.00	-0.50	0.00	0.00	0.00	0.90	0.00	0.00	0.90
24	1.00	-0.50	-0.70	0.00	0.00	0.90	0.00	0.00	0.90
25	1.50	0.00	-1.20	0.00	0.00	1.20	0.00	0.00	1.20
26	-1.50	0.00	-1.20	0.00	0.00	1.20	0.00	0.00	1.20
27	-1.00	0.50	-0.70	0.00	0.00	1.50	1.10	1.30	1.50
28	0.00	0.50	0.00	0.00	0.00	1.50	0.00	0.00	1.50
29	1.00	1.00	0.70	0.00	0.00	1.80	0.00	0.00	1.80
30	1.50	1.00	1.20	0.00	0.00	1.80	0.00	0.00	1.80
M	0.37	0.27	-0.10	.40	.40	.40	.47	.48	.49
SD	1.04	0.56	0.85	.63	.63	.63	.64	.65	.66

Table 7.

Item Parameters for 3D Noncompensatory MIRT Model for 10 Items per Dimension for 30% and 50% Complex Structures

				30% Complex Structure			50% Complex Structure		
Item	b_1	b_2	b_3	a_1	a_2	a_3	a_1	a_2	a_3
1	-1.50	-1.00	1.20	0.60	0.00	0.00	0.60	0.80	1.00
2	-1.00	-1.00	0.70	0.60	0.00	0.00	0.60	0.00	0.00
3	0.00	-0.50	0.00	0.90	1.10	1.30	0.90	1.10	1.30
4	1.00	-0.50	-0.70	0.90	0.00	0.00	0.90	0.00	0.00
5	1.50	0.00	-1.20	1.20	1.00	0.80	1.20	1.00	0.80
6	-1.50	0.00	-1.20	1.20	0.00	0.00	1.20	0.00	0.00
7	-1.00	0.50	-0.70	1.50	0.00	0.00	1.50	1.30	1.10
8	0.00	0.50	0.00	1.50	0.00	0.00	1.50	0.00	0.00
9	1.00	1.00	0.70	1.80	1.60	1.40	1.80	1.60	1.40
10	1.50	1.00	1.20	1.80	0.00	0.00	1.80	0.00	0.00
11	-1.50	-1.00	1.20	0.00	0.60	0.00	0.00	0.60	0.00
12	-1.00	-1.00	0.70	1.00	0.60	0.80	1.00	0.60	0.80
13	0.00	-0.50	0.00	0.00	0.90	0.00	0.00	0.90	0.00
14	1.00	-0.50	-0.70	0.00	0.90	0.00	0.70	0.90	1.10
15	1.50	0.00	-1.20	0.00	1.20	0.00	0.00	1.20	0.00
16	-1.50	0.00	-1.20	1.00	1.20	1.40	1.00	1.20	1.40
17	-1.00	0.50	-0.70	0.00	1.50	0.00	0.00	1.50	0.00
18	0.00	0.50	0.00	1.30	1.50	1.10	1.30	1.50	1.10
19	1.00	1.00	0.70	0.00	1.80	0.00	0.00	1.80	0.00
20	1.50	1.00	1.20	0.00	1.80	0.00	1.60	1.80	1.40
21	-1.50	-1.00	1.20	1.00	0.80	0.60	1.00	0.80	0.60
22	-1.00	-1.00	0.70	0.00	0.00	0.60	0.00	0.00	0.60
23	0.00	-0.50	0.00	1.10	1.30	0.90	1.10	1.30	0.90
24	1.00	-0.50	-0.70	0.00	0.00	0.90	0.00	0.00	0.90
25	1.50	0.00	-1.20	0.00	0.00	1.20	1.00	0.80	1.20
26	-1.50	0.00	-1.20	0.00	0.00	1.20	0.00	0.00	1.20
27	-1.00	0.50	-0.70	1.10	1.30	1.50	1.10	1.30	1.50
28	0.00	0.50	0.00	0.00	0.00	1.50	0.00	0.00	1.50
29	1.00	1.00	0.70	0.00	0.00	1.80	1.40	1.60	1.80
30	1.50	1.00	1.20	0.00	0.00	1.80	0.00	0.00	1.80
M	0.37	0.27	-0.10	0.62	0.64	0.63	0.77	0.79	0.78
SD	1.04	0.56	0.85	0.64	0.67	0.66	0.63	0.64	0.63

For NOHARM, default options are utilized for model identification, i.e., factor variances are fixed to 1, and exploratory solutions are examined for 1-, 2-, 3-, 4-, and 5-factors. The choice to model the 5-factor solution as the highest in exploratory NOHARM is such that it corresponds to DETECT's allowance of maximum of 5 clusters extraction. Additionally Promax methods are used to obtain oblique transformations that are used in analysis.

Outcome Variables

Following the literature on DETECT and NOHARM, several variables are included in the current study to evaluate the performance of these methods (e.g., Finch & Habing, 2005, 2007; Gierl, et al., 2006; Monahan, et al., 2007; Tate, 2003). Three main outcome variables reported in this study include: a) the proportion of correct selection of true dimensional structure, b) the ability to label sets of items as representing the true dimensions (dimension-like), and c) the classification consistency of items. As discussed next, these outcomes are operationalized somewhat differently for the different procedures. The final reported values for a condition are averaged across 500 successfully run replications.[4]

[4] Possible convergence issues may be encountered while fitting models in NOHARM. In conditions with nonconvergence of replications, additional replications are run to arrive to a total of 500 successfully estimated replications per condition.

The proportion of correct selection of true dimensionality. The first outcome variable is operationalized as the proportion of times within each condition that a true dimensional space is found (i.e., 2 factors in conditions where data are generated using a 2D MIRT, and 3 factors in conditions where data follow a 3D MIRT). In DETECT, this is a straight forward procedure because DETECT outputs non-overlapping clusters, hence the number of dimensions found equals the number of clusters DETECT outputs. For purposes of this study, clusters that contain 3 items or less are still considered, although they might be considered nuisance dimensions (e.g., Zhang & Stout, 1999b). Furthermore, in reporting results, for consistency in the language used, when referring to a group of items that are associated together in a cluster, the term 'factor' is used (although typically in DETECT we often refer to these groups of items as clusters).

In NOHARM, three procedures are used to determine the optimal number of factors. Each of these procedures is performed and reported separately. The first procedure is based on the NOHARM output that yields the root mean square residual (RMSR). Here, based on Tate (2003), a sequential model fitting approach to determining the number of factors is adopted. This approach suggests that models are fitted with additional factors until the change in RMSR does not exceed 10%. For example, if RMSR for a model with a single (2-, 3-, and 4-) factor(s) is .00631 (.00512, .00457, and .00422), the resulting decreases in RMSRs from a single factor solution to the second, third, and fourth dimensional

solutions are 19%, 11%, and 8%, respectively. Following the recommended rule of 10% decrease, the result of adding the fourth factor (from 3 to 4) results in decrease of 8% in the RMSR, thus the conclusion is to retain a 3D solution.

The second and third procedures used to determine the formal fit of the model and retain the optimal number of factors are based on $\chi^2_{G/D}$ and *ALR*, respectively. Here, similar to a traditional factor analytic approaches to determining the number of dimensions using a χ^2 test for the difference in test statistics. This means that a researcher starts with the fewer dimensional model and asks whether a higher dimensional model is needed based on the difference test. If the higher dimensional model provides a better fit (i.e., $p < .05$ of the difference test), the procedure continues. The optimal factor solution is found when the higher dimensional model does not improve the fit significantly (i.e., $p > .05$).

The ability to label sets of items as “dimension-like”. This outcome variable puts emphasis on answering the question of how many of sets of items could be labeled as dimension-like. In other words, once either of the methods groups a set of items together in a set, the question remains as to how often could that set of items be labeled as a dimension-like, meaning that they could be interpreted as adequately representing one of the true underlying dimensions. Prior to answering this question, items have to be grouped in some way. In DETECT, sets of items are determined and grouped automatically, as the procedure outputs non-overlapping clusters. Therefore, sets of items (clusters) are

determined by the procedure, and those sets of items that are then submitted to criteria for labeling sets of items as dimension-like.

In NOHARM, prior to investigating how often a group of items be labeled as dimension-like, items have to be grouped. In order to group items together, the following criteria are applied to the rotated factor solution from NOHARM. For an item to be grouped with a factor, the item must have an estimated loading > .40 on that particular factor and the difference between that loading and all other loadings must be > .20. If the item has an estimated loading that is > .40 and the difference between its largest loading and at least one other loading is < .20, the item is grouped separately in a group that is interpreted as complex (note this complexity is with respect to the fitted factor model, which will not necessarily correspond to whether the item is truly a factorially complex item). Alternatively, if an item does not meet either criteria (i.e., its loadings are < .40 on all factors), the item is considered to be unexplained.

For example, let us assume we have a condition that is originally generated as a true 2D condition with 10 items associated with each dimension. This condition therefore has 20 items in total. If a method based on NOHARM output determines an optimal factor solution to be 4 factors, a rotated factor loadings matrix from NOHARM output is obtained. This loading matrix is 20 (items) by 4 (factor-solution) in size. Each items for each factor is then submitted to criteria in order to determine with which factor an item is mostly associated. In order for an item to be put in a set associated with factor one, for example, the

item's estimated loading has to be > .40 on factor one, and it has to be larger than its loading on factors two, three, and four (where those loadings are all ≤ .40). The difference of estimated loading between factor one and each of the remaining factors has to be larger than .20. If an item meets these two criteria, that item is then put in a group that belongs to factor one. Alternatively, if the item meets the criterion of having an estimated loading > .40 on multiple factors or the difference between its loading on that factor and at least one other factor is < .20, the item is grouped in a complex set. Alternatively, if the item does not meet either criteria (i.e., its loadings are < .20 on all factors), the item is considered to be unexplained.

After all items are grouped, the labeling of these “item groups” or “item sets” as dimension-like begin. A set of items can be labeled as dimension-1-like set of items, dimension-2-like set of items, or dimension-3-like set of items, depending on what is the true dimensionality of the data. Additionally, each item is generated originally as factorially simple or factorially complex (see Tables 2 through 7 for item parameters used in data generation). In order for a set of items to be called dimension-1-like set, it ought to meet the following criteria. First, at least 50% of items in the set must be items that were generated as factorially simple and reflecting (the true) dimension-1. Second, dimension-1 factorially simple items ought to occupy more than half of the set of items. If both of these criteria are met, then that set of items is labeled as dimension-1-like, and all items that belong to the set in question are considered as dimension-1-like items.

Classification consistency rates of items. In order to examine consistency of the methods, c*lassification consistency* is computed by taking each item's classification (across 500 replications in each condition) and taking the proportion of times that the true classification is obtained.

For example, each item is given a classification assignment. First, the item is tracked to see which set of items it is grouped with (based on the labeling criteria discussed above). If the item is grouped in a set of items that are labeled as dimension-1-like (e.g., items in that group are mostly designated as dimension 1 items), all of the items in that set are assigned a classification of D1. Classification rates are computed for each item by taking the mean of the correct classification assignment over the 500 replications. In reporting classification rates, items of the same type (e.g., all factorially simple or all factorially complex) are pooled.

Chapter 4

DATA ANALYSIS AND RESULTS

In Chapter 4, results of the current study are reported. Nonconvergence issues are discussed at the beginning. Then, results are presented for conditions when methods selected a one-factor solution as being optimal. Results concerning the three main outcome variables are discussed next. Results for the number of factors extracted by each method are presented, followed by the marginal proportions of the methods' ability to label a set of items associated with a factor or cluster as a dimension-like, given the pre-specified criteria. Finally, the consistency of the methods in classifying factorially simple and factorially complex items is examined via classification rates. Given the symmetry of the study's design, in order to compute consistency rates for different types of items, items of the same type are pooled. Also, for the purposes of this study, when referring to a factor solution or a factor model, it is in reference to what the particular method yielded as an optimal or favorable solution.

For clarity of presentation, the results are presented separately for compensatory and noncompensatory MIRT data, for different tests lengths of 10 and 20 items per dimension, and for two- dimensional, 2D, and three-dimensional, 3D, structures. Useful comparisons are made when appropriate throughout the results. Lastly, the effects of the number of items per dimension, used to organize most of the presentation were summarized.

Nonconvergence in NOHARM

As stated in Chapter 3, NOHARM uses least squares estimation to arrive to the optimal estimates of item parameters. Recall that for each condition, when fitting exploratory models in NOHARM, a total of 2,500 replications were submitted to NOHARM for parameter estimation (i.e., 500 replications for fitting one-, two-, three-, four-, and five-factors). Additionally, two different levels of test lengths were considered. This resulted in a total of 480 conditions, 240 of which included 10 items per dimension and 240 of which included 20 items per dimension.

In this study, 215 conditions encountered some degree of nonconvergence. The number of nonconvergent replications within a condition ranged from one to 461. Over 90% of the conditions with failed convergence included cases with 20 items per dimension.

Nonconvergence issues were observed in several different ways. First, nonconvergent issues were found in cases with 10 items per dimension. Here, within a condition, replications that failed to converge appeared to be tied to the specific dataset. That is, if a particular replication did not converge for fitting a one-factor solution, then that same replication failed to converge for fitting subsequent two-, three-, four-, and five-factor models. If only one such instance occurred in a condition, a total of five nonconvergent runs would be counted (i.e., one for each of the five fitted models for that replication).

Second, for conditions with 20 items per dimension, nonconvergent replications occurred mostly when fitting four- or five-factor models, although for a few replications, fitting one-, two, and three-factor models also appeared problematic. Third, problems occurred in estimation where there was a "perfect" response vector for an item in a dataset (e.g., an item was answered incorrectly by all simulees); this occurred only in conditions with 20 items per dimension and noncompensatory data-generating structures. The next sections describe the degree of nonconvergence problems as well as how each issue was resolved.

Nonconvergence of datasets with ten items per dimension. Nonconvergence that occurred for all factor models fit to a particular dataset of appeared in 21out of 240 conditions, where the number of nonconvergent replications varied in size from one dataset (5 total replications equaling 0.2% of total replications in that condition) up to 19 datasets (95 total replications equaling 3.8% of total replications in that condition). Nineteen out of 21 nonconvergent conditions were conditions with $N = 500$ and 3D structures, with various correlation levels and complexity.

In Figure 4, the total numbers of attempts needed to achieve 500 convergent replications for 15 out of these 21 conditions are plotted. These conditions are all $N = 500$ and included three levels of complexity (0%, 10%, and 30%) and five levels of correlations between dimensions (.00 through .90). Note that similar number of attempts to achieve successful 500 runs was required for any one of these conditions.

Six additional conditions (not plotted) with convergence issues included two conditions of 50% complexity with .30 and .90 correlation; two conditions with N = 1000 and correlation of .60, with 0% and 10% complexity levels, and two conditions of 10% complexity with .30 and .75 correlation for 2D structures. For any of these six conditions, only one extra replication was needed to achieve 500 successful replications.

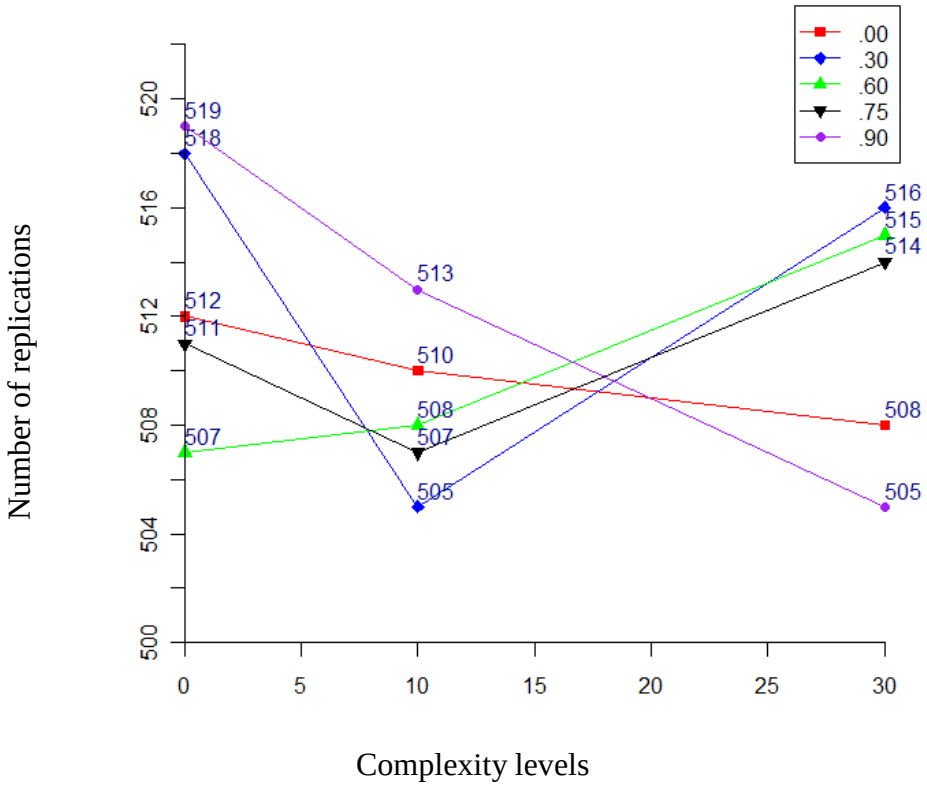

Figure 4. Summary of nonconvergent conditions with various complexity level and correlations among dimensions. Test length (10 items per dimension), small sample size (N = 500), and dimensional structure (3D) were held constant in the plotted conditions. Numbers associated with each data point represent the total number of attempts to achieve successful 500 replications. Colored lines represent various levels of correlations among dimensions.

For each nonconvergent replication, a new dataset with the same characteristics (as defined by the condition) was generated. As Figure 4 displays, the reruns using the additional replications were largely successful, such that vast majority of nonconvergent conditions required only a single additional replication. One exception to that was a condition with N = 500, 0% complexity, correlations of .90, and 3D structure, where in order to achieve 500 successful replications, two additional replications were required. Note that these newly created datasets used for reanalysis in NOHARM were then used in reanalysis in DETECT.

Nonconvergence related to the fitted model with twenty items per dimension. NOHARM failed to successfully converge in 194 out of 240 conditions in conditions with 20 items per dimension. The number of nonconvergent replications varied within conditions. Nonconvergence occurred primarily in replications when fitting a four- or a five-factor solution. Thus, in some cases, replications that converged while fitting a one- or two-factor model, failed to converge in fitting higher-dimension models. However, there were instances when fitting a one-, two-, or three-factor model that also resulted in nonconvergence.

The number of nonconverging replications in those 194 conditions is plotted as a histogram in Figure 5. Out of 194 conditions, many conditions had fewer than 50 nonconvergent replications (interquartile range equaled 3.00 to 40.75). There were, however, several conditions that had large numbers of

unsuccessful replications. The range of nonconvergent replications across these 194 conditions was 1 to 461, with a mean (standard deviation) of 44.05 (83.92), and a median of 11.50.

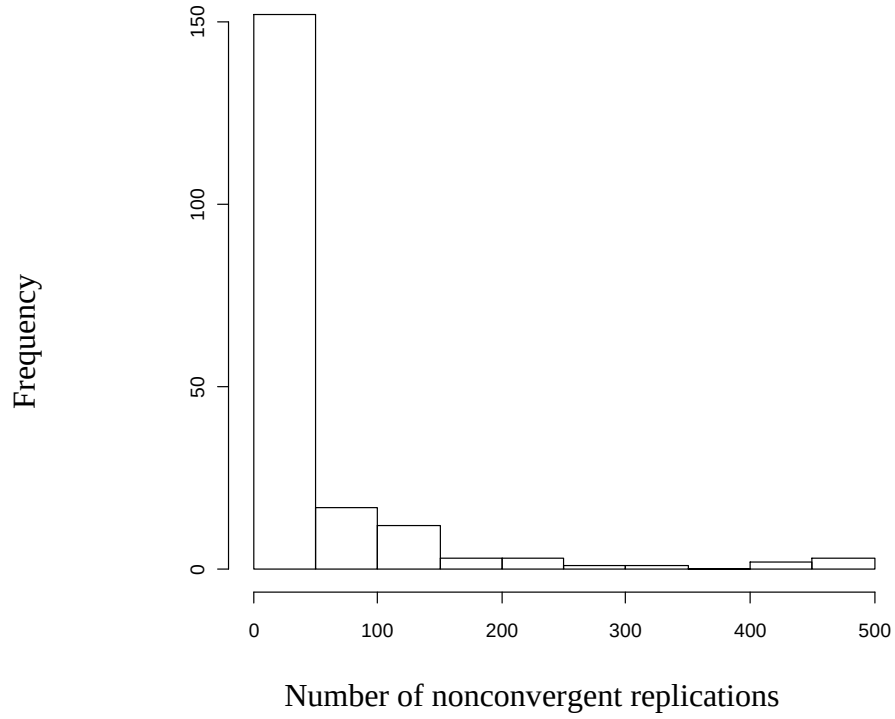

Figure 5. Histogram of the nonconvergent replications in 194 conditions with longer test lengths (note that these are out of 2,500 runs due to fitting 500 replications to five exploratory NOHARM models).

The convergence issues in these conditions were dealt in the following manner. First, the nonconvergent replications within a condition for any of the five models were identified. Second, default options in NOHARM were changed such that maximum function cell was increased and the convergence criterion was

decreased.[5] Lastly, the nonconvergent replications were rerun in NOHARM such that a successful 500 runs for any one model within a condition were achieved.

Nonconvergence related to presence of perfect items. In addition to the nonconvergent replications discussed above, a total of 355 datasets across 30 different conditions had one or more replications that contained at least one perfect item. All of these instances occurred in conditions with 20 items per dimension where data followed a noncompensatory MIRT model. Five of 30 conditions were 2D conditions; the remaining 25 were 3D conditions. All 2D and most of the 3D (20 out of 25) conditions with perfect items were conditions with $N = 500$. The remaining five 3D conditions had $N = 1000$.

Conditions in which problems with estimation due to perfect item(s) varied across complexity and correlation levels. In any one of the 30 conditions, the number of replications with perfect items varied from one to 33 replications (mean number of replications with perfect item equaled 11.83, with standard deviation of 11.46). This type of convergence issue was corrected by removing

[5] The NOHARM user's guide (Fraser & McDonald, 2003) recommends that in cases where nonconvergence is an issue, a user should change the default options for the maximum function cell and/or the criterion value. In this study, the number of maximum function cells was increased from default 2000 to 4000 and criterion was decreased from .000001 to .0001. This solved the convergence issues encountered in this study.

the perfect item(s) from the dataset, and refitting the exploratory models in NOHARM.[6]

Unidimensional Solution in NOHARM

The current study focused on examining the performance of methods in the presence of factorial complexity in multidimensional data. Heuristic and statistical methods based on the NOHARM output (RMSR, $\chi^2_{G/D}$, and the *ALR*,) resulted in favoring a single-factor solution in some replications for several conditions. DETECT analyses in exploratory or cross-validated modes never resulted in a single factor solution. Therefore, results and discussion of single-factor solutions concern only the methods based on NOHARM. In the text below, only general trends in selected conditions are highlighted. In particular, conditions where the methods tended to favor unidimensional solution frequently are discussed. Tables 8 through 15 report proportions of replications (out of 500) that selected unidimensional solution for each method across the studied conditions.

Compensatory multidimensional data.

Tests with ten items per dimension. The proportions that the methods yielded unidimensional solutions for conditions with ten items associated per dimension were investigated for 2D and 3D. For each dimensional structure, there were a total of 60 conditions.

[6] Conditions in which a perfect item was removed were not then rerun in DETECT.

2D structures. A complete tabulation of proportions for 2D compensatory conditions with 10 items per dimension is shown in Table 8. RMSR only selected a unidimensional solution in two conditions, where the complexity levels were 30% and correlations of .90, for N = 500 and N = 1000. $\chi^2_{G/D}$ and *ALR* performed similarly to each other, yielding unidimensional solutions to one or more replications in 11 conditions and 17 conditions (respectively). Most of these replications appeared in conditions where complexity levels were 30% or 50%, and the correlation between dimensions was 90. Additionally, most of the one-factor solution appeared in conditions with N = 500 and N = 1000. The highest proportions of replications within a condition that favored one-factor solution by $\chi^2_{G/D}$and *ALR* were .89 (444 out of 500) and .49 (245 out of 500), respectively. Both of these high proportions were found in a condition with 30% complexity and N = 500.

Table 8.

Proportion of Replications Across Complexity Levels for Conditions with Two-dimensional Compensatory MIRT and 10 Items per Dimension

		Method											
		ALR				$\chi^2_{G/D}$				RMSR			
Complex		0%	10%	30%	50%	0%	10%	30%	50%	0%	10%	30%	50%
N	ρ												
	.00	-	-	-	-	-	-	-	-	-	-	-	-
	.30	-	-	**	-	-	-	-	-	-	-	-	-
500	.60	-	-	**	**	-	-	-	-	-	-	-	-
	.75	-	-	.02	.05	-	-	**	.09	-	-	-	-
	.90	.11	.21	.49	.20	.27	.47	.89	.55	-	-	**	-
	.00	-	-	-	-	-	-	-	-	-	-	-	-
	.30	-	-	-	-	-	-	-	-	-	-	-	-
1000	.60	-	-	-	-	-	-	-	-	-	-	-	-
	.75	-	-	-	**	-	-	-	-	-	-	-	-
	.90	**	.01	.23	.06	**	.05	.59	.01	-	-	**	-
	.00	-	-	-	-	-	-	-	-	-	-	-	-
	.30	-	-	-	-	-	-	-	-	-	-	-	-
2000	.60	-	-	-	-	-	-	-	-	-	-	-	-
	.75	-	-	-	-	-	-	-	-	-	-	-	-
	.90	**	-	.02	**	-	-	.03	-	-	-	-	-

Note: Each condition has a total of 500 replications. "-" sign indicates that zero replications in a condition selected unidimensional solution. "**" sign indicates that less than 1% of replications in a condition selected unidimensional solution.

3D structures. A complete tabulation of proportions for 3D compensatory conditions with 10 items per dimension is shown in Table 9. *ALR* favored a one-factor solution in at least one replication for 36 out of 60 conditions. The unidimensional solutions were particularly favored as optimal in conditions with 30% or 50% of complexity, for different sample sizes and correlation values. The other two methods, $\chi^2_{G/D}$and RMSR, favored one-factor solutions to a lesser extent. For both methods, a one-factor solution was selected for at least one replication in 9 conditions, only. These 9 conditions came primarily in cases when correlations were at .90, for a variety of the sample sizes and complexity levels.

For $\chi^2_{G/D}$, the largest number of replication within a condition that favored a one-factor solution occurred in the condition with 30% complexity, *N* = 500, and .90 correlations (385 out of 500 replications). The maximum number of replications within a condition for RMSR was 268 (out of 500), in the condition with 50% complexity, *N* = 500, and correlations of .90.

Table 9.

Proportion of Replications Across Complexity Levels for Conditions with Three-dimensional Compensatory MIRT and 10 Items per Dimension

		Method											
		ALR				$\chi^2_{G/D}$				RMSR			
Complex		0%	10%	30%	50%	0%	10%	30%	50%	0%	10%	30%	50%
N	ρ												
	.00	.02	-	**	**	-	-	-	-	-	-	-	-
	.30	-	-	.10	.09	-	-	-	-	-	-	-	-
500	.60	-	-	.29	.37	-	-	-	**	-	-	-	**
	.75	-	-	.28	.14	-	-	-	-	-	-	**	-
	.90	.05	.15	.22	.10	.12	.33	.77	.03	.03	.05	.54	**
	.00	.02	-	-	**	-	-	-	-	-	-	-	-
	.30	-	-	.06	.03	-	-	-	-	-	-	-	-
1000	.60	-	-	.34	.36	-	-	-	-	-	-	-	-
	.75	-	-	.29	.13	-	-	-	-	-	-	-	-
	.90	**	.02	.09	.05	-	**	.50	-	-	-	.50	-
	.00	**	-	**	-	-	-	-	-	-	-	-	-
	.30	-	-	.01	**	-	-	-	-	-	-	-	-
2000	.60	-	-	.42	.42	-	-	-	-	-	-	-	-
	.75	-	-	.30	.15	-	-	-	**	-	-	-	**
	.90	**	-	.06	.03	-	-	.10	-	-	-	.34	-

Note: Each condition has a total of 500 replications. "-" sign indicates that zero replications in a condition selected unidimensional solution. "**" sign indicates that less than 1% of replications in a condition selected unidimensional solution.

Tests with twenty items per dimension. The frequencies that the methods yielded unidimensional solutions for conditions with twenty items associated per dimension were investigated for 2D and 3D. For each dimensional structure, there were a total of 60 conditions.

2D structures. A complete tabulation of proportions for 2D compensatory conditions with 20 items per dimension is presented in Table 10. *ALR* yielded a unidimensional solution as preferred in 19 out of 60 conditions. Most of these 19 conditions had a correlation of .90 and various complexity levels. Three of the 19 conditions yielded a nontrivial proportion of replications that favored a unidimensional solution when the dimensions were uncorrelated and no complexity was present in the data. These conditions reported a one-factor solution in proportions of .18, .19, and .20 for $N = 500$, $N = 1000$, and $N = 2000$, respectively. $\chi^2_{G/D}$ and RMSR yielded a unidimensional solution in only one and five conditions, respectively.

Table 10.

Proportion of Replications Across Complexity Levels for Conditions with Two-dimensional Compensatory MIRT and 20 Items per Dimension

		Method											
		ALR				$\chi^2_{G/D}$				RMSR			
Complex		0%	10%	30%	50%	0%	10%	30%	50%	0%	10%	30%	50%
N	ρ												
	.00	.18	-	-	-	-	-	-	-	-	-	-	-
	.30	-	-	-	-	-	-	-	-	-	-	-	-
500	.60	-	-	**	-	-	-	-	-	-	-	-	-
	.75	-	-	.02	.13	-	-	-	-	-	-	-	-
	.90	.21	.36	.78	.34	-	-	.07	-	-	**	.20	.21
	.00	.19	-	-	-	-	-	-	-	-	-	-	-
	.30	-	-	-	-	-	-	-	-	-	-	-	-
1000	.60	-	-	-	-	-	-	-	-	-	-	-	-
	.75	-	-	**	.01	-	-	-	-	-	-	-	-
	.90	.01	.06	.43	.05	-	-	-	-	-	-	**	.01
	.00	.20	-	-	-	-	-	-	-	-	-	-	-
	.30	-	-	-	-	-	-	-	-	-	-	-	-
2000	.60	-	-	-	-	-	-	-	-	-	-	-	-
	.75	-	-	-	-	-	-	-	-	-	-	-	-
	.90	-	**	.10	**	-	-	-	-	-	-	-	-

Note: Each condition has a total of 500 replications. "-" sign indicates that zero replications in a condition selected unidimensional solution. "**" sign indicates that less than 1% of replications in a condition selected unidimensional solution.

3D structures. A complete tabulation of proportions for 3D compensatory conditions with 20 items per dimension is presented in Table 11. *ALR* yielded a unidimensional solution as preferred in 58 out of 60 conditions with 3D compensatory models with 20 items per dimension. The highest proportions of replications were found in conditions with highly correlated dimensions or in conditions where data exhibited higher complexity. $\chi^2_{G/D}$ and RMSR yielded a unidimensional solution in only two and six conditions, respectively.

Table 11.

Proportion of Replications Across Complexity Levels for Conditions with Three-dimensional Compensatory MIRT and 20 Items per Dimension

		Method											
		ALR				$\chi^2_{G/D}$				RMSR			
Complex		0%	10%	30%	50%	0%	10%	30%	50%	0%	10%	30%	50%
N	ρ												
	.00	.13	.13	.08	.06	-	-	-	-	-	-	-	-
	.30	.03	.03	.15	.16	-	-	-	-	-	-	-	-
500	.60	.01	.02	.38	.38	-	-	-	-	-	-	-	-
	.75	.03	.05	.28	.12	-	-	-	-	-	-	-	-
	.90	.35	.42	.11	.06	-	-	.06	-	.25	.46	.94	-
	.00	.14	.12	.06	.05	-	-	-	-	-	-	-	-
	.30	.01	.03	.12	.08	-	-	-	-	-	-	-	-
1000	.60	**	.01	.41	.39	-	-	-	-	-	-	-	-
	.75	.01	.02	.27	.09	-	-	-	-	-	-	-	-
	.90	.21	.32	.09	.04	-	-	.02	-	-	.01	.79	-
	.00	.12	.10	.05	.03	-	-	-	-	-	-	-	-
	.30	.01	.01	.06	.03	-	-	-	-	-	-	-	-
2000	.60	**	**	.45	.48	-	-	-	-	-	-	-	-
	.75	-	-	.34	.12	-	-	-	-	-	-	-	-
	.90	.20	.17	.06	.02	-	-	-	-	-	-	.45	-

Note: Each condition has a total of 500 replications. "-" sign indicates that zero replications in a condition selected unidimensional solution. "**" sign indicates that less than 1% of replications in a condition selected unidimensional solution.

Noncompensatory multidimensional data.

Tests with ten items per dimension. The frequencies that the methods yielded unidimensional solutions for conditions with ten items associated per dimension were investigated for 2D and 3D. For each dimensional structure, there were a total of 60 conditions.

2D structures. A complete tabulation of proportions for 2D noncompensatory conditions with 10 items per dimension is presented in Table 12. RMSR selected one-factor solution in at least one replication in only 9 out of total of 60 conditions. Within any condition, no more than six replications selected one factor. *ALR* and $\chi^2_{G/D}$ methods tended to favor unidimensional structures more often than RMSR. *ALR* selected one factor in at least one replication in 35 conditions; 32 of which were conditions with correlation of .60 or larger, and 25 of which were in conditions where N = 500 and N = 1000 (12 and 13, respectively). When N = 500, one-factor solutions were selected across all levels of complexity, although larger number of such replications within a condition increased as complexity levels reached 30%. For example, with N = 500 and correlations of .60 or larger, $\chi^2_{G/D}$ had a considerable number of replications that favored one-factor solution.

Table 12.

Proportion of Replications Across Complexity Levels for Conditions with Two-dimensional Noncompensatory MIRT and 10 Items per Dimension

		Method											
		ALR				$\chi^2_{G/D}$				RMSR			
Complex		0%	10%	30%	50%	0%	10%	30%	50%	0%	10%	30%	50%
N	ρ												
	.00	-	-	-	-	-	-	-	-	-	-	-	-
	.30	-	-	.04	**	-	-	**	**	-	-	-	-
500	.60	.02	.05	.32	.26	**	**	.29	.27	-	-	**	-
	.75	.23	.26	.45	.46	.21	.21	.48	.50	-	-	**	-
	.90	.47	.42	.40	.40	.46	.36	.40	.43	.01	**	**	**
	.00	-	-	-	-	-	-	-	-	-	-	-	-
	.30	-	-	-	-	-	-	-	-	-	-	-	-
1000	.60	-	**	.07	.03	-	-	.05	.02	-	-	-	-
	.75	.03	.04	.19	.17	.01	.01	.16	.16	-	-	-	**
	.90	.32	.14	.17	.12	.32	.09	.07	.07	**	-	-	**
	.00	-	-	-	-	-	-	-	-	-	-	-	-
	.30	-	-	**	-	-	-	-	-	-	-	-	-
2000	.60	-	-	.01	-	-	-	-	-	-	-	-	-
	.75	**	.01	.05	.02	-	-	**	**	-	-	-	-
	.90	.11	.01	.04	.01	.08	-	**	-	-	-	-	-

Note: Each condition has a total of 500 replications. "-" sign indicates that zero replications in a condition selected unidimensional solution. "**" sign indicates that less than 1% of replications in a condition selected unidimensional solution.

A similar pattern was found when *N* = 1000. Fewer conditions and replications within a condition yielded a one-factor solution as the preferred solution. $\chi^2_{G/D}$ chose one factor in at least one replication in 28 out of 60 conditions. The types of conditions as well as the number of replications within those conditions were very similar to that of *ALR*. Conditions with largest number of replications with one-factor solutions tended to be those with *N* = 500.

3D structures. A complete tabulation of proportions for 3D noncompensatory conditions with 10 items per dimension is presented in Table 13. The RMSR method found one-factor solution in conditions across all levels of complexity, particularly when *N* = 500. The condition with largest proportion of replications (50 out of 500 replications) with preferred unidimensional solutions had a complexity level of 50%, *N* = 500, and correlations of .60. *ALR* and $\chi^2_{G/D}$ selected the one-factor model as optimal more frequently than RMSR. *ALR* selected the unidimensional solution for at least one replication in 56 out of 60 total conditions. In many of these conditions, however, the number of replications was much higher than one (median of 75). Conditions with 0% of complexity and correlations of .75 and .90 across all three sample sizes contained the highest numbers of replications that *ALR* chose the one-factor solution.

Table 13.

Proportion of Replications Across Complexity Levels for Conditions with Three-dimensional Noncompensatory MIRT and 10 Items per Dimension

		Method											
		ALR				$\chi^2_{G/D}$				RMSR			
Complex		0%	10%	30%	50%	0%	10%	30%	50%	0%	10%	30%	50%
N	ρ												
	.00	.09	.06	.39	.35	**	**	.25	.09	-	-	.01	**
	.30	.18	.17	.52	.39	.06	.08	.44	.16	-	-	.03	.07
500	.60	.47	.47	.56	.40	.34	.30	.38	.13	**	**	.04	.10
	.75	.57	.47	.45	.27	.52	.29	.30	.09	.02	.01	.02	.03
	.90	.68	.41	.32	.14	.53	.17	.14	.01	.05	**	**	-
	.00	.01	**	.18	.14	-	-	.06	.02	-	-	-	**
	.30	.04	.05	.37	.25	-	**	.33	.09	-	-	**	.03
1000	.60	.27	.32	.39	.20	.21	.19	.23	.05	-	**	.01	.03
	.75	.48	.33	.25	.08	.50	.17	.09	**	**	-	**	**
	.90	.62	.16	.08	**	.63	.02	**	-	.04	-	-	-
	.00	-	-	.04	.02	-	-	**	-	-	-	-	-
	.30	**	**	.19	.10	-	-	.07	**	-	-	**	**
2000	.60	.07	.11	.13	.05	.02	.02	.04	.01	-	-	**	**
	.75	.37	.10	.04	**	.36	.02	-	-	-	**	-	-
	.90	.60	.01	-	-	.69	-	-	-	.04	-	-	-

Note: Each condition has a total of 500 replications. "-" sign indicates that zero replications in a condition selected unidimensional solution. "**" sign indicates that less than 1% of replications in a condition selected unidimensional solution.

The results for $\chi^2_{G/D}$ method followed a similar pattern to that of *ALR*, although to a slightly lesser degree ($\chi^2_{G/D}$had fewer replications within conditions that selected one-factor). In conditions with 0% of complexity when the correlation was .60 or larger, the number of replications within conditions that favored a one-factor solution increased. In a condition with 0% of complexity and sample size of 2000, almost 70% of replications favored one-factor solution.

Tests with twenty items per dimension. The frequencies that the methods yielded unidimensional solutions for conditions with twenty items associated per dimension were investigated for 2D and 3D. For each dimensional structure, there were a total of 60 conditions.

2D structures. A complete tabulation of proportions for 2D noncompensatory conditions with 20 items per dimension is presented in Table 14. *ALR* chose a one-factor solution in at least one replication in 44 out of 60 conditions. Large numbers of replications that favored unidimensional solution were found in conditions with N = 500 and N = 1000 and correlation levels of .60 across all levels of complexity. On average, in these conditions, *ALR* selected a one-factor solution almost 300 times (median number of replications across these conditions was 186.5). $\chi^2_{G/D}$ and RMSR selected one-factor solution in fewer conditions than *ALR*.

Table 14.

Proportion of Replications Across Complexity Levels for Conditions with Two-dimensional Noncompensatory MIRT and 20 Items per Dimension

		Method											
		ALR				$\chi^2_{G/D}$				RMSR			
Complex		0%	10%	30%	50%	0%	10%	30%	50%	0%	10%	30%	50%
N	ρ												
	.00	.39	.13	**	.03	-	-	-	-	-	-	-	-
	.30	**	-	.11	.16	-	-	-	-	-	-	-	-
500	.60	.11	.13	.74	.81	-	-	-	-	-	-	.02	**
	.75	.71	.74	.89	.93	-	-	**	**	.01	.02	.17	.10
	.90	.95	.92	.88	.85	.02	**	**	-	.70	.22	.05	.05
	.00	.41	.04	-	-	-	-	-	-	-	-	-	-
	.30	-	-	-	**	-	-	-	-	-	-	-	-
1000	.60	-	-	.40	.49	-	-	-	-	-	-	-	-
	.75	.27	.35	.69	.72	-	-	-	-	-	-	.01	-
	.90	.90	.71	.63	.53	-	-	-	-	.34	**	-	-
	.00	.41	**	-	-	-	-	-	-	-	-	-	-
	.30	-	-	-	-	-	-	-	-	-	-	-	-
2000	.60	-	-	.02	.06	-	-	-	-	-	-	-	-
	.75	**	.02	.35	.32	-	-	-	-	-	-	-	-
	.90	.73	.21	.14	.05	-	-	-	-	**	-	-	-

Note: Each condition has a total of 500 replications. "-" sign indicates that zero replications in a condition selected unidimensional solution. "**" sign indicates that less than 1% of replications in a condition selected unidimensional solution.

The $\chi^2_{G/D}$method identified at least one replication with a preferred unidimensional solution in only 5 out of 60 conditions, with a maximum of 9 replications within any of the five conditions. These conditions all had *N* = 500, correlation of .75 or .90, across the levels of complexity. The RMSR method resulted in selection of 14 out of 60 conditions that yielded preferred one-factor solutions to at least one replication. Large numbers of replications that favored one-factor solutions were found in conditions with .90 correlation and complexity levels of 0% and 10%, with *N* = 500 and *N* = 1000 (mean and median number of replications in those conditions were 158 and 140, respectively).

3D structures. A complete tabulation of proportions for 3D noncompensatory conditions with 20 items per dimension is presented in Table 15. *ALR* yielded at least one replication that favored a unidimensional solution in all of 60 conditions. A large number of replications within conditions that favored one-factor solution were found across sample sizes and complexity levels. With only a few exceptions, the same trend was observed across all levels of correlation; as correlation among dimensions increased, the number of replications also increased. RMSR method selected one-factor model in 27 out of 60 conditions; most of which were with *N* = 500 and *N* = 1000.

The largest proportions of unidimensional selection within a condition were found in conditions with .90 correlations and no complexity, although large proportions were also found in conditions with *N* = 500 and complexity level of 50%. $\chi^2_{G/D}$ favored a one-factor model in 15 out of 60 conditions; the fewest out

of the three NOHARM-based methods. Additionally, many of these 15 conditions contained few replications that favored unidimensional solution.

Table 15.

Proportion of Replications Across Complexity Levels for Conditions with Three-dimensional Noncompensatory MIRT and 20 Items per Dimension

		Method											
		ALR				$\chi^2_{G/D}$				RMSR			
Complex		0%	10%	30%	50%	0%	10%	30%	50%	0%	10%	30%	50%
N	ρ												
	.00	.40	.35	.71	.70	-	-	**	-	-	-	**	.03
	.30	.47	.54	.92	.83	**	**	.01	-	-	-	.36	.57
500	.60	.85	.85	.95	.87	-	**	.01	-	.10	.23	.58	.54
	.75	.96	.92	.93	.81	.06	**	-	-	.65	.50	.14	.14
	.90	.98	.92	.84	.54	.12	**	**	-	.94	.12	-	**
	.00	.27	.24	.47	.39	-	-	-	-	-	-	-	-
	.30	.36	.32	.75	.65	-	-	-	-	-	-	.04	.13
1000	.60	.65	.68	.85	.67	-	-	**	-	-	.01	.17	.15
	.75	.89	.80	.73	.38	.02	-	-	-	.22	.07	-	-
	.90	.98	.73	.42	.11	.21	-	-	-	.95	**	-	-
	.00	.19	.16	.31	.20	-	-	-	-	-	-	-	-
	.30	.25	.24	.48	.39	-	-	-	-	-	-	-	-
2000	.60	.39	.41	.55	.24	-	-	-	-	-	-	**	-
	.75	.78	.61	.24	.02	-	-	-	-	**	-	-	-
	.90	.96	.27	.04	**	.25	-	-	-	.92	-	-	-

Note: Each condition has a total of 500 replications. "-" sign indicates that zero replications in a condition selected unidimensional solution. "**" sign indicates that less than 1% of replications in a condition selected unidimensional solution.

Synthesis. Generally, the investigation of unidimensional solutions revealed that *ALR* and $\boldsymbol{\chi^2_{G/D}}$ tended to favor one-factor solution more often than RMSR. It was also generally found that an increase in either the correlation or complexity resulted in a more frequent selection of one-factor model, particularly for $\boldsymbol{\chi^2_{G/D}}$ and *ALR*, and in conditions where the generating 3D MIRT model was noncompensatory (one exception was in conditions with 3D noncompensatory MIRT and 20 items per dimension conditions using $\boldsymbol{\chi^2_{G/D}}$, where fewer conditions and lower proportions within a condition were observed).

Multidimensional Solutions to Multidimensional Data

The following section discusses in depth results with a focus on the three main outcome variables: a) the proportions of selection of the correct dimensional solution, b) the ability to label sets of items as dimension-like, and c) the consistency of the methods in classifying items according to their generating assignment (see Chapter 3 for details on criteria used to label sets of items as dimension-like and classify items). Most of the results are presented in graphical form for easier identification of the main patterns. Results presented in a tabular form for proportions correct across conditions can be found in Appendix A.

Compensatory multidimensional data.

Tests with ten items per dimension in 2D structures.

The proportion of correct dimensional selection. Figure 6 plots the proportions of times within a condition that a method selected the correct 2D solution across complexity levels. The figure contains 15 graphs, which represent

various combinations of the sample size and correlations between dimensions. In each graph, five lines represent the five methods: DETECTe, DETECTcv, RMSR, $\chi^2_{G/D}$, and *ALR*. Rows represent five levels of correlations, while columns represent three different sample sizes. Within each graph, the y-axis represents the proportion correct and ranges from 0 to 1 and the x-axis represents the complexity levels, and includes 0%, 10%, 30%, and 50% complexity.

As observed in Figure 6, the methods had different rates of success in recovering the correct 2D solutions. The RMSR method performed very poorly; it maximally selected the correct solution less than four percent of time; in all conditions, 70% or more of replications yielded a five-factor solution. The performance of other methods depended on the complexity levels, sample size, correlation levels, or some combination thereof.

$\chi^2_{G/D}$ performed quite well, particularly with when N = 500 and N = 1000 with 30% or less complexity in the data and correlation of .75 or less. Its performance tended to diminish at 50% of complexity, with more extreme drop off when N = 2000 and increased correlation. An extreme result was obtained in the condition with N = 2000 and correlation of .75 when data exhibited 50% complexity. Here, $\chi^2_{G/D}$ selected incorrectly a three-factor solution 100% of the time. Another interesting observation was made for conditions with .90 correlation where across all levels of sample size, $\chi^2_{G/D}$ tended to be more accurate at lower (0% and 10%) and higher (50%) levels of complexity than at the

middle 30%. *ALR* performed worse than $\chi^2_{G/D}$method in most occasions; however, its pattern of performance was very similar to that of the $\chi^2_{G/D}$ method.

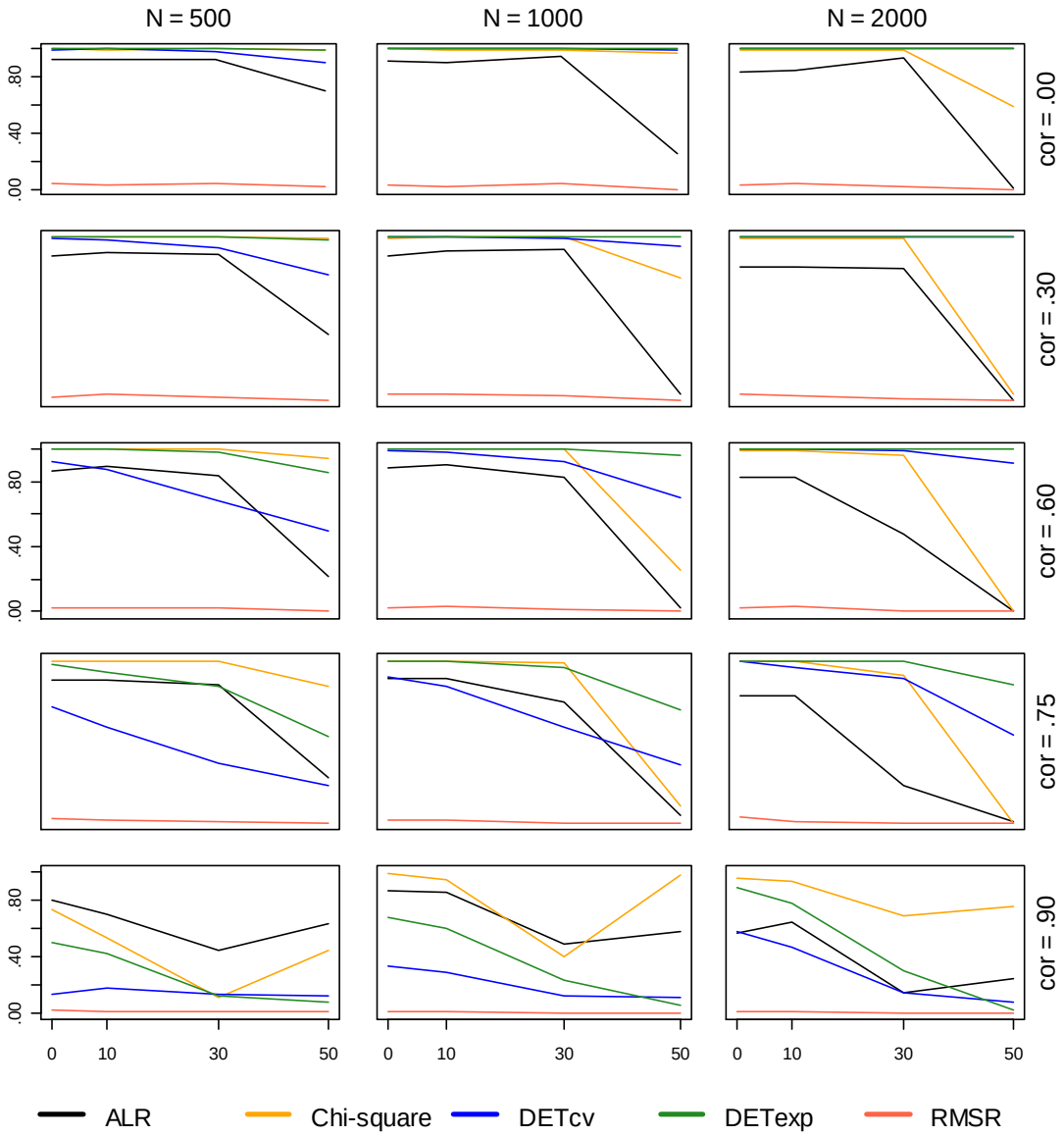

Figure 6. Proportion correct across complexity levels when the data follow a compensatory 2D MIRT model with 10 items per dimension.

DETECTe accurately selected the two-factor solution almost every time for all complexity levels and sample sizes when correlation was .60 or less.

DETECTe was less accurate at correlation of .75 and $N = 500$; particularly when data exhibited 50% complexity. This was true to a lesser extent for $N = 1000$ and $N = 2000$. At correlation levels of .90, DETECTe was performing above .90 in a condition with $N = 2000$ and at 0% of complexity. In all other cases, as complexity in the data increased, DETECTs ability to identify the 2D solution diminished. Similar patterns were found for DETECTcv, with noticeable differences in deterioration for DETECTcv with $N = 500$ and when the correlation was .75. Generally, DETECTe was more accurate than DETECTcv.

The proportion of dimensional labeling. In order to examine the performance of the methods further, marginal proportions of the methods' ability to label a set of items as dimension-like were computed. This variable does not condition on correct selection of the true dimensionality. Results for the dimensional recognition address the question of how often a particular method yields a group of items that facilitate an interpretation of the groups as reasonably representing a true underlying dimension (see Chapter 3 for more details regarding the criteria used to define a set of items as dimension-like).

In 2D conditions, a method could label two (both), (any) one, or none of the sets of items as dimension-like, regardless if the selection of optimal factor solution was correct (i.e., 2), or incorrect (3, 4, or 5). The marginal proportions were calculated across different factor solutions and plotted for easier identification of patterns. Note that in some conditions and for some methods, marginal proportions do not add up to 1. This occurs when a method selected a

unidimensional factor solution as optimal (see section on *Unidimensional solutions* at the beginning of the chapter).

Figures 7 and 8 present the marginal proportions that each method identified sets of items as dimension-like for 30% and 50% complexity levels across the sample sizes and correlations. The results for 0% and 10% complexities were quite similar to the results for 30% complexity, thus only a graph for 30% is shown (see Appendix B for 0% and 10% complexity graphs). As seen from Figure 7, when data exhibited 30% complexity or less, the methods were highly successful at labeling two sets of items as dimension-like across sample size with correlation levels of .75 or less (note the “L” shaped lines for most of the conditions). An exception was found with RMSR and N = 500 at .75 correlation, where the method had fewer instances of selecting two sets as dimension-like.

At a correlation of .90, the methods' abilities to group items in terms of sets that can be labeled as the underlying dimensions diminished, particularly at N = 500. When N = 500, the methods had more success labeling one or none of the sets of items as dimension-like than two. As the sample size increased, marginal proportions for two and none sets of items as dimension-like increased, while labeling only one set as dimension-like decreased.

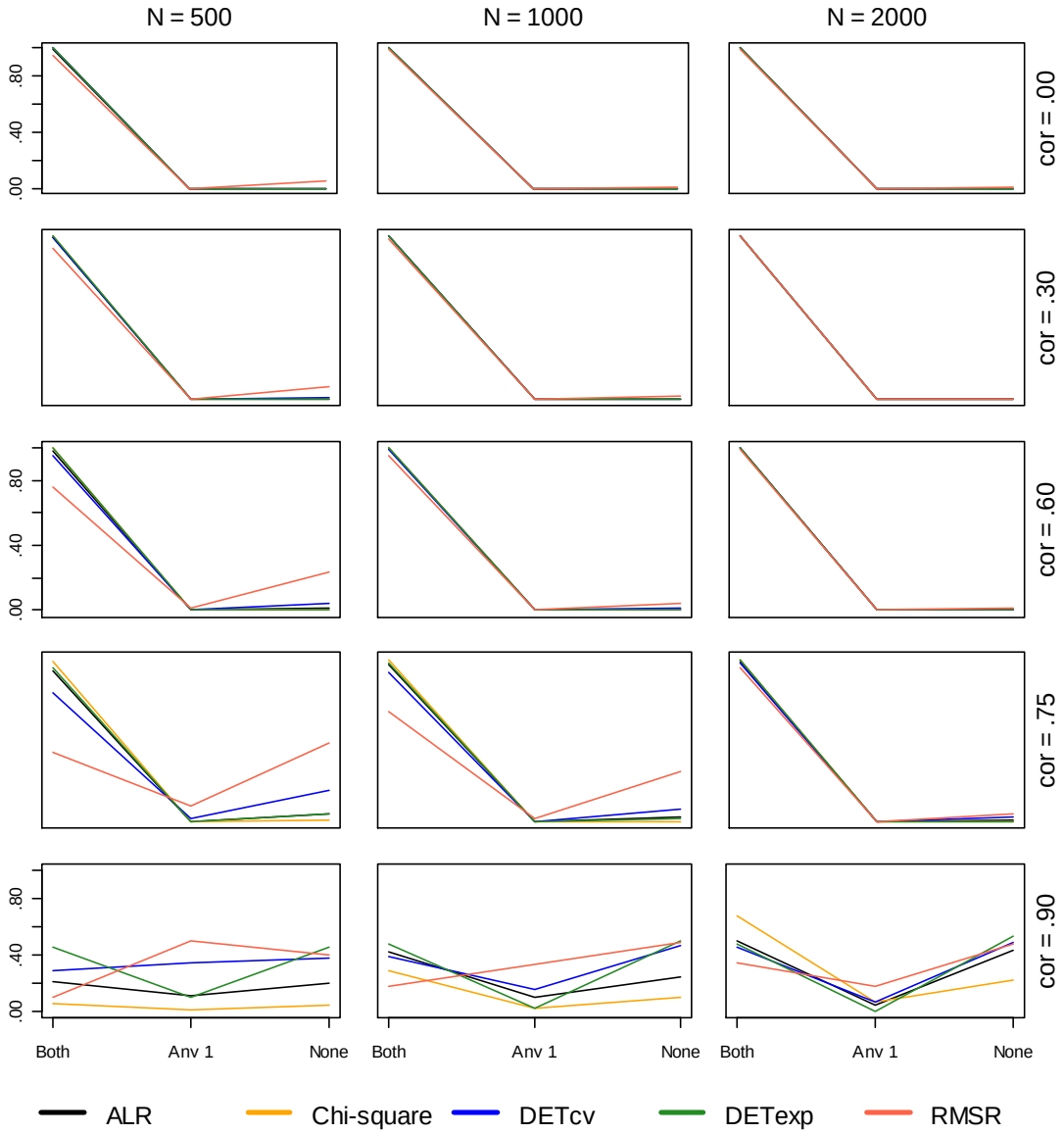

Figure 7. Marginal proportions across 500 replications that a method identified two (both), (any) one, or none of the sets of items as dimension-like (x-axis) when the data exhibit 30% complexity and follow a compensatory 2D MIRT model with 10 items per dimension.

At 50% of complexity, the patterns of performance varied for DETECT-based and NOHARM-based methods (see Figure 8). $\chi^2_{G/D}$, *ALR*, and RMSR were generally more likely to label either two or none sets of items as dimension-like

when correlations were .75 or less (note the “V” shapes for orange, black, and red lines). An exception to this occurred at N = 500 for *ALR* and RMSR. At a correlation of .90, however, the NOHARM-based methods were more likely to label one set of items as being like one of the dimensions. The DETECT-based methods generally failed to label two sets of items as dimension-like across correlation levels and sample size. In only a few conditions did the DETECT methods, particularly DETECTe, have success in labeling any one set as dimension-like. This most often occurred in conditions with N = 2000 and a correlation of .30 or less.

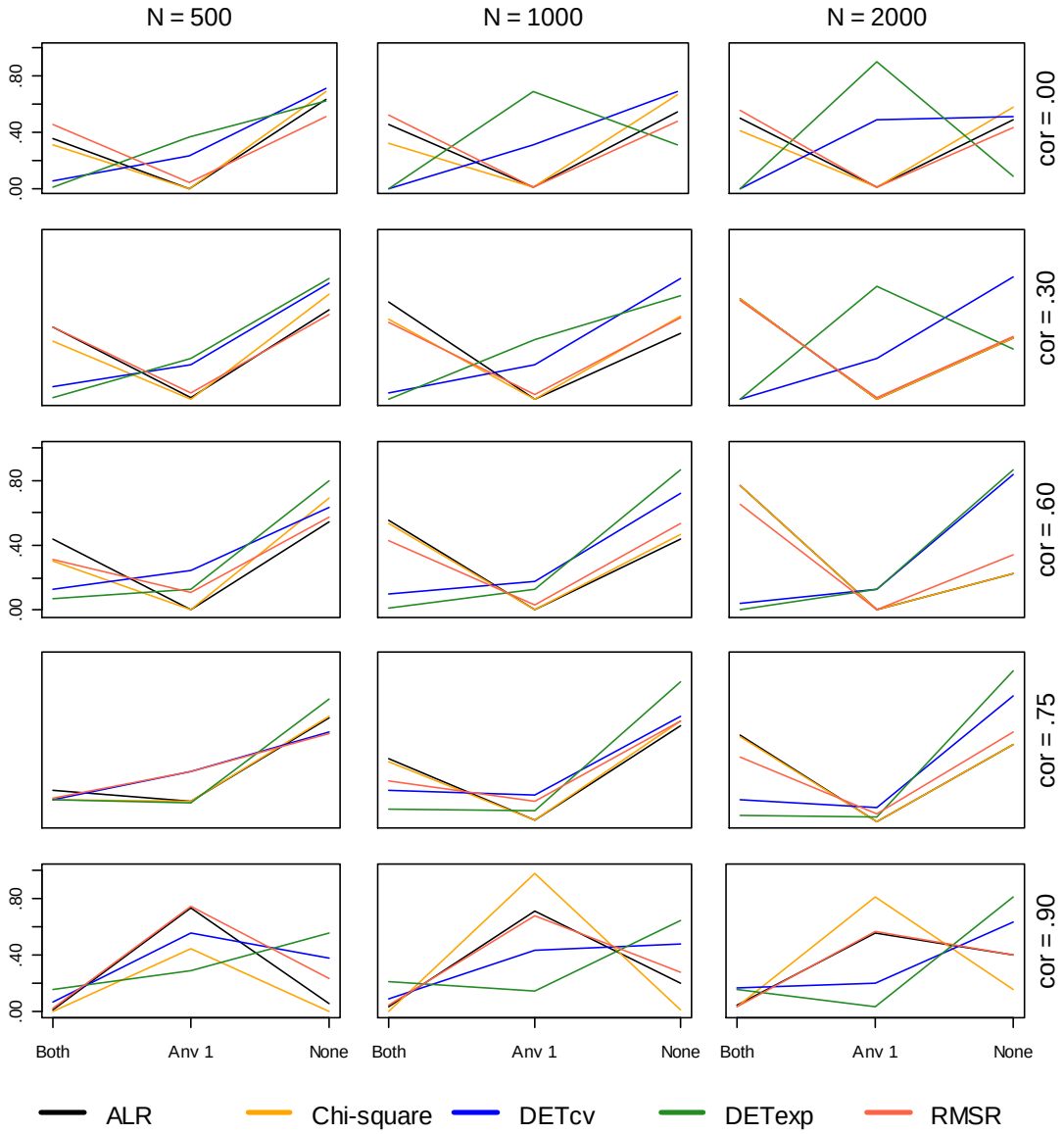

Figure 8. Marginal proportions across 500 replications that a method identified two (both), (any) one, or none of the sets of items as dimension-like (x-axis) when the data exhibit 50% complexity and follow a compensatory 2D MIRT model with 10 items per dimension.

The consistency of item classification. Figure 9 plots the classification consistencies for factorially simple items across complexity levels (x-axis) when the data follow a compensatory 2D MIRT model with 10 items per dimension.

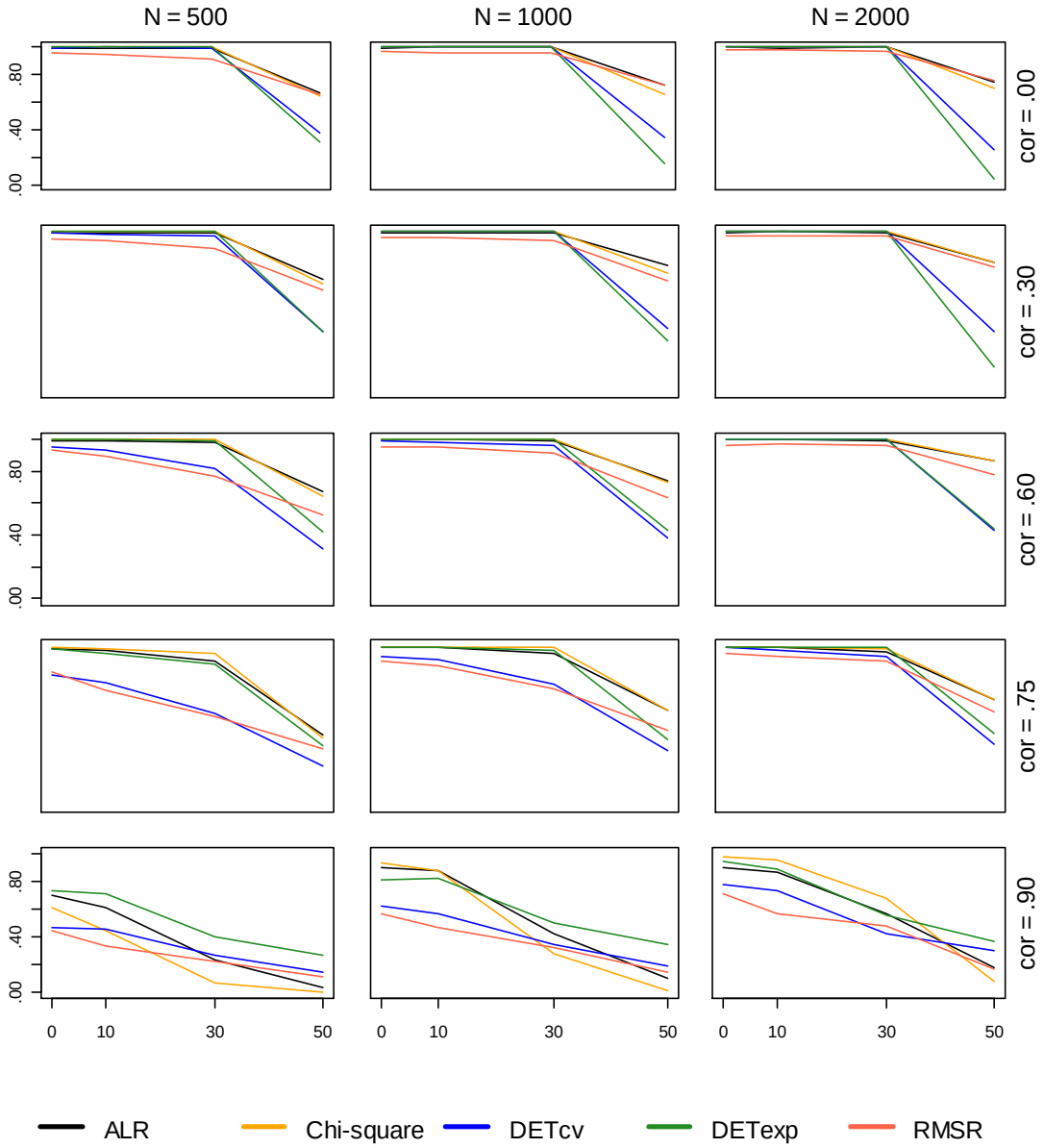

Figure 9. Consistency of factorially simple items across complexity levels when the data follow a compensatory 2D MIRT model with 10 items per dimension.

In Figure 9, it was observed that the consistency for factorially simple items generally improved for all methods, as the sample size increased. All five methods were consistent in classifying factorially simple items at rates of .82 or higher (most at or around 1) for complexity levels of 30% or less, except when

correlation was .90. At a correlation of .90, only $\chi^2_{G/D}$, DETECTe, and *ALR* had rates of above .80 for *N* = 1000 and *N* = 2000.

Although higher rates were found at lower levels of complexity, when data exhibited 50% of complexity, the methods varied in how consistently they classified factorially simple items. At 50% complexity and lower level of correlation (0 or .30), the most successful methods were $\chi^2_{G/D}$, *ALR*, and RMSR. DETECTe was the least consistent, particularly with *N* = 2000 when its rates were .05 and .16, respectively. Though DETECTcv performed slightly better than DETECTe with *N* = 500 and a correlation of .60, as the correlation increased to .75 or .90, DETECTe became more consistent than DETECTcv for all sample sizes.

Figure 10 plots the classification consistencies for factorially complex items across complexity levels (x-axis) when the data follow a compensatory 2D MIRT model with 10 items per dimension. Note that on these graphs, only conditions with complexity were plotted, hence, the x-axis included only levels of 10%, 30%, and 50%.

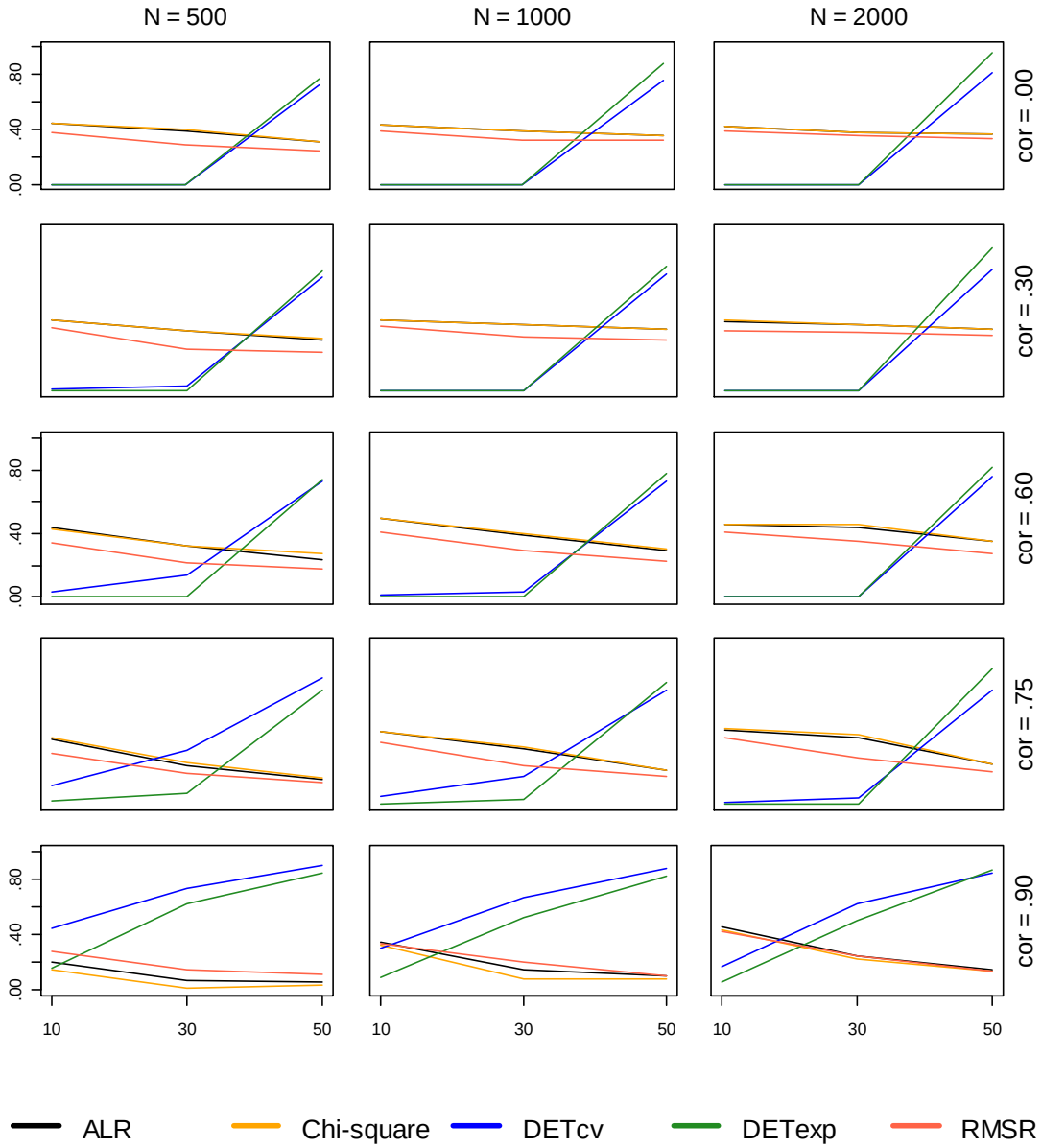

Figure 10. Consistency of factorially complex items across complexity levels when the data follow a compensatory 2D MIRT model with 10 items per dimension.

In Figure 10, two interesting patterns were noted. First, the classification rates of factorially complex items for the NOHARM methods were quite similar, as indicated by close proximity of these lines on most of the graphs. The largest

differences were noted at correlation levels of .60 and .75, where $\chi^2_{G/D}$ and *ALR* outperformed the RMSR at 30% complexity (on average by 10% to 12%). A second interesting pattern was observed at 50% complexity. Although NOHARM based methods were more successful than DETECT-based methods in classification at complexity levels of 10% and 30%, the opposite was found for 50% complexity levels across all correlation levels and sample sizes.

When data exhibited 50% complexity, DETECTe and DETECTcv had higher classification rates, ranging from .72 to .96 for various sample sizes and correlation levels. At a correlation level of .90, this type of switch was noted even earlier; for N = 500, The DETECT-based methods at correlation level of .90 had comparable or higher classification rates than NOHARM-based methods. At N = 1000 and N = 2000, notable differences occurred at 30% complexity.

Tests with ten items per dimension in 3D structures.

The proportion of correct dimensional selection. Figure 11 plots the proportions of times within a condition that a method selected the correct 3D solution across different complexity levels (x-axis).

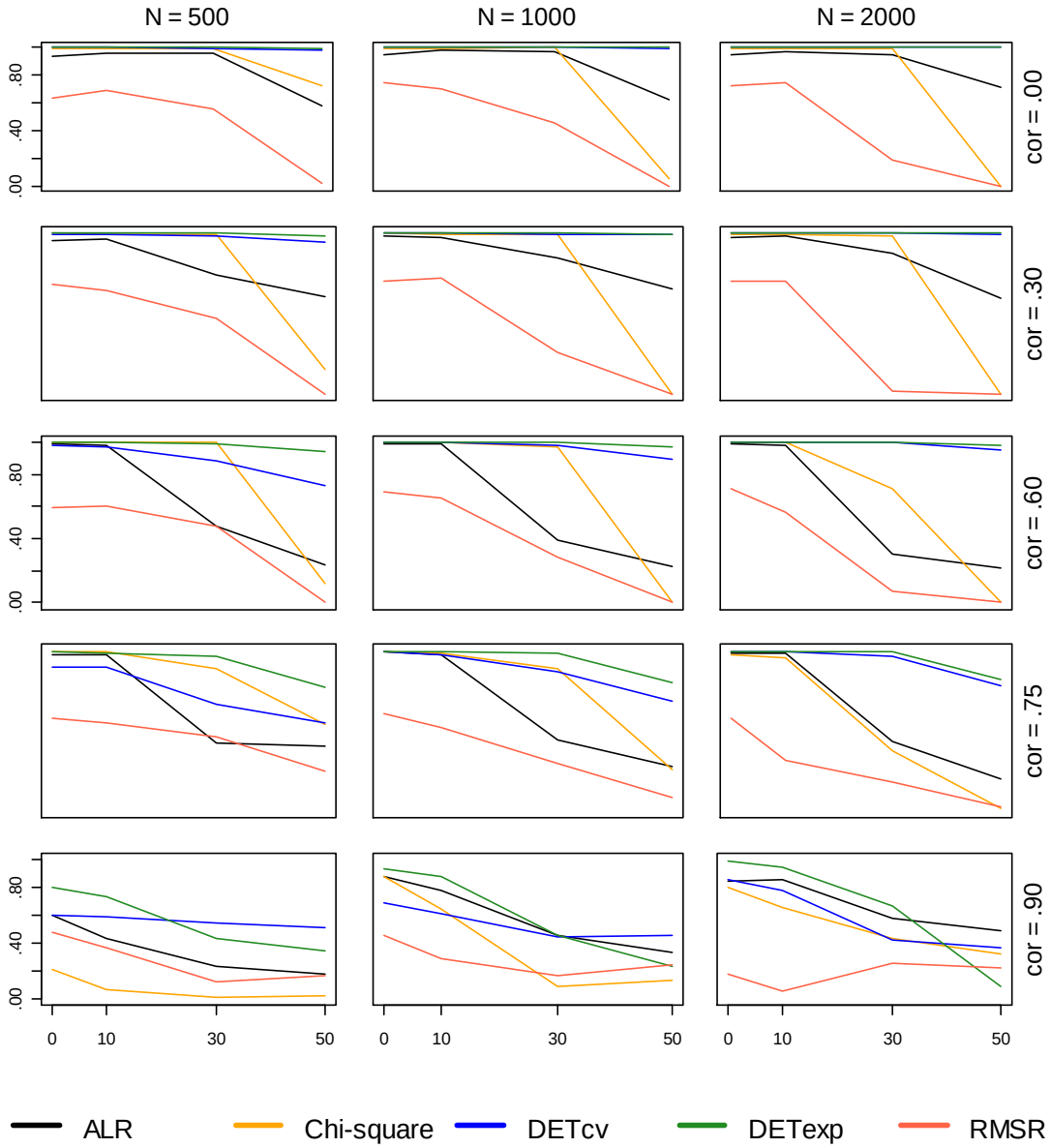

Figure 11. Proportion correct across complexity levels when the data follow a compensatory 3D MIRT model with 10 items per dimension.

RMSR tended to perform poorly across conditions when the data followed a compensatory 3D MIRT model with 10 per dimension (Figure 11). A slight improvement was noted in conditions with correlation of .00 and N = 1000 and N = 2000, and in conditions with 0% or 10% complexity. DETECTe tended to

perform the best out of the methods examined across all levels of complexity and sample sizes when correlation levels were .75 or less. Similar behavior was observed for DETECTcv, with larger discrepancies noted in smaller sample size and correlation of .75.

$\chi^2_{G/D}$ tended to select the correct factor solution for N = 500 and N = 1000, correlations of .75 or less, for complexity levels of 30% or lower. *ALR* also performed well for complexity levels of 30% or less but only for correlation of .30 or lower. An increase in the correlation resulted in *ALR* performing less accurately (selecting the correct solution only half of the time) even when the data exhibited 30% of complexity. When correlation was at .90, all methods performed less accurately especially as the complexity levels increased.

The proportion of dimensional labeling. In 3D conditions, a method could label three, any two (both), (any) one, or none of the sets of items as dimension-like, regardless whether the optimal factor solution was a two-, three-, four-, or five-factors. As in 2D conditions, marginal proportions were calculated across different factor solutions and were plotted. Figures 12 and 13 present the proportions of times that each method identified sets of items as dimension-like for 30% and 50% complexity level across sample size and correlation (note that 0% and10% complexity conditions had similar patterns to 30% conditions; see Appendix B for 0% and 10% complexity graphs).

Beginning with Figure 12, it was observed that the methods were highly successful in labeling three sets of items as dimension-like when the data

exhibited 30% complexity or less, and the correlations were .30 or smaller for all sample sizes. As complexity or the correlations increased, the methods were less successful in identifying three sets of items as dimension-like, but identified any one set as dimension-like more often.

As illustrated in Figure 12, when data exhibited 30% of complexity, the methods tended to identify three sets of items as dimension-like more often when correlations were lower and sample sizes were larger. At correlations of .60 or larger, the NOHARM-based methods were more likely to identify three sets as dimension-like, but were less likely to label any two or one set. The DETECT-based methods on the other hand tended to successfully label any one set as dimension-like most often. The DETECT-based methods' ability to label any one set of items as dimension-like particularly increased as the sample size and correlations increased; more so for DETECTe than DETECTcv (note the inverted "V" shapes of blue and green lines).

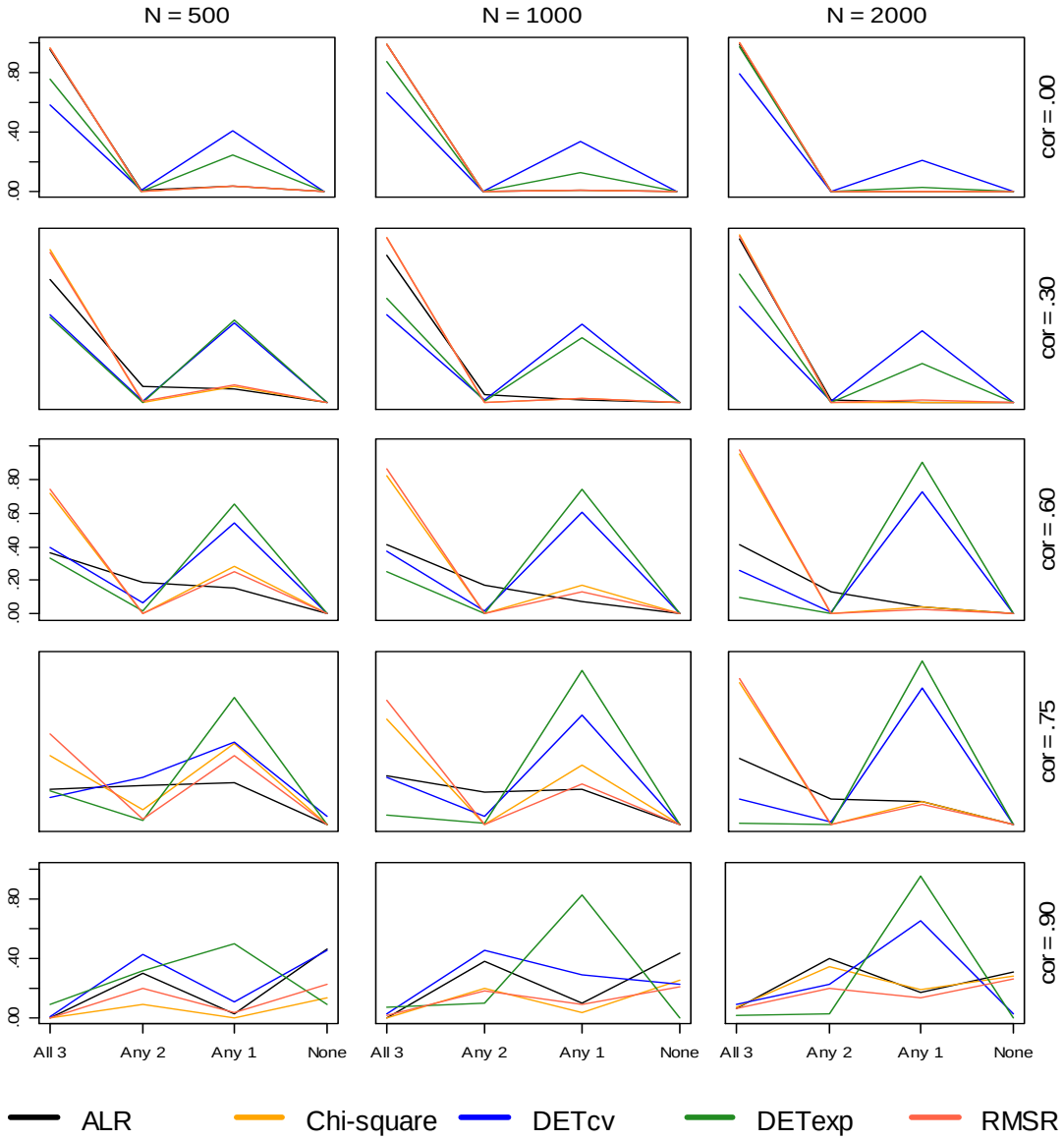

Figure 12. Marginal proportions across 500 replications that a method identified (all) three, (any) two, (any) one, or none of the sets of items as dimension-like (x-axis) when the data exhibit 30% complexity and follow a compensatory 3D MIRT model with 10 items per dimension.

In Figure 13, a somewhat opposite pattern was observed for lower values of the correlations compared to conditions with 30% of complexity. The DETECT-based methods were more successful in labeling any two sets of items

as dimension-like, while the NOHARM-based methods labeled three sets or any one set as dimension-like more often. As the correlation increased to .60 and particularly with N = 2000, the methods behaved more similarly, increasing the relative frequency of labeling any one of the sets of items as dimension-like. When correlation was at .90, the methods were most likely not to label any of the sets as dimension-like, and only the DETECT-based methods were likely to label three, any two, or any one set.

Note that DETECTe labeled any one set as dimension-like more often than any other method when the correlations were .90 across sample size, while DETECTcv labeled any two sets more often than any other method when N = 1000 or N = 2000. In conditions with .90 correlation, the NOHARM-based methods did not successfully label any of the sets as dimension-lie (i.e., large marginal proportions in the last category “none” on x-axis in the figure).

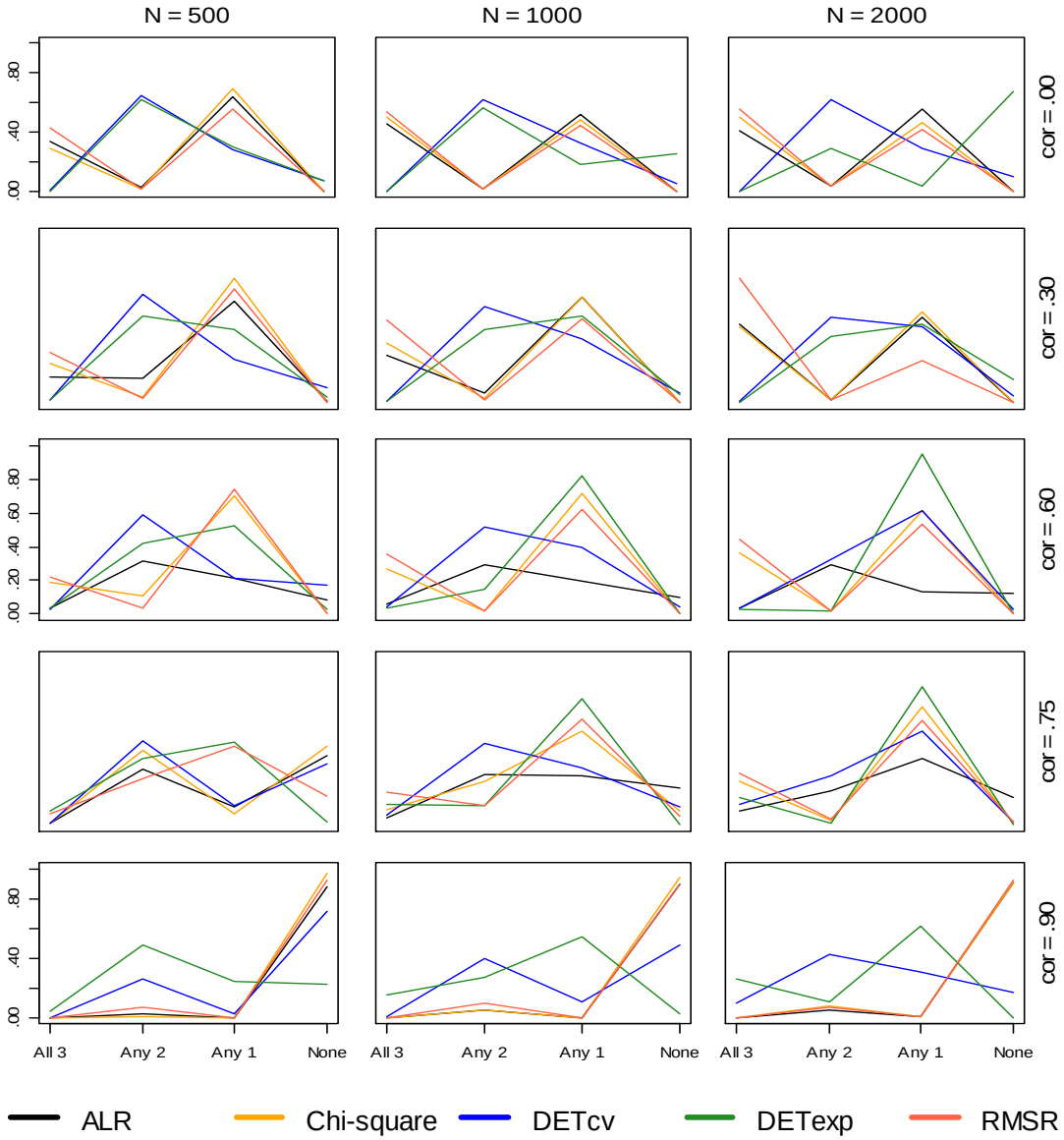

Figure 13. Marginal proportions across 500 replications that a method identified (all) three, (any) two, (any) one, or none of the sets of items as dimension-like (x-axis) when the data exhibit 50% complexity and follow a compensatory 3D MIRT model with 10 items per dimension.

The consistency of item classification. Figure 14 plots the classification consistencies for factorially simple items across complexity levels (x-axis) when the data follow a compensatory 3D MIRT model with 10 items per dimension.

DETECTe and DETECTcv generally were more successful in classification of factorially simple items in 3D structures than their NOHARM counterparts. This was particularly true when N = 500 across correlation levels and complexity levels of 30% or less.

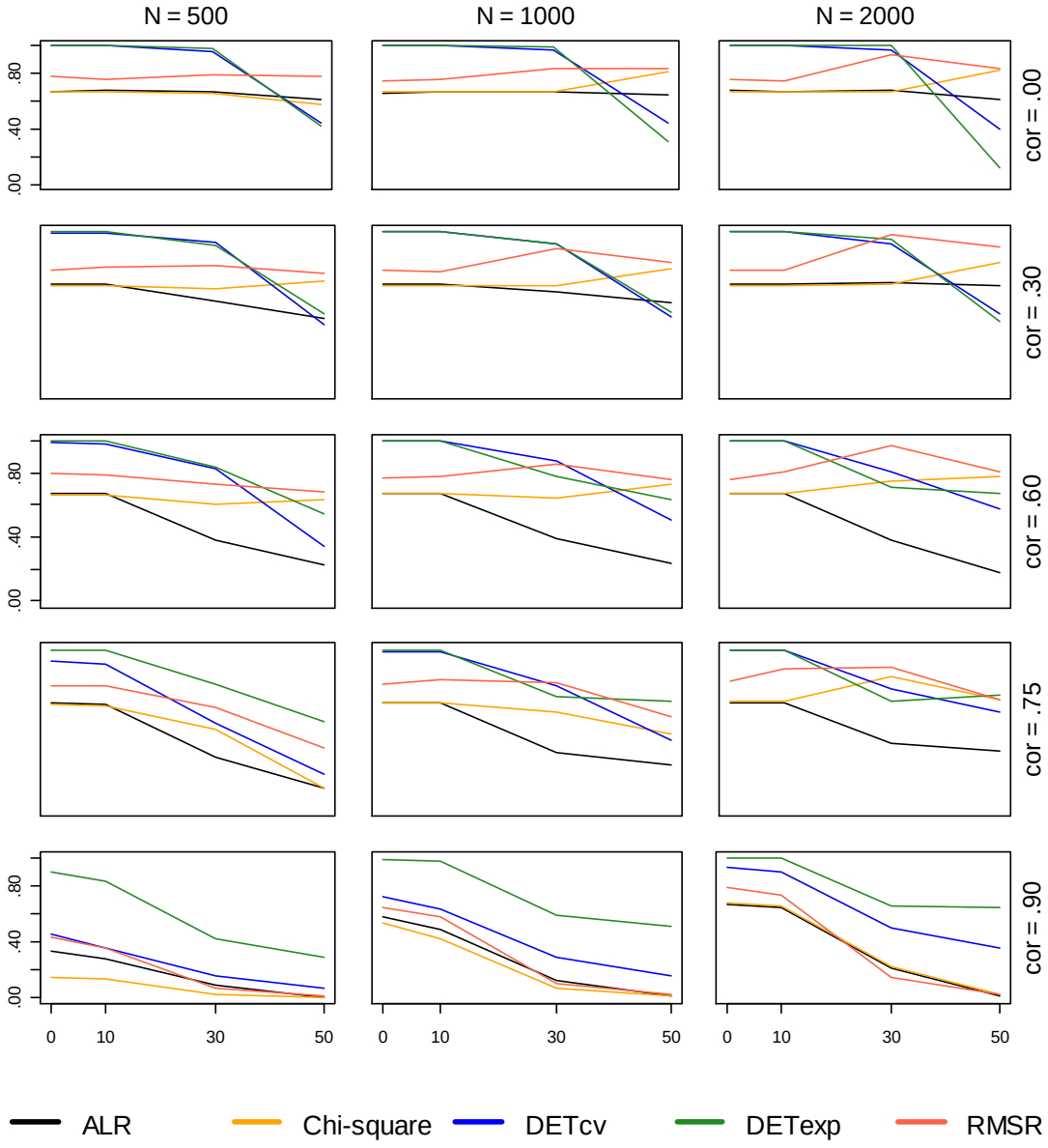

Figure 14. Consistency of factorially simple items across complexity levels when the data follow a compensatory 3D MIRT model with 10 items per dimension.

Similar patterns were observed in conditions with N = 1000 and N = 2000; with an exception of RMSR, for which classification rates improved as the complexity levels increased for correlations of .60 or smaller. *ALR* classification rates were around .67 for correlation of .00 and all sample sizes; however, rates decreased with the increase of complexity. At correlations of .90, only DETECTe had acceptable classification rates (particularly with N = 2000). Its rates were close to 1 at 0% and 10% of complexity; however, the rates dropped down to around .65 as complexity increased to 30% and 50%. Similar observations were noted in cases where N = 500 and N = 1000 for DETECTe at correlation of .90.

The classification of factorially complex items in 3D structures when data follow a 3D compensatory MIRT with 10 items per dimension is plotted across complexity levels in Figure 15. It was noted that the NOHARM-based methods tended to classify complex items better for complexity levels of 30% or less for correlation levels of .00 and .30. However, at 50% complexity, DETECT-based methods strictly outperformed *ALR*, RMSR, and $\chi^2_{G/D}$.

The largest differences were found at correlation levels of .60 or higher for all sample sizes. When the correlations were.90, differences in classification rates were notable even at N = 500, and at lower levels of complexity. For example, at 30% complexity and N = 500 and N = 1000, DETECTe reported .93 and .95 classification rates, while NOHARM methods were all at around zero. Between the two DETECT methods, most notable differences in classification rates were observed in following conditions. DETECTe performed better at 30% complexity

and correlation of .60 and .75 when *N* = 1000 and *N* = 2000; the difference was of .28. DETECTcv however outperformed DETECTe in 10% complexity, *N* = 500 when correlation was .90, where its classification rate was at .58 and DETECTe was at .09.

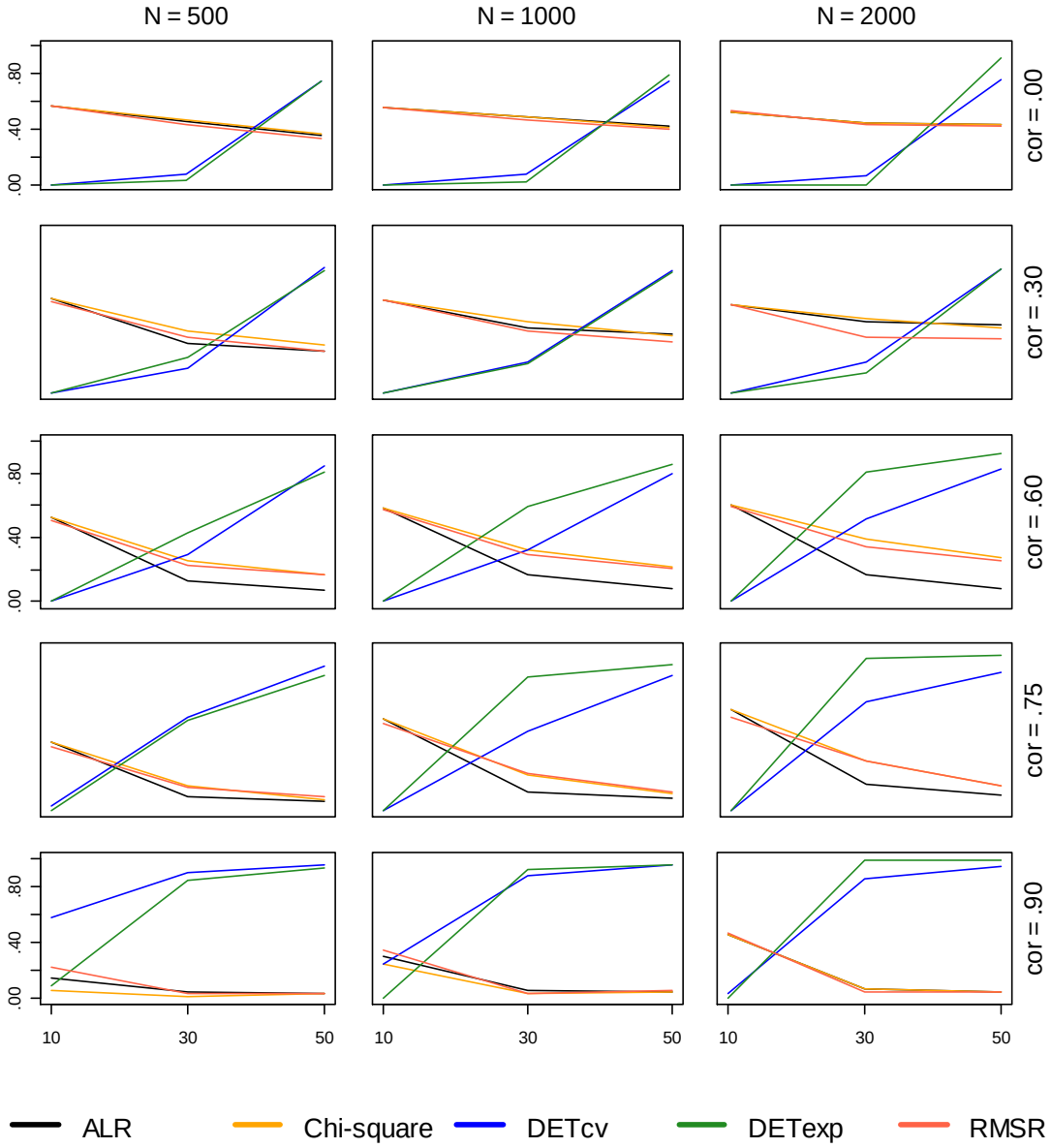

Figure 15. Consistency of factorially complex items when the data follow a compensatory 3D MIRT model with 10 items per dimension.

Tests with twenty items per dimension in 2D structures.

The proportion of correct dimensional selection. Figure 16 plots the proportions that the methods correctly selected a two-factor model when the data follow a compensatory 2D MIRT model with 20 items per dimension. DETECTe outperformed the other four methods in most of the cases. Good performance was noted across various levels of complexity. DETECTe selected the correct dimensional structure virtually always when $N = 2000$, and correlation was .75 or smaller. When $N = 500$ or $N = 1000$, DETECTe performed somewhat well; however, at $N = 500$ and correlation of .90, the DETECT-based methods suffered. In all of the conditions, DETECTe selected the correct solution in larger proportions that DETECTcv across all levels of complexity. Both methods seemed to improve with the increase in sample size, but suffer as the correlations increased.

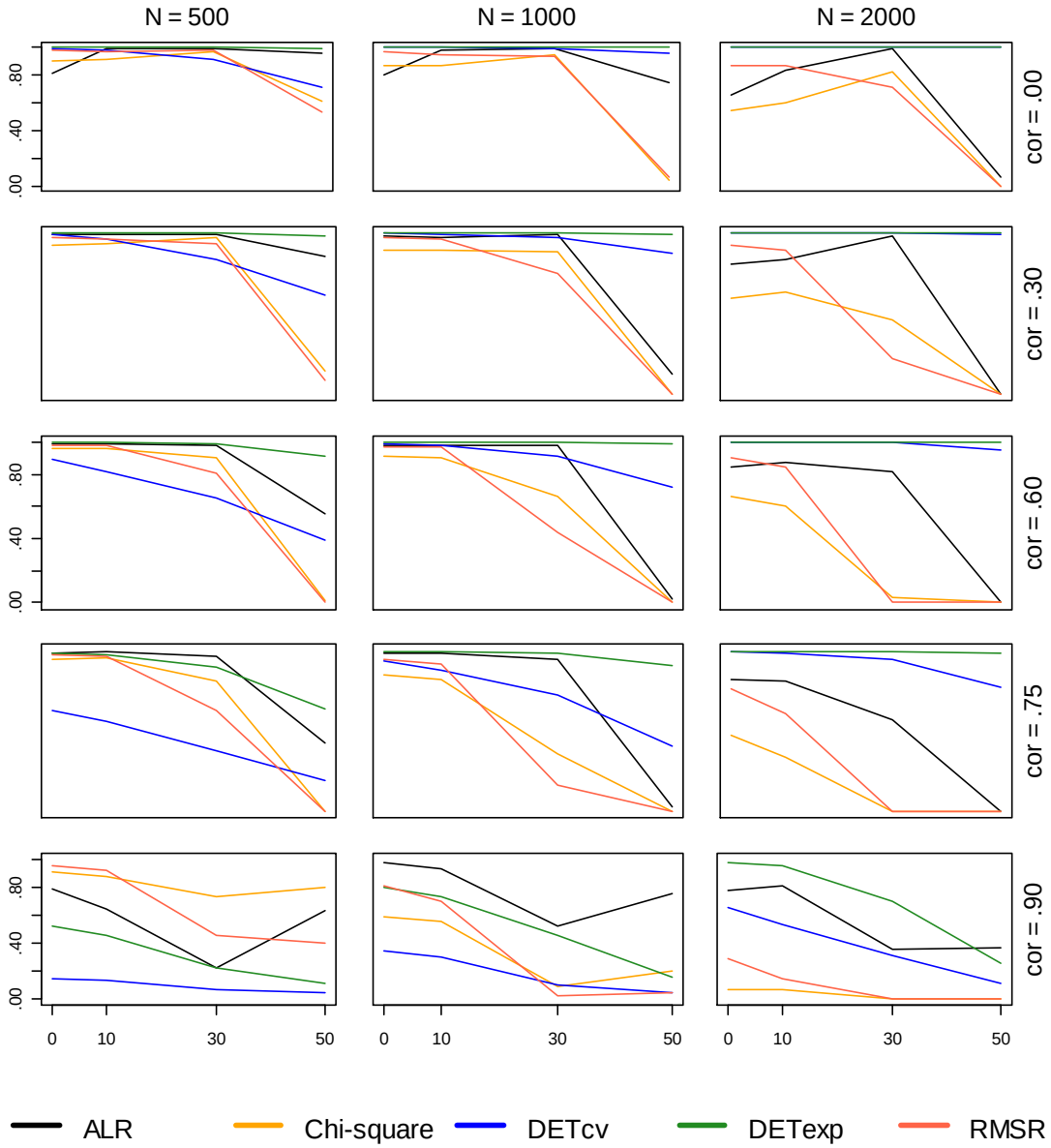

Figure 16. Proportion correct across complexity levels when the data follow a compensatory 2D MIRT model with 20 items per dimension.

ALR performed equally well or better than its NOHARM counterparts across all levels of complexity and sample sizes, expect in the condition with N = 500 and .90 correlations, where $\chi^2_{G/D}$ outperformed *ALR* across all levels of complexity. Generally speaking, when N = 500, the methods based on NOHARM

output performed well for complexity levels of 30% or less when the correlations between dimensions did not exceed .60. As the correlations increased to .75 or .90, *RMSR* performed somewhat satisfactory only for 0% and 10% complexity levels. *ALR* did not perform well in conditions with larger correlation levels; at .75 the degradation in performance occurred at 50% complexity, while at .90, *ALR* seemed to have performed better at the extreme ends of complexity (0% and 50%).

The proportion of dimensional labeling. Figure 17 illustrates the marginal proportions of labeling sets of items as dimension-like for conditions where the data exhibit 30% of complexity, following a true 2D compensatory structure with 20 items per dimension (note that figures for 0% and 10% look very similar to 30% complexity and are included in Appendix B).

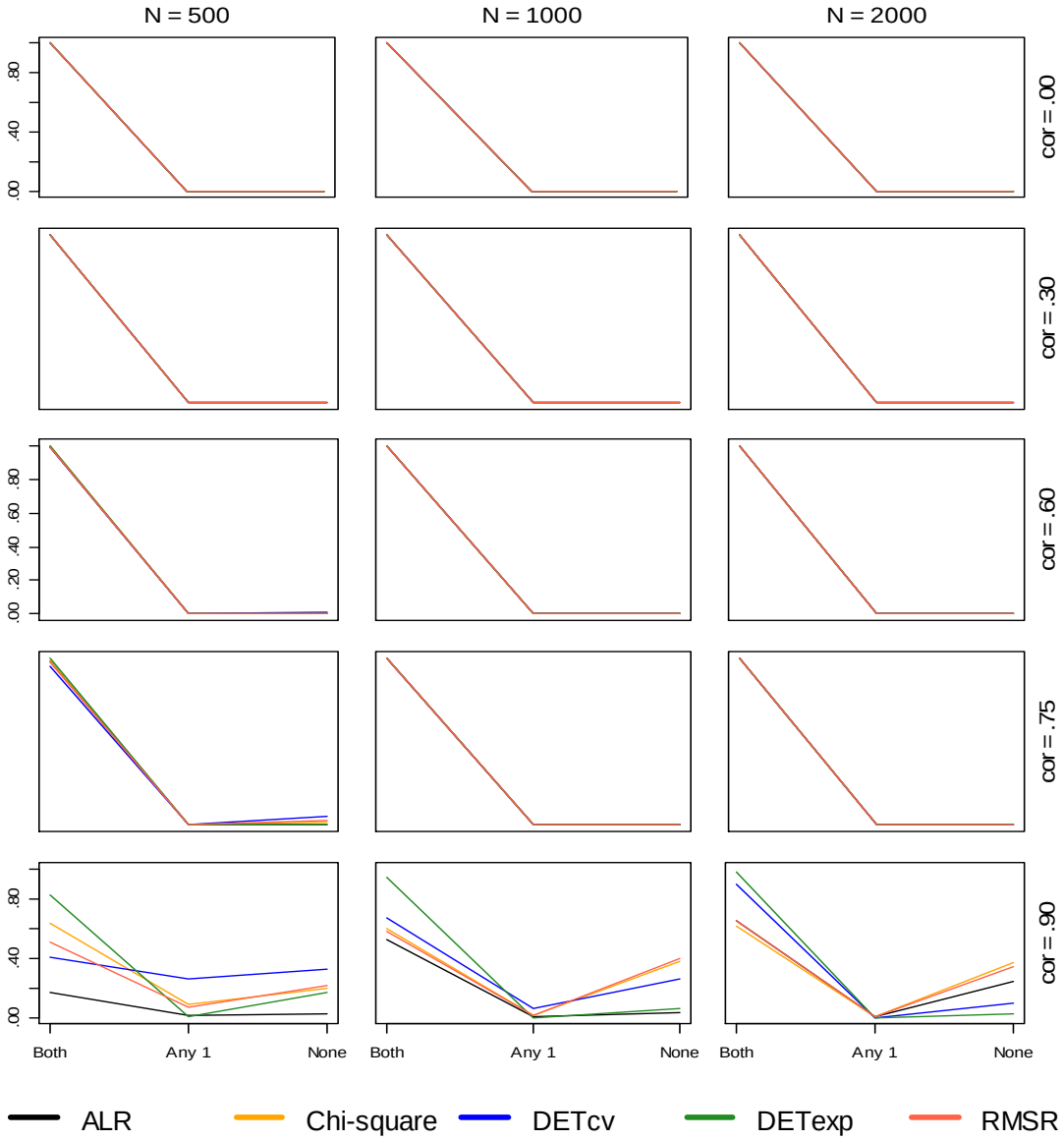

Figure 17. Marginal proportions across 500 replications that a method identified two (both), (any) one, or none of the sets of items as dimension-like (x-axis) when the data exhibit 30% complexity and follow a compensatory 2D MIRT model with 20 items per dimension.

When data exhibited 30% complexity or less, all the methods were highly successful in labeling two sets of items as dimension-like across sample size and correlation levels, except in conditions with a correlation of .90 and N = 500,

where the methods tended to label two dimensions less often (note the “L” shaped lines in the graphs).When correlations were .90, an increase in sample size resulted in the DETECT-based methods (particularly DETECTe) to label two sets of items as dimension-like more frequently, while NOHARM methods resulted in increases in labeling none of the sets as dimension-like. Generally in conditions with high correlation, the methods were either identifying two sets or none as dimension-like (marginal proportions for labeling any one set as dimension-like were low or zero throughout the conditions with up to 30% complexity).

Figure 18 illustrates the marginal proportions of labeling sets of items as dimension-like for conditions where the data exhibit 50% of complexity, following a true 2D structure with 20 items per dimension. As seen in Figure 18, when complexity was at 50%, the NOHARM-based methods were much more likely to label either two or none of sets of items as dimension-like when the correlation was zero. As the correlation levels increased to .60, the marginal proportions for labeling two sets of items as dimension-like for the NOHARM-based methods increased, while at the same time the marginal proportions for labeling none of the sets as dimension-like decreased. A similar effect was found for increases in sample size.

In all of these conditions, the DETECT-based methods were rather unlikely to label two sets of items as dimension-like. As the correlation increased, the DETECT-based methods yielded higher marginal proportions for identifying both sets of items as dimension-like; however, those never rose above .27. At a

correlation of .90 and *N* = 500, all the methods tended to label one set of items as dimension-like.

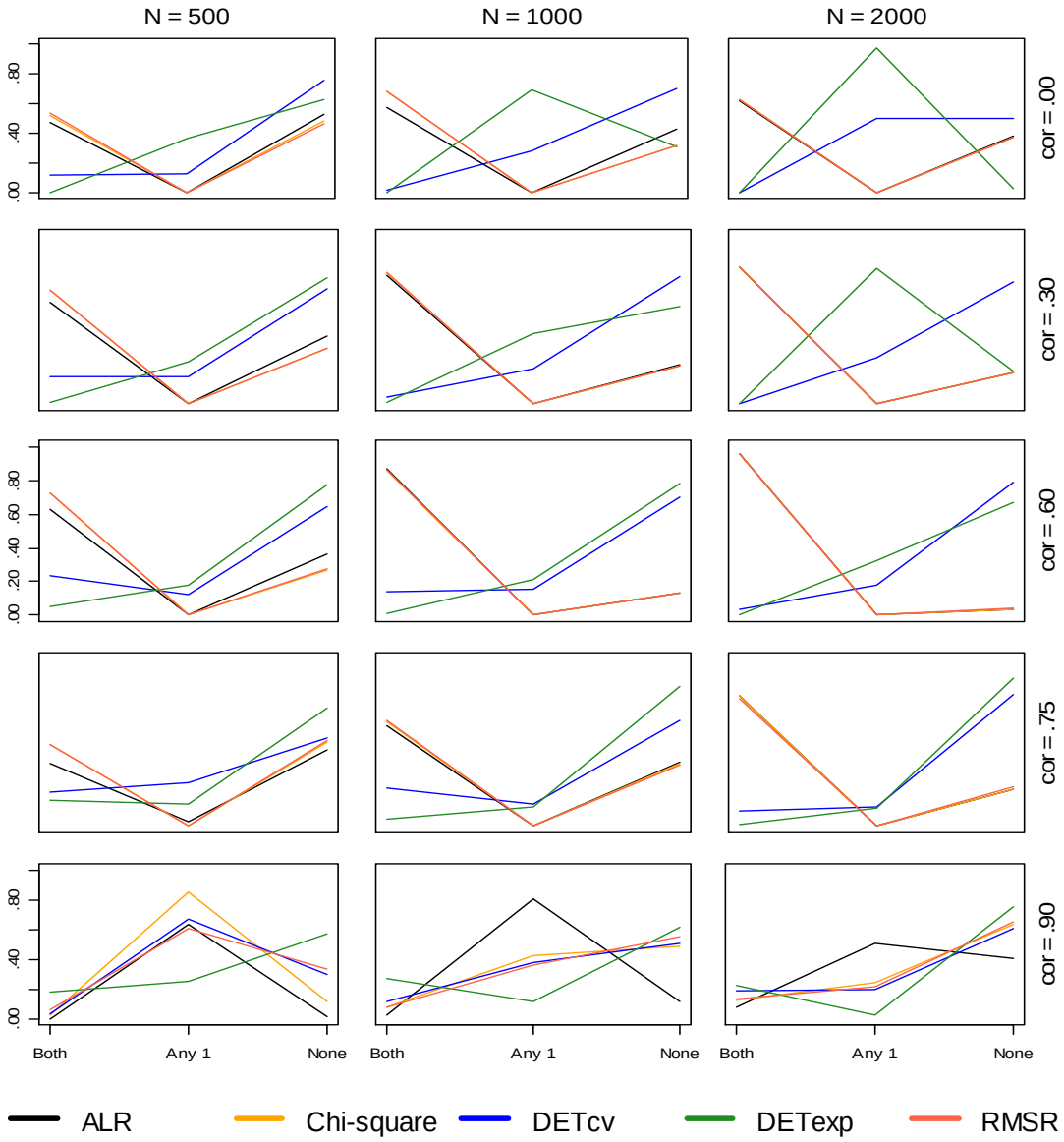

Figure 18. Marginal proportions across 500 replications that a method identified two (both), (any) one, or none of the sets of items as dimension-like (x-axis) when the data exhibit 50% complexity and follow a compensatory 2D MIRT model with 20 items per dimension.

The consistency of item classification. Figure 19 plots the classification consistencies for factorially simple items across complexity levels (x-axis) when the data follow a compensatory 2D MIRT model with 20 items per dimension. The methods were successful in classifying factorially simple items across different sample sizes and correlation levels of .75 or less, when 30% of less complexity existed. Additionally, the NOHARM-based methods yielded high classification rates even for 50% of complexity and correlations of .60 and .75. DETECTe yielded high classification rates when $N = 2000$ and correlation of .90 for complexity levels of 0%, 10%, and 30% of .99, .97, and .89, respectively.

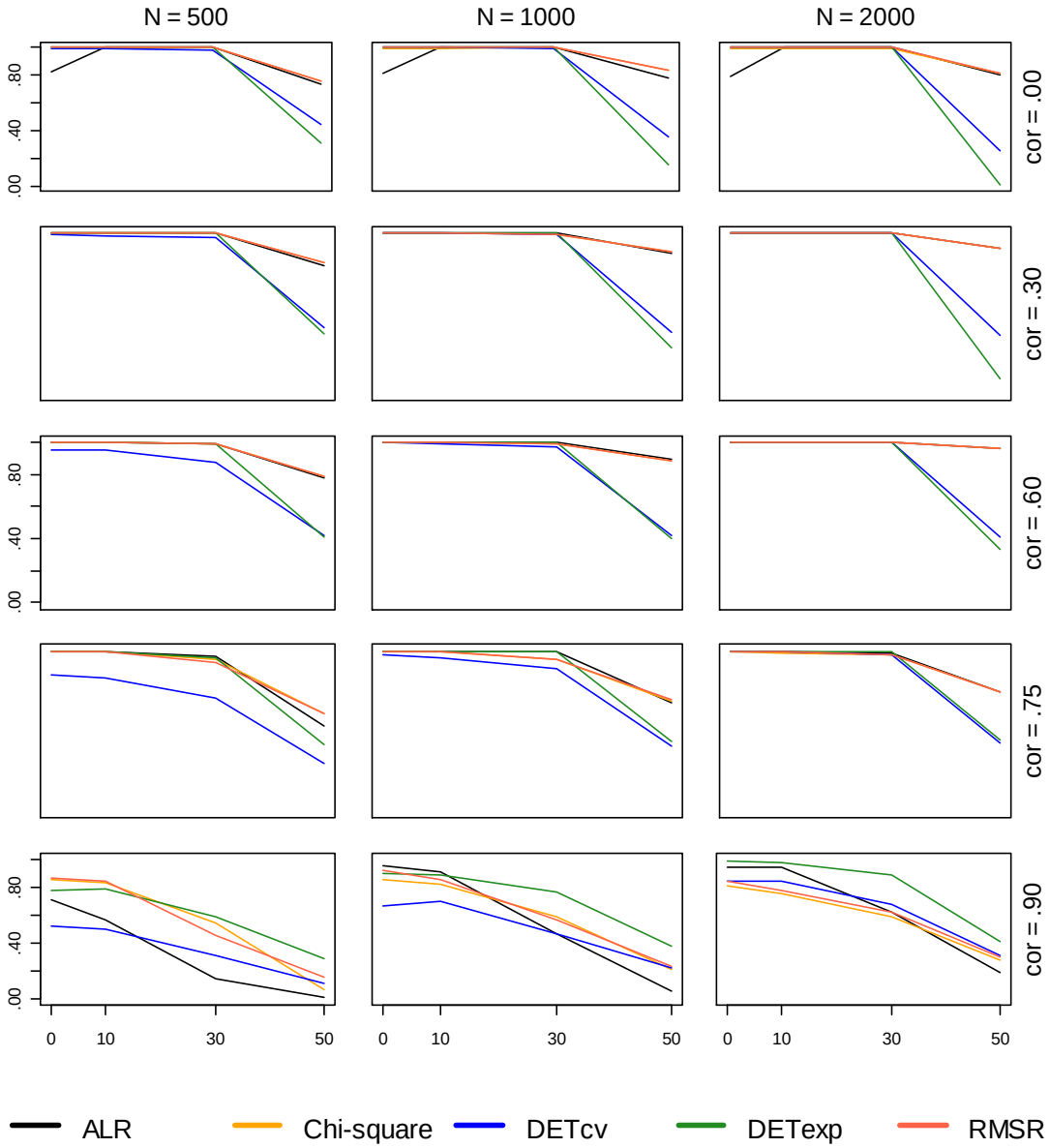

Figure 19. Consistency of factorially simple items when the data follow a compensatory 2D MIRT model with 20 items per dimension.

Figure 20 plots the classification consistencies for factorially complex items across complexity levels (x-axis) when the data follow a compensatory 2D MIRT model with 20 items per dimension. For the NOHARM-based methods, the

classification rates never exceeded .62 (note mostly horizontal orange, red, and black lines), and were largely at .45 or below.

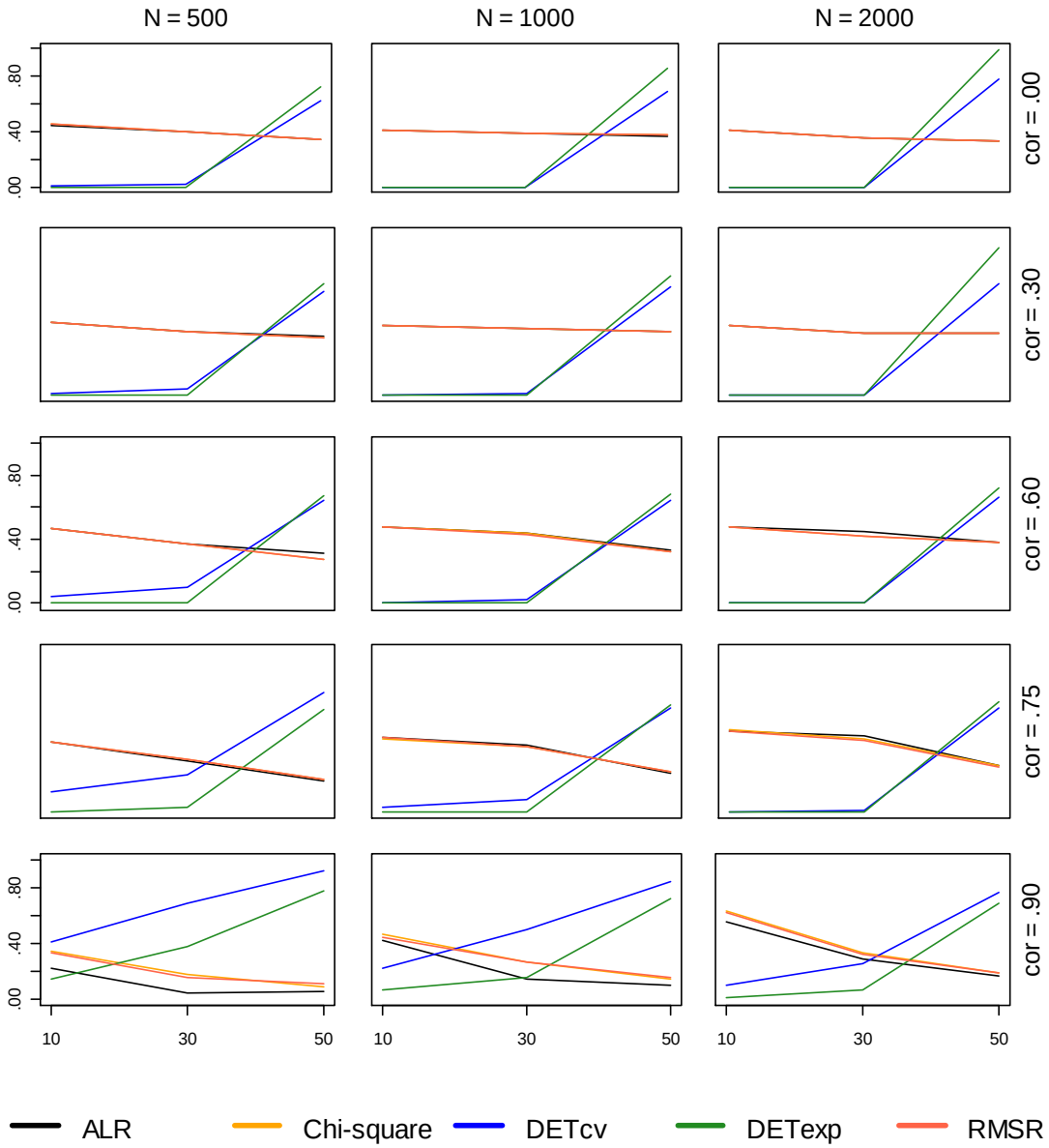

Figure 20. Consistency of factorially complex items when the data follow a compensatory 2D MIRT model with 20 items per dimension.

The methods in DETECT were most consistent in their classifications of complex items when data exhibited a large degree of complexity (50%), across

the levels of sample size and correlations. For example, when the correlation was .90 and N = 1000, DETECTcv classification rates increased as the level of complexity increased from .40 to .92 in conditions with N = 500, and from .22 to .84 in conditions with N = 1000. Similar classification rates of factorially complex items and associated increases were noted in other conditions for DETECT-based methods.

Tests with twenty items per dimension in 3D structures.

The proportion of correct dimensional selection. Figure 21 plots the proportions of correct dimensional selection across complexity levels when the data follow a compensatory 3D MIRT model with 20 items per dimension. Overall, the DETECT-based methods outperformed the NOHARM-based counterparts in correctly identifying the number of dimensions across all levels of complexity, sample size, and correlation. DETECTe was particularly robust in conditions with the high correlations among the dimensions, where it only suffered to larger extent at 50% complexity with any sample size.

ALR suffered in accuracy of selection as early as .60 correlation and 30% of complexity for all sample sizes. $\chi^2_{G/D}$ tended to correctly identify the true dimensional structure only in conditions with correlation of .30 or lower and 30% or lower complexity levels. As the sample size increased, within the correlational level, $\chi^2_{G/D}$ generally yielded lower proportions correct.

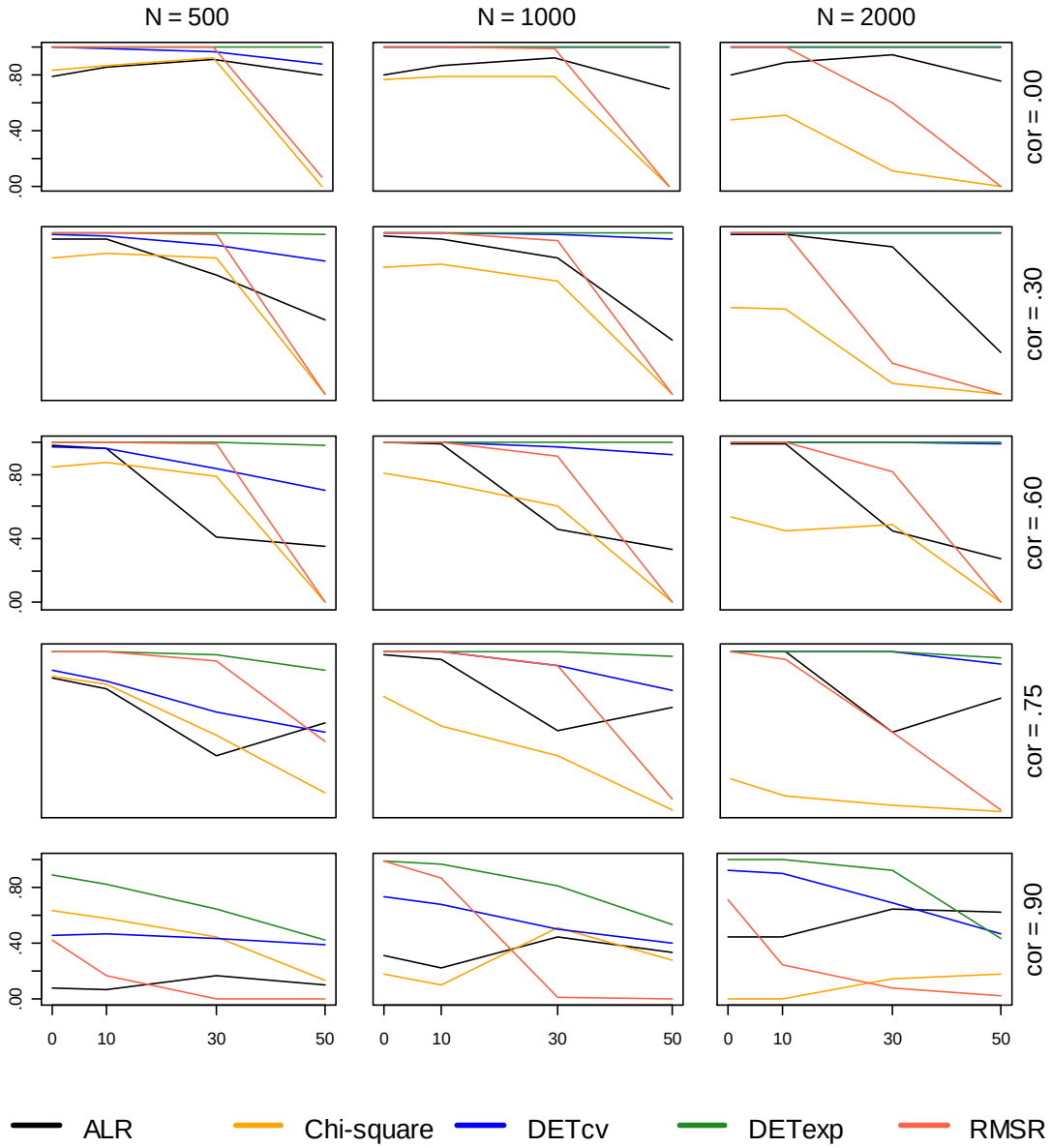

Figure 21. Proportion correct when the data follow a compensatory 3D MIRT model with 20 items per dimension.

The proportion of dimensional labeling. Figure 22 plots the marginal proportions across 500 replications that a method labeled three, two, one, or none of the sets of items as dimension-like (x-axis) when the data exhibit 30% complexity and follow a compensatory 3D MIRT model with 20 items per

dimension (conditions with 0% and 10% complexity yielded similar results to 30% complexity; for the remaining with 0% and 10%, see Appendix B).

In conditions across sample size and with correlation of .75 or smaller, the NOHARM-based methods were most likely to label three sets of items as dimension-like, while the DETECT-based methods tended to have somewhat lower rates for labeling three sets of items as dimension-like. The DETECT-based methods had higher proportions of labeling one set of items as dimension-like than the NOHARM-based methods. Generally, all methods were successful in identifying three sets of items as dimension-like when the data exhibited 30% or less complexity across sample size and correlation levels of .75 or less (except *ALR*, whose performance diminished at .75 correlation and $N = 500$). With correlations of .90, the methods had some success in labeling mostly either two or one set of items as dimension-like.

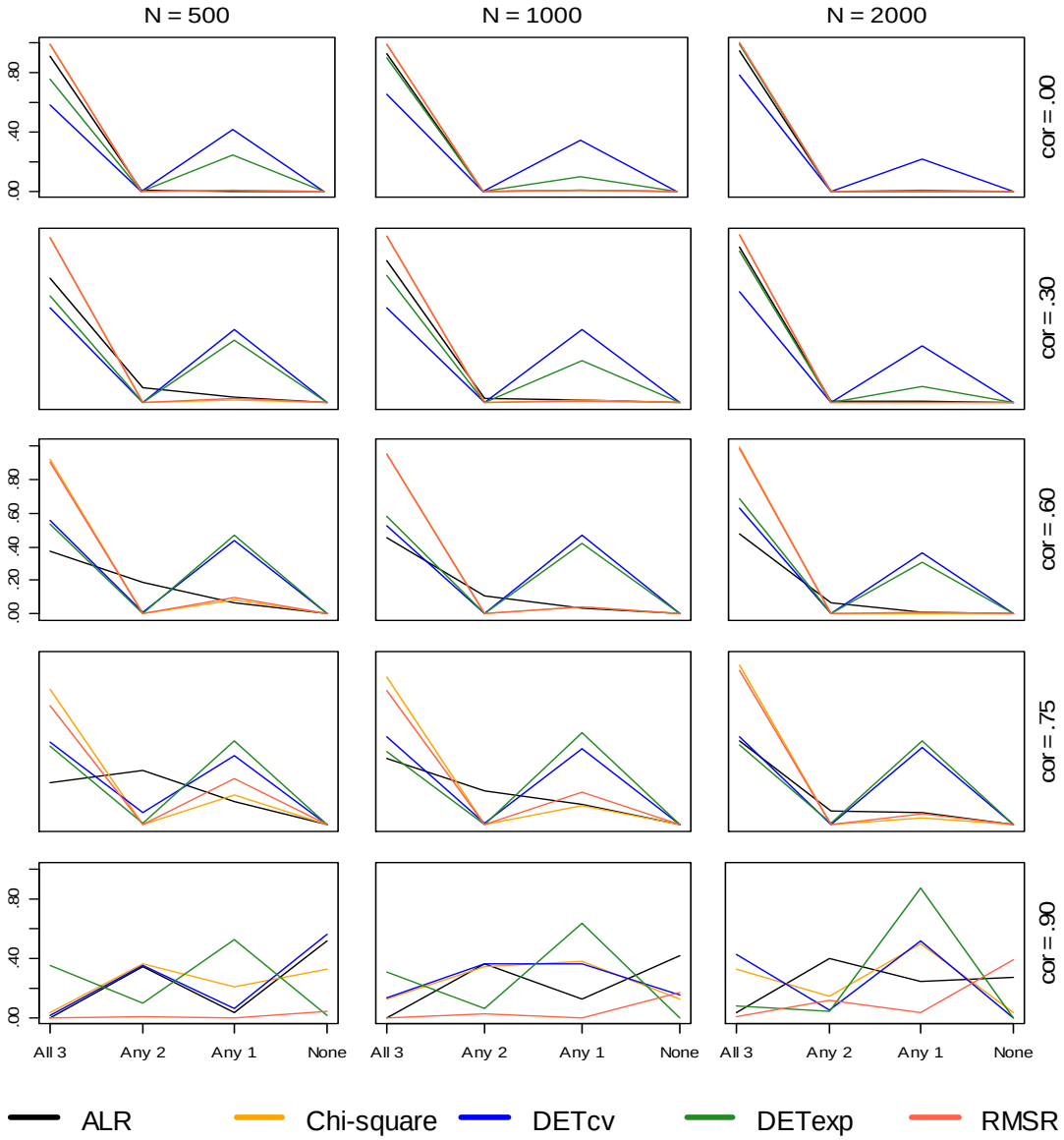

Figure 22. Marginal proportions across 500 replications that a method identified three, any two (both), (any) one, or none of the sets of items as dimension-like (x-axis) when the data exhibit 30% complexity and follow a compensatory 3D MIRT model with 20 items per dimension.

Figure 23 plots the marginal proportions across 500 replications that a method identified three, any two (both), (any) one, or none of the sets of items as

dimension-like (x-axis) when the data exhibit 50% complexity and follow a compensatory 3D MIRT model with 20 items per dimension.

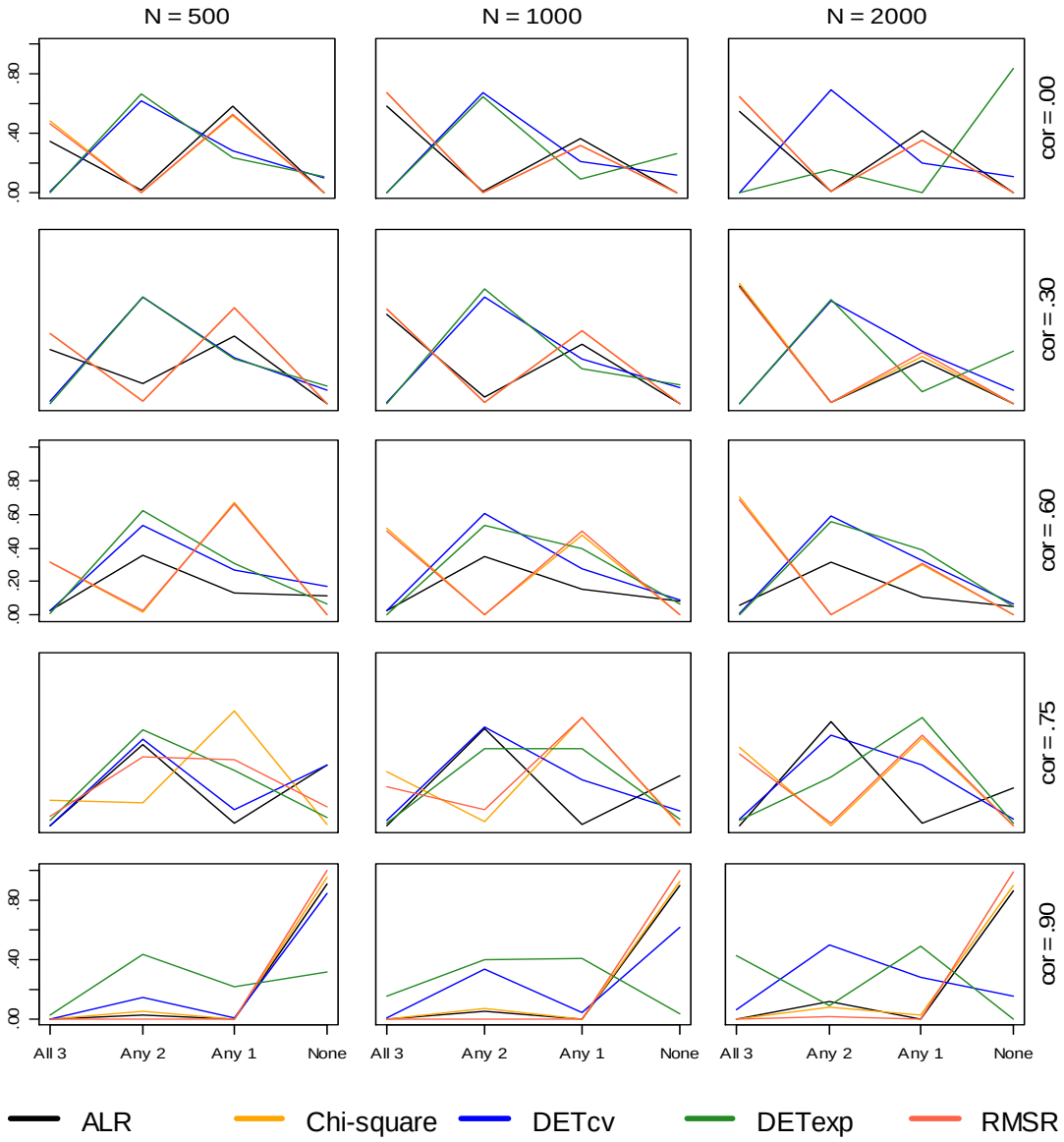

Figure 23. Marginal proportions across 500 replications that a method identified three, any two (both), (any) one, or none of the sets of items as dimension-like (x-axis) when the data exhibit 50% complexity and follow a compensatory 3D MIRT model with 20 items per dimension.

In conditions with N = 500 and correlation of .75 or lower, the NOHARM-based methods tended to successfully label one set of items as dimension-like and the DETECT-based methods tended to label any two sets of items as dimension-like. Within a correlation level, as the sample size increased, $\chi^2_{G/D}$and RMSR (and to some extent *ALR*) yielded higher marginal proportions for labeling of three sets of items as dimension-like. DETECT methods failed to label three sets as dimension-like across all correlation and sample size levels.

The consistency of item classification. Figure 24 plots the classification consistencies for factorially simple items across complexity levels (x-axis) when the data follow a compensatory 3D MIRT model with 20 items per dimension. From Figure 24, it was observed that classification of factorially simple items with 20 items per dimension resulted in DETECT-based methods obtaining high classification rates (above .95) for complexity levels of 30% or less. However, at 50% complexity in the data, DETECTe and DETECTcv reported lower classification rates. This was observed consistently across both the sample size and correlation levels of .75 or lower.

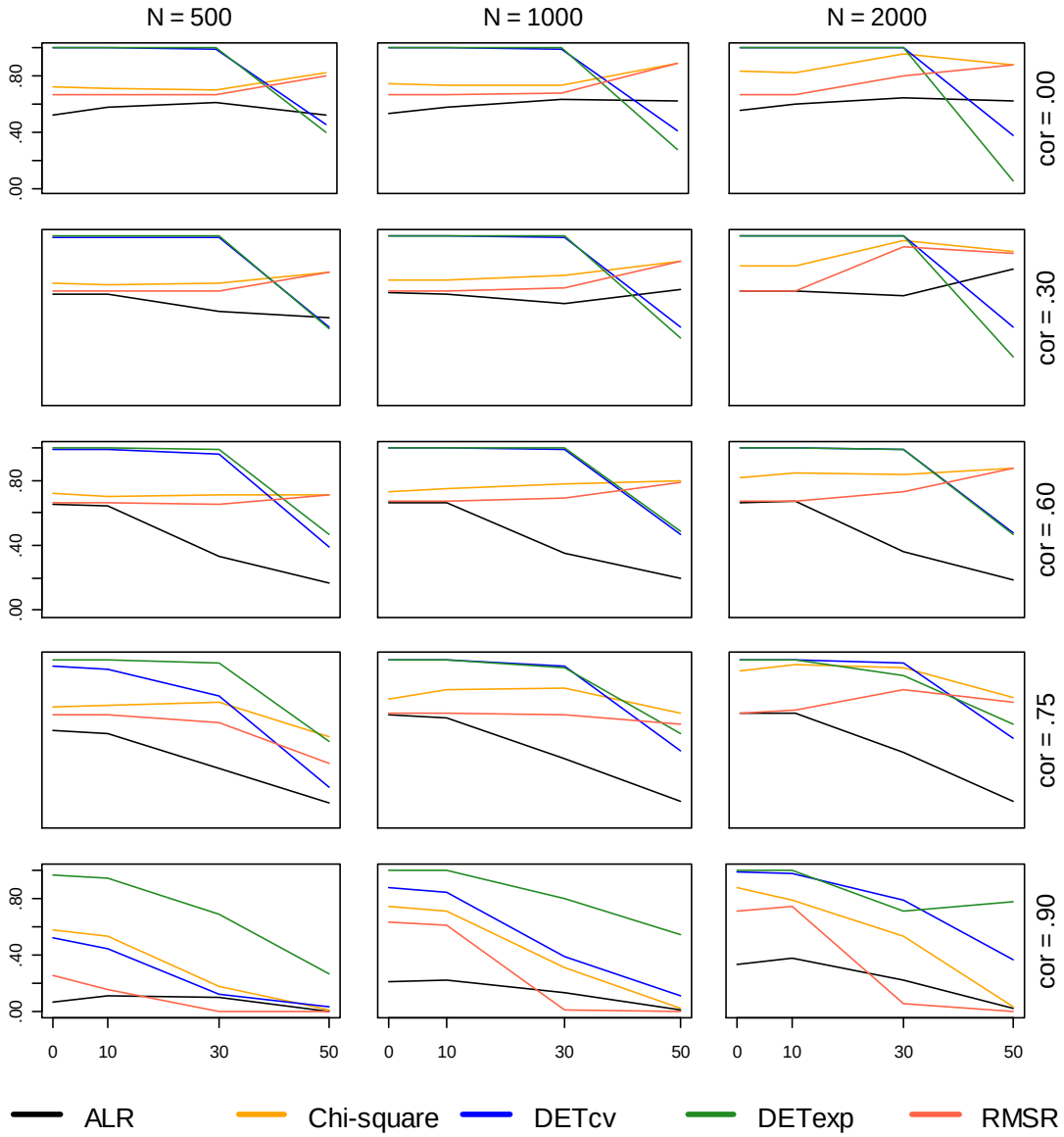

Figure 24. Consistency of factorially simple items when the data follow a compensatory 3D MIRT model with 20 items per dimension.

Of the three NOHARM-based methods, $\chi^2_{G/D}$ and RMSR reported higher rates than *ALR* for most of the conditions. Interestingly, the level of complexity in the data or the sample size had little effect on the methods' classification, as they remained between .74 and .89 for different levels of correlations (up to .75). *ALR*,

however, was notably affected by the correlation level and complexity, as its classification rates decreased greatly at correlations of .60 and complexity levels beyond 10%.

Figure 25 plots the classification consistencies for factorially complex items across complexity levels (x-axis) when the data follow a compensatory 3D MIRT model with 20 items per dimension. The DETECT-based methods had higher consistency rates for complex items than methods based on NOHARM output only at complexity levels of 50% for all sample sizes and correlation levels of .75 or lower. They also more consistently classified items at .90 correlations across sample size levels at 30% and 50% complexity. DETECTcv had notably higher classification rates in at 50% complexity and $N = 500$ at .96, while DETECTe performed similarly when $N = 2000$ with classification rate of .89.

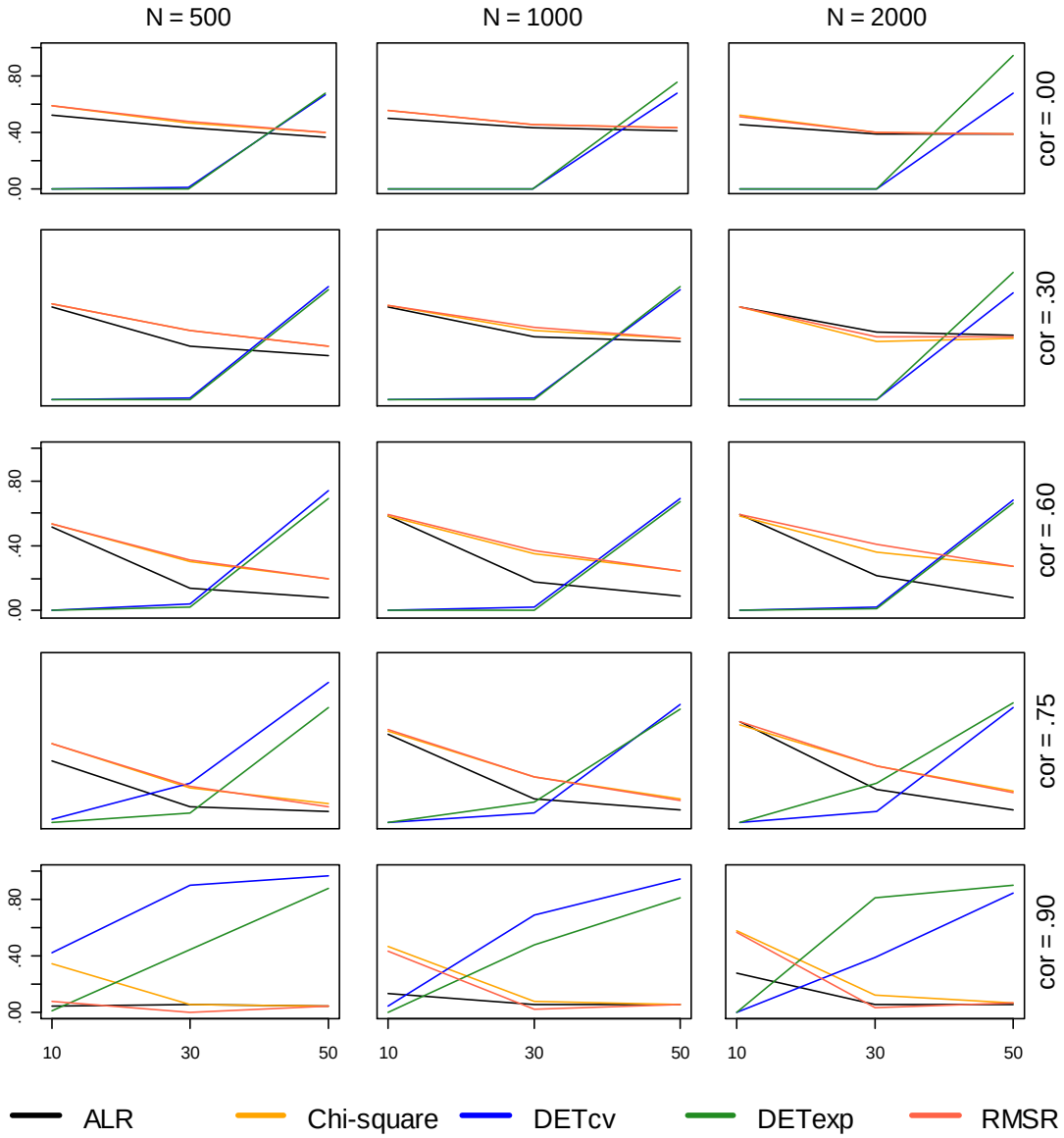

Figure 25. Consistency of factorially complex items when the data follow a compensatory 3D MIRT model with 20 items per dimension.

Overall, it can be concluded that methods' classification rates somewhat depended on the correlation level along with the complexity level. With lower complexity levels, and with correlation of no larger than .75, NOHARM-based methods reported higher classification rates than DETECT-based methods,

although those rates were never higher than .60, As complexity of data increased to 50%, the NOHARM-based methods were unable to consistently classify items across conditions, and their rates dropped to essentially zero.

Effects due to the number of items on determining correct dimensionality. The preceding presentation has displayed results separately by the number of items associated with each dimension. Additional plots were conducted to illustrate the effects of the number of items on the method's abilities to obtain the correct number of dimensions. Figures 26 through 31 correspond to analyses of the effects for varying the number of items for all sample size levels and dimensional structures. The figures plot the proportion of times within a condition (i.e., out of 500 replications) that each method accurately selected the correct dimensional structure in compensatory models.

In the graphs, the y-axis ranges from 0 to 1 and represents the proportion of replications that the method correctly identified the true number of dimensions. The x-axis denotes having 10 and 20 items per dimension. Connected lines on the graphs (from 10 to 20 items per dimension) are drawn only for illustration purposes, not to imply any function between the two categories. Within a graph, different colors represent the five methods of interest.

Conditions that follow a 2D compensatory MIRT model were plotted for all sample sizes. Figure 26 plots the proportion correct when the data follow a compensatory, 2D MIRT model for 10 and 20 items per dimension for $N = 500$. It was observed that the differences in methods' performance to identify the correct

dimensional structure when items per dimension increased from 10 to 20 were found in NOHARM-based methods, particularly for RMSR. RMSR reported very low proportions in all conditions with 10 items per dimension. Increasing the number of items resulted in RMSR to perform better, as proportions of correct selection increased greatly. This improvement was found in almost all conditions across all complexity and correlation levels. RMSR did not improve as much or at all in conditions with 50% complexity and correlation ranging between .30 and .75.

An increase in the number of items when N = 500 had the opposite effect on $\chi^2_{G/D}$ in some conditions. When complexity was at 30% or less and correlation was .60 or smaller, $\chi^2_{G/D}$ seemed not to be affected by the increase in the number of items. However, at complexity levels of 50%, with correlations between .00 to .75, an increase in the number of items resulted in worse performance for $\chi^2_{G/D}$. When the correlation was .90, $\chi^2_{G/D}$ showed improvement from 10 to 20 items across all levels of complexity, although the largest differences in improvement were found at higher levels of complexity.

ALR showed only slight improvement as the number of items increased for conditions with correlations of .75 or less, with most notable improvement in conditions with 50% complexity. When the correlation was .90, *ALR* did not seem to benefit from the increase in items (in fact, with 30% complexity, an increase in items resulted in a decrease in proportion correct). The DETECT-based methods

seemed not to be affected much by the increase in the number of items when N = 500.

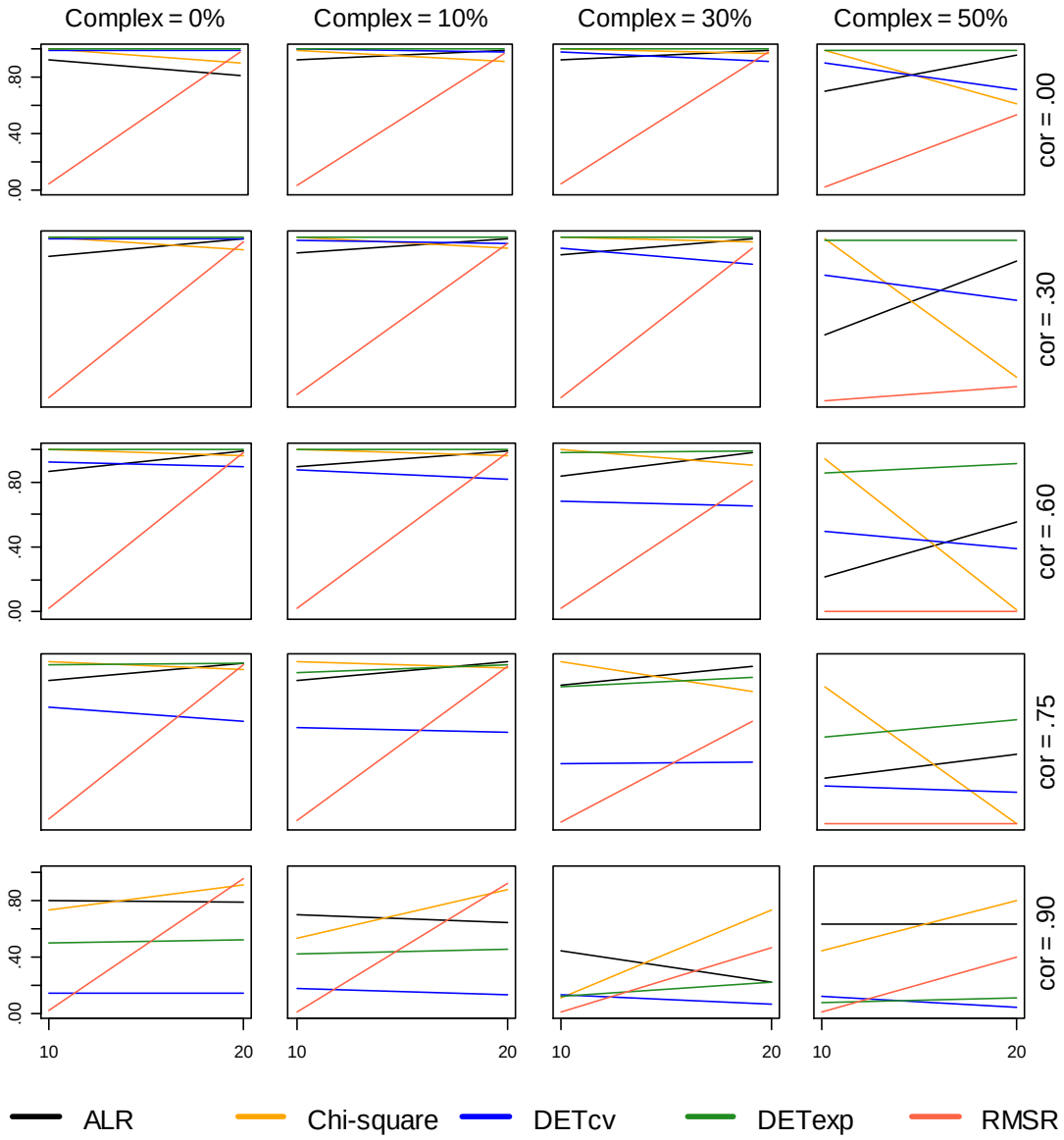

Figure 26. Proportion correct when the data follow a compensatory 2D MIRT model for 10 and 20 items per dimension for N = 500.

Figure 27 plots the proportion correct when the data follow a compensatory, 2D MIRT model for 10 and 20 items per dimension for N = 1000.

In Figure 27, RMSR had patterns similar to those for the previously discussed conditions when N = 500. Within a complexity level, RMSR yielded better performance with 20 items per dimension than with 10 items. This was noted across all correlation levels in conditions with 0%, 10%, and 30% of complexity.

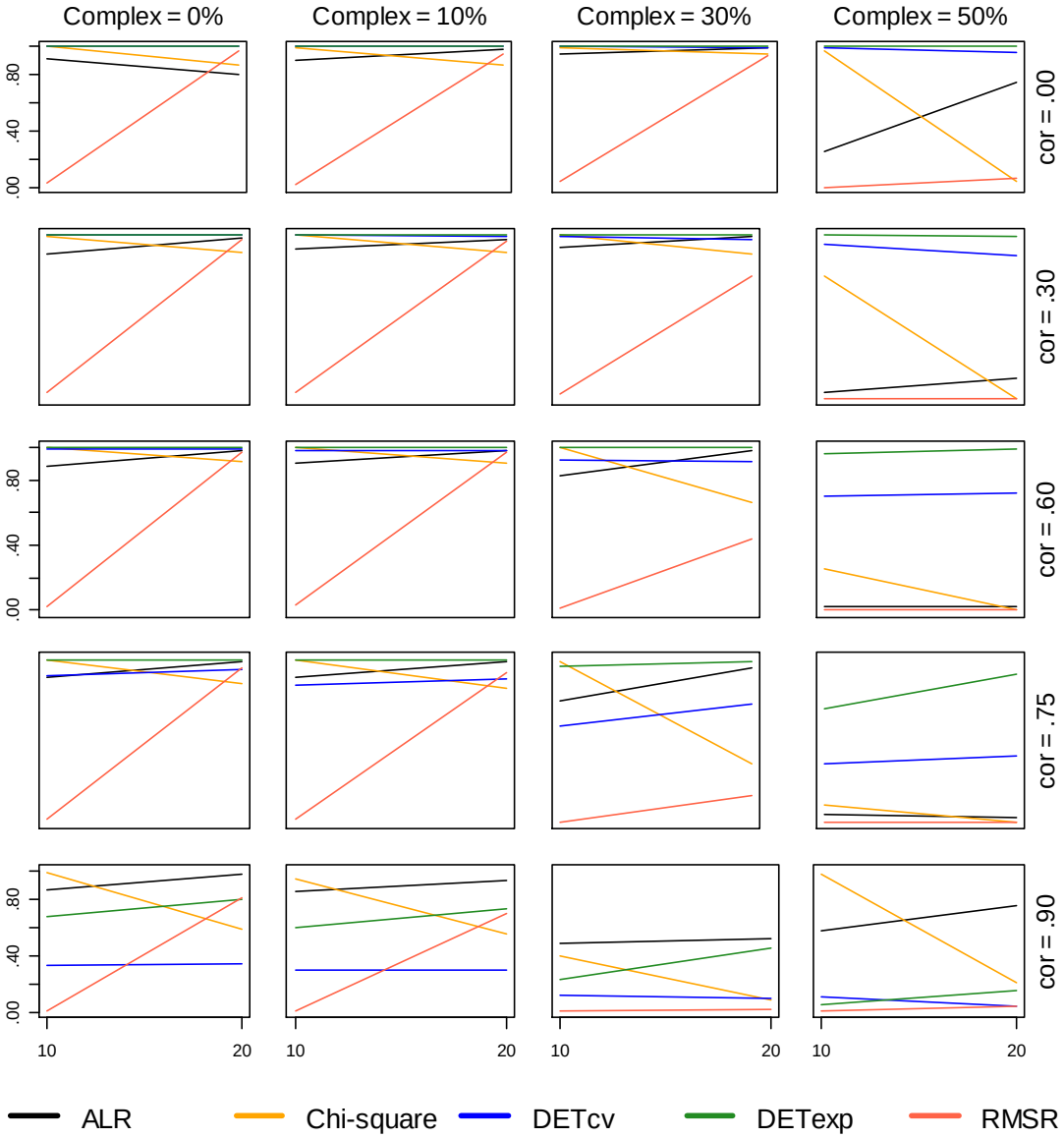

Figure 27. Proportion correct when the data follow a compensatory 2D MIRT model for 10 and 20 items per dimension for N = 1000.

When complexity was at 50%, RMSR performance was poor for both 10 and 20 items per dimension. When N = 1000, an increase in items per dimension resulted in poorer performance of $\chi^2_{G/D}$, particularly when complexity or correlation levels increased. *ALR* as well as the DETECT methods seemed to be only slightly affected by the increase in the number of items.

Figure 28 plots the proportion correct when the data follow a compensatory, 2D MIRT model for 10 and 20 items per dimension for N = 2000. Here, RMSR generally improved in selecting the correct 2D factor solution as the number of items increased; this was particularly found at complexity levels of 0% or 10%. As the complexity level increased to 30%, an increase in the number of items seemed to affect RMSR performance only at low levels of the correlation.

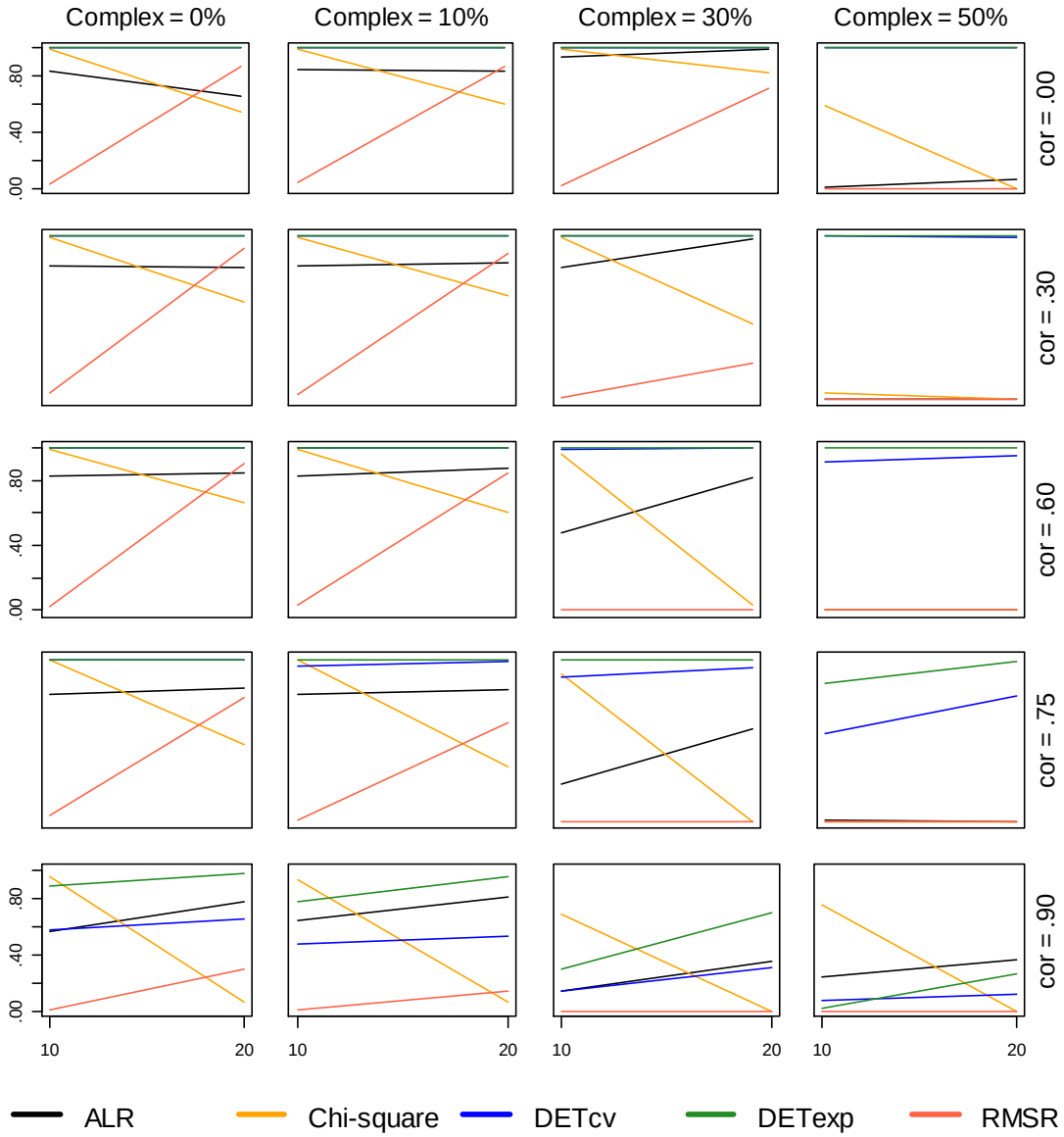

Figure 28. Proportion correct when the data follow a compensatory 2D MIRT model for 10 and 20 items per dimension for N = 2000.

$\chi^2_{G/D}$ never benefited from the increase in items when data followed a compensatory 2D MIRT and N = 2000. It actually performed worse or equally poor across the complexity and correlation levels when the number of items increased. The remaining three methods, *ALR*, DETECTe and DETECTcv, were less affected by the increase in the number of items. Among the three methods,

most notable improvements were found for *ALR* with 30% complexity when correlations were .60 or larger. DETECTcv also showed some improvement when complexity was 30% or larger, and correlation was at .90. RMSR for conditions with low complexity tended to benefit most from the increase in the number of items.

Similar analyses were conducted for conditions in which the data follow a 3D MIRT. Figures 29 through 31 illustrate the effects of increase in the number of items across all levels of complexity and correlation for all sample sizes in 3D cases.

Figure 29 plots the proportion correct when the data follow a compensatory, 3D MIRT model for 10 and 20 items per dimension for $N = 500$. RMSR seemed to be positively affected by the increase in items in 3D conditions as it was in 2D conditions previously discussed. The increase in proportion correct for RMSR was mostly observed when complexity levels were 30% or less. When correlations were at .90, RMSR performed slightly worse when the number of items increased and complexity was at 30% or 50%.

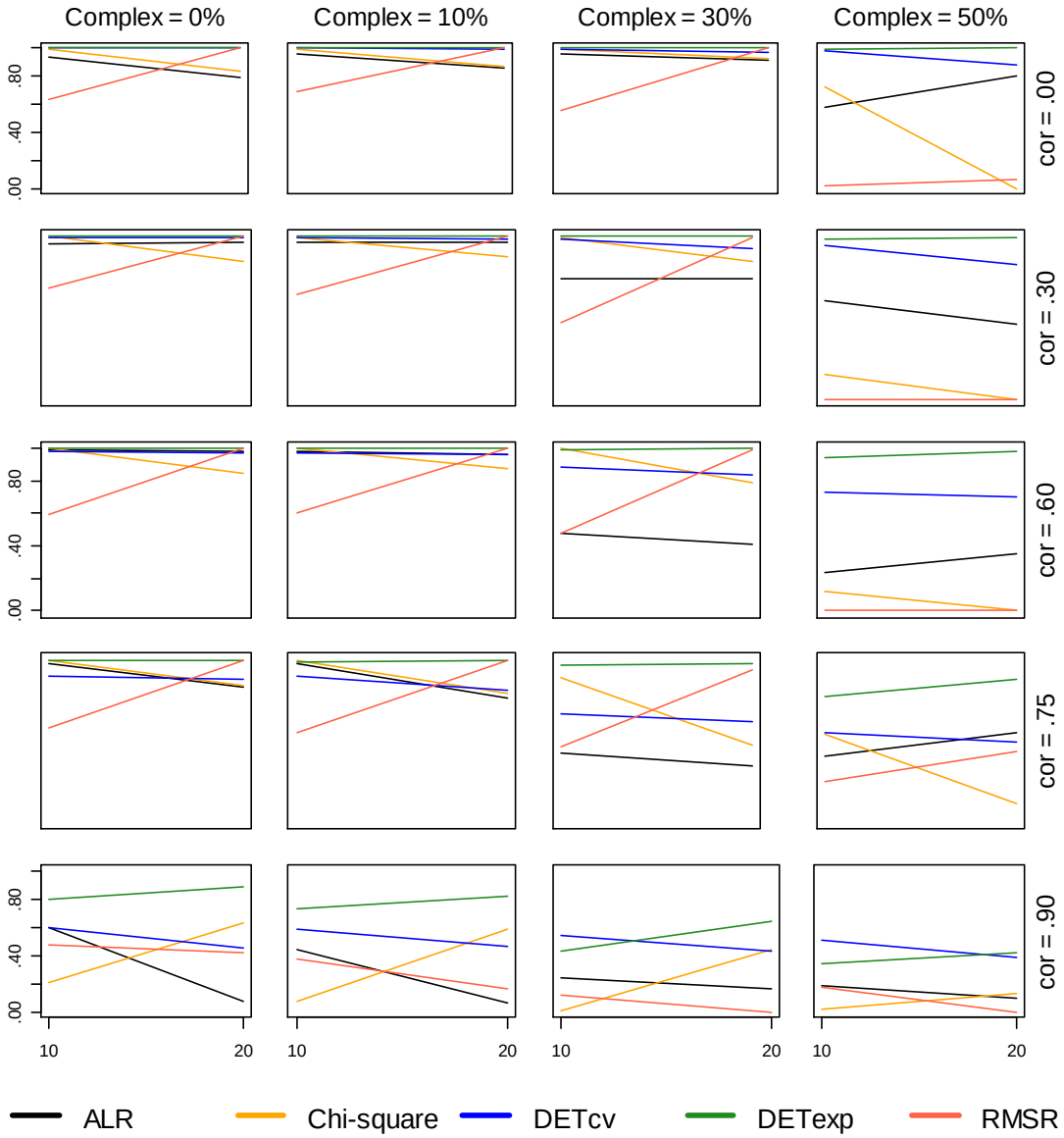

Figure 29. Proportion correct when the data follow a compensatory 3D MIRT model for 10 and 20 items per dimension for N = 500.

ALR and $\chi^2_{G/D}$ were affected less by the increase in items when N = 500. The effect was also not uniform in one direction. *ALR* showed the largest effect in condition where data exhibited 50% complexity and correlation was .75; in this condition, an increase in the number of items had a positive effect on *ALR*'s

performance. However, when complexity was low (e.g., 0% or 10%) and the correlations were .90, an increase in items led to decrease in proportion correct for *ALR*. $\chi^2_{G/D}$performed slightly worse in conditions with more items when correlations were at .75 or smaller. The degree of degradation in performance increased as the complexity levels increased. Increase in the number of items when correlation was .90 however resulted in $\chi^2_{G/D}$ obtaining higher proportion correct (the opposite effect than in conditions with .75 or smaller correlations).

The DETECT-based methods were mostly unaffected by the increase in items when *N* = 500 in the 3D compensatory conditions; DETECTe reported somewhat higher proportion correct in conditions with 20 items when complexity was at 30% and 50% and correlations were .90.

Figure 30 plots the proportion correct when the data follow a compensatory, 3D MIRT model for 10 and 20 items per dimension for *N* = 1000. Similar effects of increase in the number of items were observed in conditions with *N* = 1000 as were noted for the conditions with *N* = 500. For example, DETECT-based methods and *ALR* tended to be only slightly impacted by the increase in the number of items. RMSR tended to be positively impacted by the increase in the number of items for conditions with correlations of .75 or less and complexity levels of 30% or less.

Most notable effects of increased number of items were noted for $\chi^2_{G/D}$method. When *N* = 1000, an increase in items from 10 to 20 per dimension did not result in $\chi^2_{G/D}$ to improve in conditions with low complexity and high

correlations (as it did when N = 500). Similarly, with N = 1000, at 50% complexity and correlation of .00, $\chi^2_{G/D}$ did not perform worse with the increase in the number of items (as it was the case when N = 500).

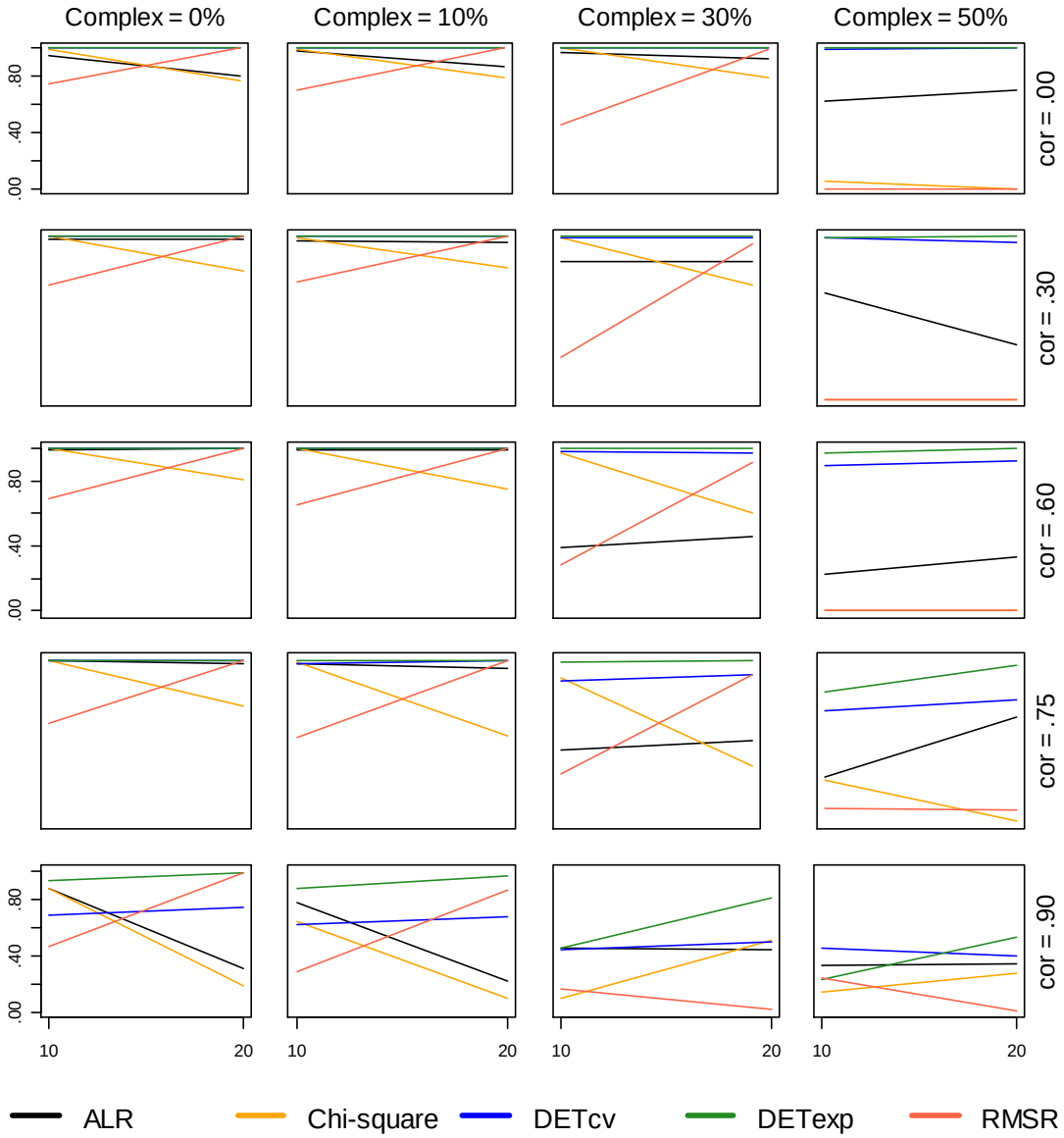

Figure 30. Proportion correct when the data follow a compensatory 3D MIRT model for 10 and 20 items per dimension for N = 1000.

Figure 31 plots the proportion correct when the data follow a compensatory, 3D MIRT model for 10 and 20 items per dimension for $N = 2000$. The effects of the number of items in 3D compensatory conditions were again similar to those in the previously discussed smaller sample sizes. Generally, increases in the number of items per dimension led to increases in proportion correct for RMSR. This was observed for complexity levels of 30% or less. Just the opposite was found for $\chi^2_{G/D}$; an increase in the number of items led to worse performance across the levels of the correlations and for complexity levels of 30% or less.

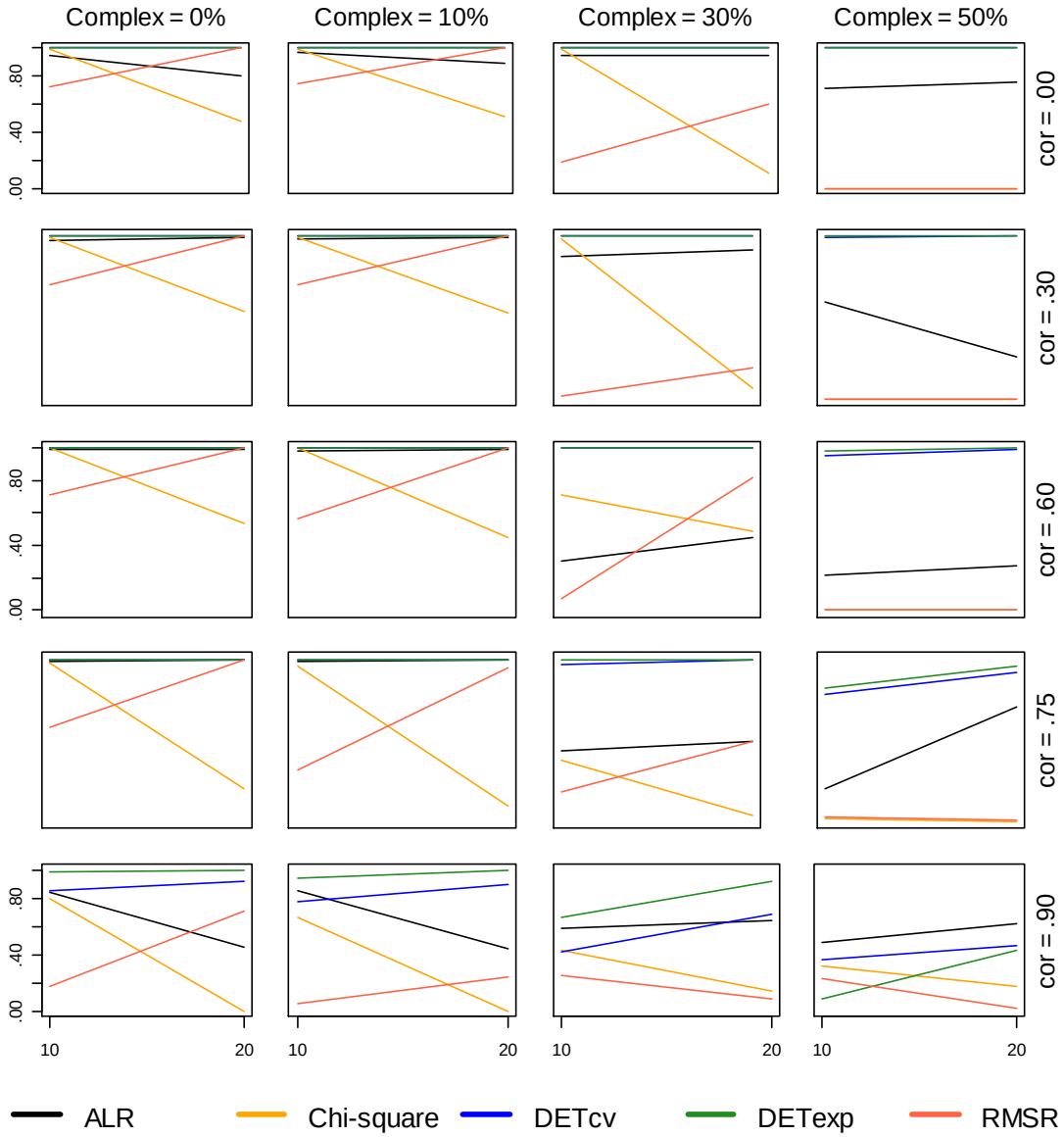

Figure 31. Proportion correct when the data follow a compensatory 3D MIRT model for 10 and 20 items per dimension for N = 2000.

Overall, the DETECT-based methods as well as *ALR* were minimally affected by the number of items when N = 2000. A couple exceptions to that were found. One exception was for *ALR* in the condition with 50% complexity and correlations of .75, where the increase in the number of items led to higher

proportion correct. The other exception involved DETECTe in condition with 50% complexity and .90 correlations, where again, the increase of number of items positively impacted DETECTe performance. Both of these trends however, were previously noted in conditions when $N = 1000$.

Effects due to the number of items on methods' ability to label sets of items as dimension-like. A comparison of results for 2D conditions where data follow compensatory MIRT model suggests that the number of items per dimension did not meaningfully affect the methods proportions of labeling sets of items as dimension-like across all levels of complexity and sample size (e.g., Figure 7 and Figure 17 were compared, as were remaining matching figures for 10 and 20 items per dimension for each complexity level).

A comparison of results for 3D conditions where data follow compensatory MIRT model suggested that the number of items per dimension did not meaningfully affect the proportions of labeling sets of items as dimension-like for the NOHARM-based methods. However, the DETECT-based methods seemed to be positively affected by the increase in items when complexity level was at 30%. In conditions where data exhibited 30% complexity, as correlations and sample size increased, DETECT-based methods increased in proportions of labeling three sets of items as dimension-like in conditions with 20 items per dimension compared to conditions with 10 items per dimension. These effects, again, were only noted in conditions with 30% complexity.

Effects due to the number of items on methods' ability to consistently classify items. A **c**omparison of results for conditions where data follow a 2D compensatory MIRT model suggests that the number of items per dimension did not meaningfully affect the methods constancy rates for factorially simple items. Only two slight effects were noted; RMSR and DETECTe increased their consistency rates for factorially simple items when items per dimension increased from 10 to 20, in conditions with correlation of .90 and $N = 500$ and $N = 1000$ (see Figures 9 and 19). Effects of the increase in number of items on consistency rates for factorially complex items in 2D conditions were very slight (only at .90 correlation and $N = 2000$) and not meaningful. In other words, the methods were not meaningfully affected by the increase in number of items per dimension in conditions with a 2D compensatory MIRT model, across levels of complexity, sample size, and correlations (see Figures 10 and 20).

A comparison of results for conditions where data follow a 3D compensatory MIRT model suggest that increase in number of items did not meaningfully affect methods in their ability to classify factorially simple items. An exception was *ALR*, which yielded lower classification rates of factorially complex items with 20 items per dimension in conditions with correlation levels of .75 or .90 (see Figures 14 and 24). A comparison of classification results for factorially complex items suggested that increase in the number of items had a negative effect on classification rates of DETECT-based methods. Namely, in conditions with a 3D compensatory MIRT model, the DETECT-based methods

yielded higher classification rates when 10 items were associated with each dimension than when there were 20 items per dimension. NOHARM-based methods tended not to be affected by the number of items per dimension when it came to classification of factorially complex items (as seen by comparing Figures 15 and 25).

Noncompensatory multidimensional data.

Tests with ten items per dimension with 2D structures.

The proportion of correct dimensional selection. Figure 32 plots the proportions of times within a condition that a method selected the correct 2D solution across complexity levels. In Figure 32, a strong pattern of performance for the methods emerged. In all but one condition, $\chi^2_{G/D}$ and *ALR* outperformed the other three methods. Large discrepancies in performance were particularly noted when N = 500 and N = 1000 across all levels of complexity and correlations. While maintaining larger proportions of correct selection of the dimensional structure, in N = 2000, the performance of *ALR* and $\chi^2_{G/D}$ shifted downward across all levels of correlation, except when correlation was .90.

In conditions with a correlation of .90, increases in complexity resulted in better performance of the NOHARM-based methods, particularly $\chi^2_{G/D}$. Within a sample size, $\chi^2_{G/D}$ and *ALR* had somewhat uniform performance; $\chi^2_{G/D}$ yielded slightly higher proportions correct in some of the conditions with N = 500 and N = 1000.

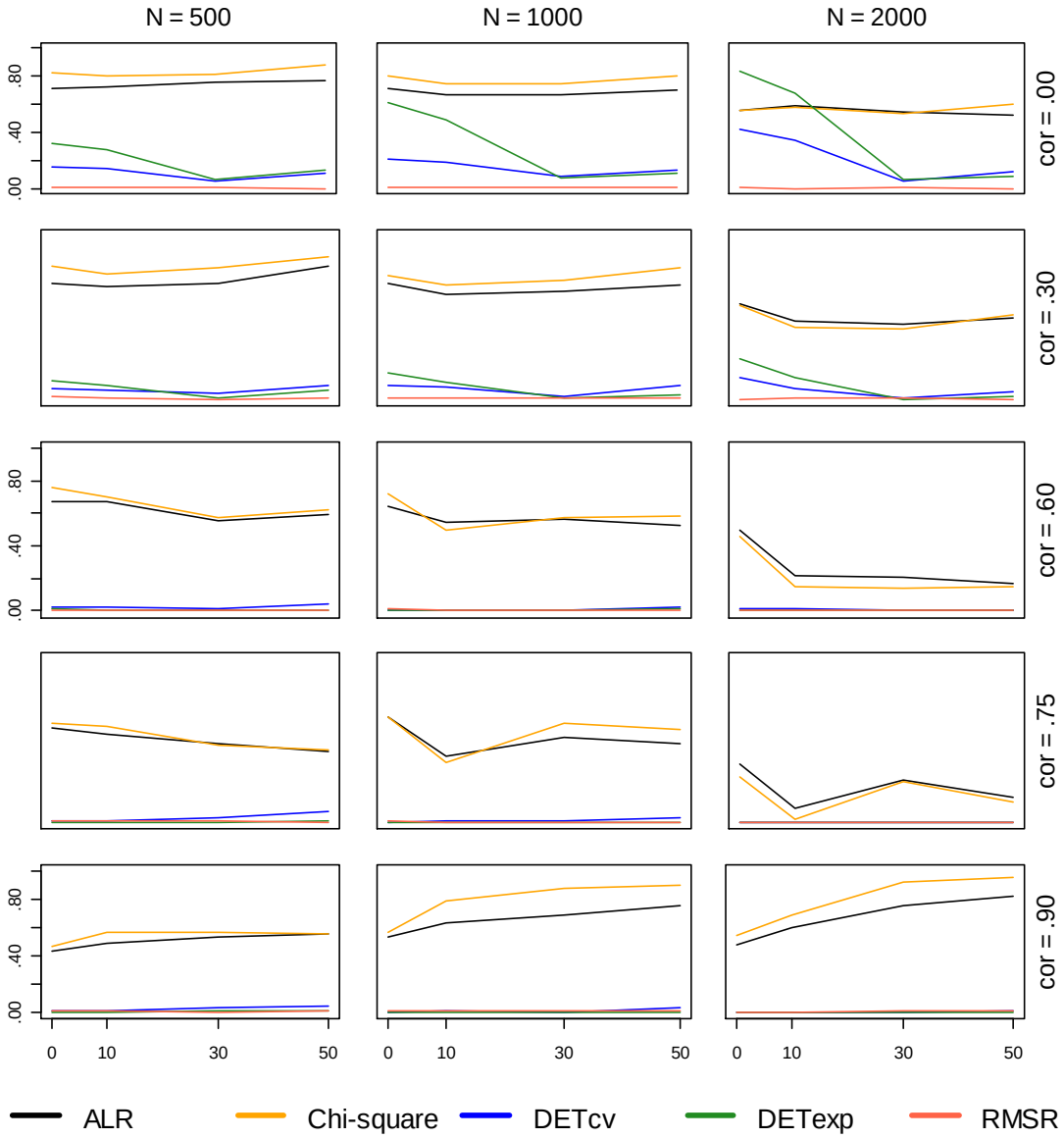

Figure 32. Proportion correct across complexity level when the data follow a noncompensatory 2D MIRT model with 10 items per dimension.

DETECTe yielded the highest proportion correct in conditions with 0% of complexity, correlation of .00, and *N* = 2000 (84% of time it correctly identified the number of factors); however, in the remainder of conditions, DETECTe largely erred. DETECTcv had similar pattern of performance to DETECTe,

except it erred even more often. The two DETECT-based methods and the RMSR method failed to correctly select 2D factor solution in most conditions (note the flatness of the green, blue, and red lines when correlation levels were .30 or larger across complexity and sample size).

The proportion of dimensional labeling. In order to examine performance of the methods further, we computed the marginal proportions of the methods' rates of labeling a set of items as dimension-like. Here again, in 2D conditions, a method could label two, one, or none of the sets of items as dimension-like, regardless of the selection of optimal factor solution. The marginal proportions are calculated across different factor solutions and are plotted for easier identification of patterns. Figures 33 through 36 plot the marginal proportions of the methods' ability to label two (both), (any) one, or none of the sets of items as dimension-like for various levels of complexity when data follow a 2D noncompensatory MIRT with 10 items associated with a dimension.

Figure 33 plots the marginal proportions that each method labeled sets of items as dimension-like for 0% complexity across the sample sizes and correlations. It was observed that when correlation levels were .60 or lower, all the methods except RMSR yielded high marginal proportions for identifying two sets of items as dimension-like, across different sample sizes. Additionally, when N = 2000, *ALR* and $\chi^2_{G/D}$ reported somewhat lower marginal proportions than the DETECT-based methods for these correlation levels. Note that the conditions of correlation of .60 or lower (across sample sizes), are marked by the "L" shaped

lines in the graphs suggesting high proportion for labeling the two sets of items as dimension-like.

At a correlation of .75, the DETECT-based methods, particularly when *N* = 2000, also yielded high marginal proportions for labeling two sets of items as dimension-like (DETECTe yielded higher means that DETECTcv across most conditions). However, the DETECT-based methods had less success in labeling any one of the sets of items as dimension-like in conditions with correlations of .90. As the sample size increased, DETECTcv and DETECTe reported higher marginal proportions for labeling none of the sets of items as dimension-like.

When the correlation was at .75 or .90, RMSR method yielded the highest marginal proportions for identification of one set of items as dimension-like; a pattern that was noted with the other two NOHARM-based methods ($\chi^2_{G/D}$ and *ALR*) at .90 correlation and *N* = 1000 and *N* = 2000.

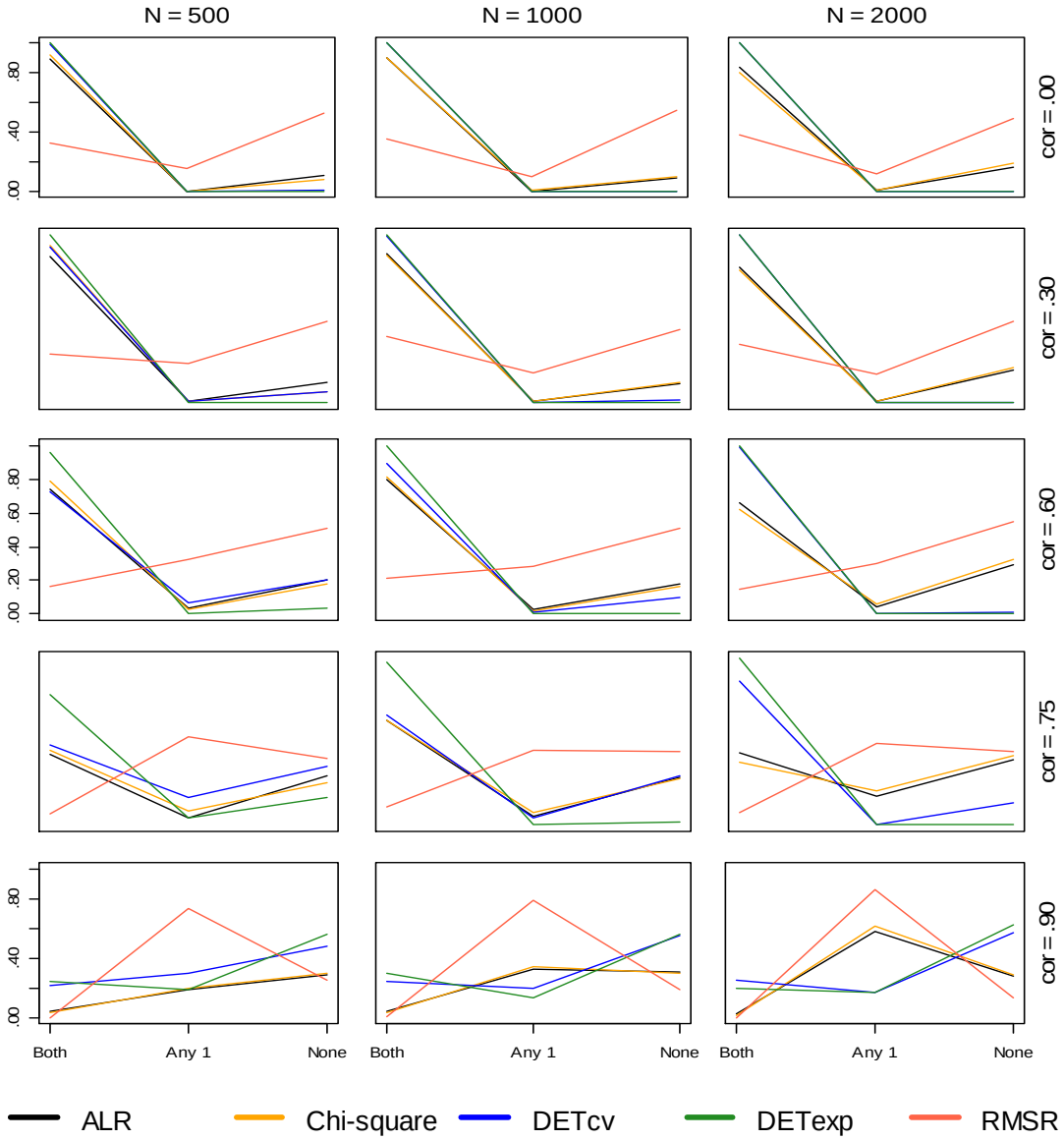

Figure 33. Marginal proportions across 500 replications that a method identified two (both), (any) one, or none of the sets of items as dimension-like (x-axis) when the data exhibit 0% complexity and follow a noncompensatory 2D MIRT model with 10 items per dimension.

Figure 34 plots the marginal proportions that each method identified sets of items as dimension-like for 10% complexity across the sample size and correlations.

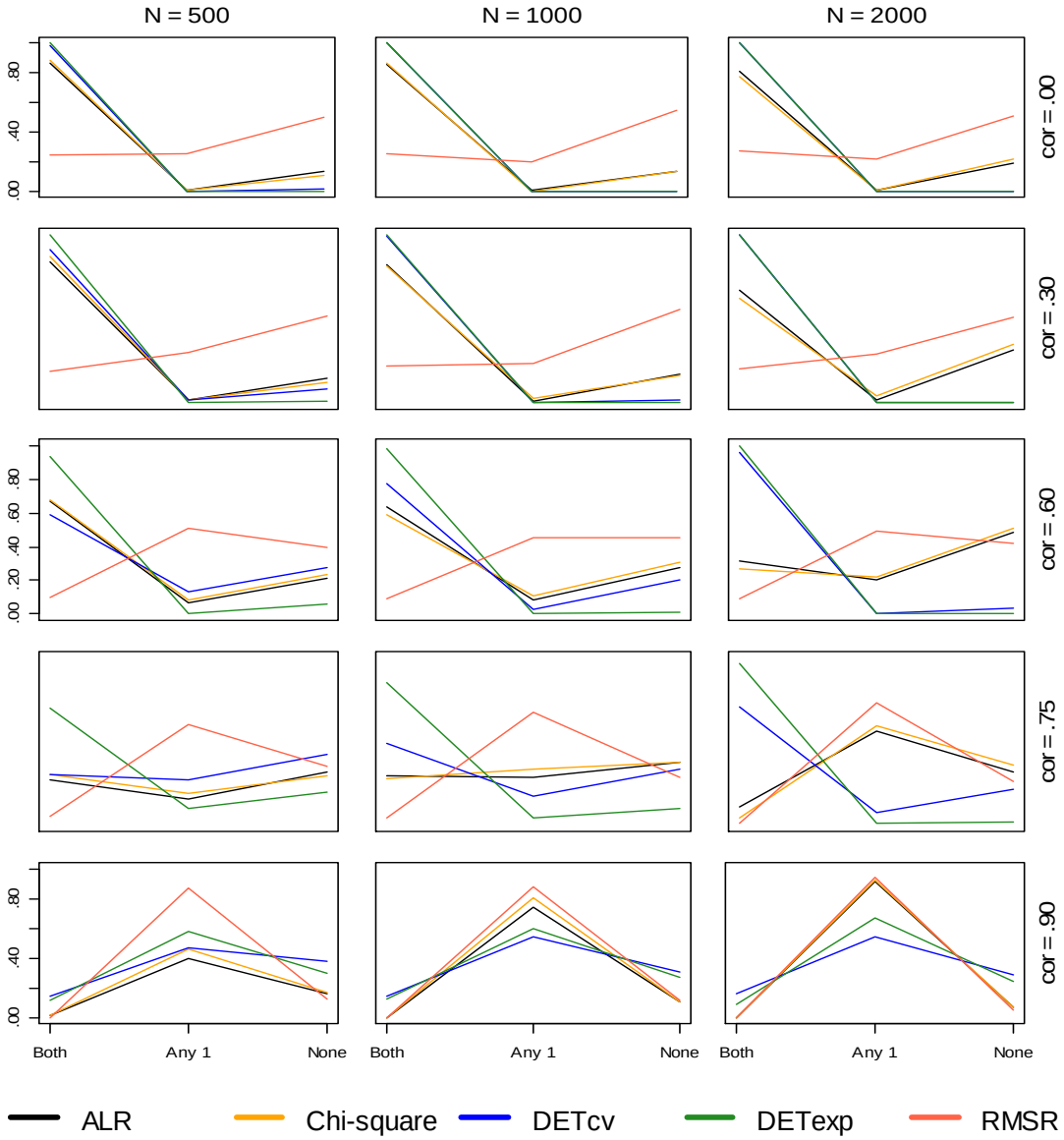

Figure 34. Marginal proportions across 500 replications that a method identified two (both), (any) one, or none of the sets of items as dimension-like (x-axis) when the data exhibit 10% percent complexity and follow a noncompensatory 2D MIRT model with 10 items per dimension.

From Figure 34, it was observed that *ALR* and $\chi^2_{G/D}$ recorded large marginal proportions for identifying two sets of items as dimension-like, in conditions with .30 correlation or less across all three sample sizes (note the "L"

shaped lines represented in the graphs). However as the correlation levels increased, the proportions that $\chi^2_{G/D}$ and *ALR* identified two sets of items as dimension-like decreased; more so when *N* = 2000 than when *N* = 500.

Generally, the DETECT-based methods (especially DETECTe) identified the two sets of items as dimension-like most often in conditions across sample size and correlation of .75 or smaller. RMSR tended to identify two sets of items as dimension-like seldom; it was most successful in labeling any one set as dimension-like in conditions with .60 correlation or higher.

The overall effect of an increase in correlation was observed as well; for all methods, increases in the correlation (up to .75) led to an increase in marginal proportions for none of the sets of items to be labeled as dimension-like. At a correlation of .90, all methods tended to successfully label any one set as dimension-like; marginal proportions increased as the sample size increased (note higher inverted "V" shapes for the conditions in *N* = 2000).

Figure 35 plots the marginal proportions of labeling sets of items as dimension-like for 30% complexity across the sample sizes and correlations. In these conditions, RMSR tended to be the most successful in labeling any one set as dimension-like across all correlation and sample size levels. The DETECT-based methods reported high marginal proportions for identifying two or none of the sets of items as dimension-like in conditions with .00 or .30 correlation across all sample sizes. At correlation of .60, however, the DETECT-based methods decreased in their ability to identify two or any one sets as dimension-like. As the

correlation increased (for all sample sizes, but more so in the conditions with *N* = 2000), the methods tended to yield higher marginal proportions for identifying only one set as dimension-like (note the inverted "V" shapes particularly in conditions with high correlation).

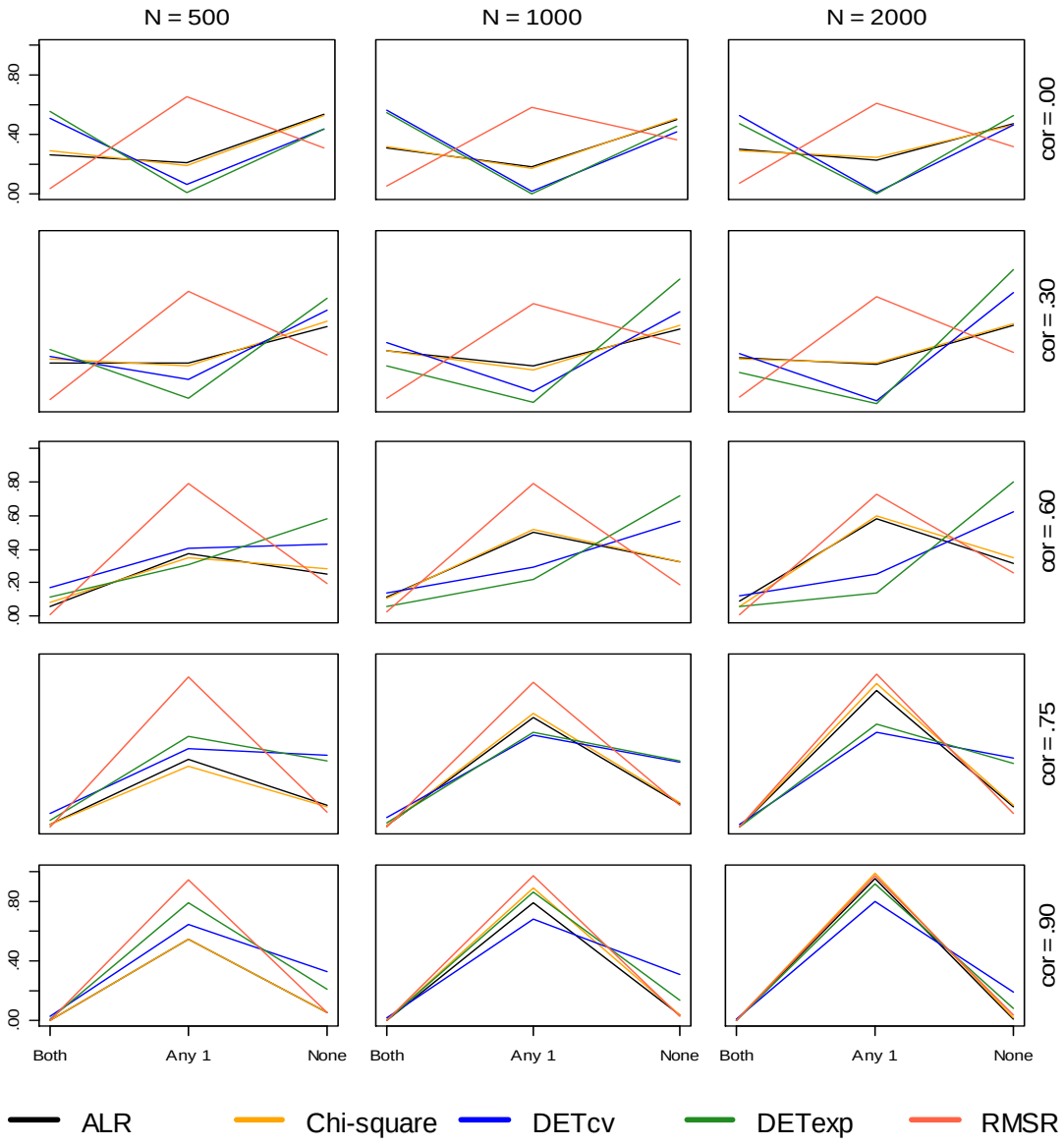

Figure 35. Marginal proportions across 500 replications that a method identified two (both), (any) one, or none of the sets of items as dimension-like (x-axis) when the data exhibit 30% percent complexity and follow a noncompensatory 2D MIRT model with 10 items per dimension.

Figure 36 plots the marginal proportions that each method identified sets of items as dimension-like for 50% complexity across the sample sizes and correlations. All methods, yielded low marginal proportions for labeling two sets of items as dimension-like. The highest marginal proportions were observed for labeling any one set of items as dimension-like. A couple of exceptions were found for the DETECT-based methods, which did not report as high of marginal proportions as the other methods in conditions with high correlations and N = 1000 and N = 2000. Lastly, it was observed that as correlations increased, methods typically reported lower marginal proportions for identifying none of the sets of items as dimension-like.

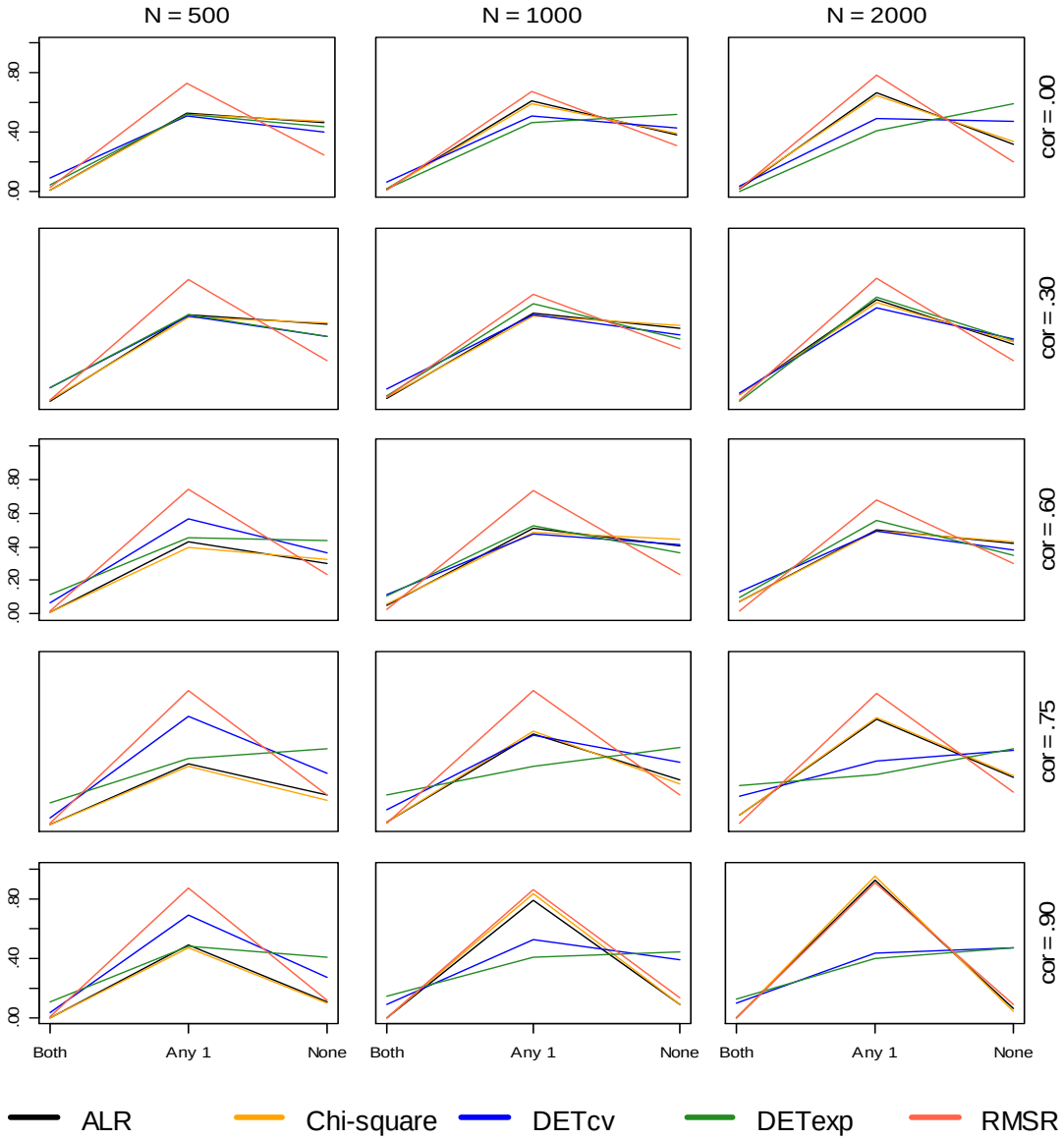

Figure 36. Marginal proportions across 500 replications that a method identified two (both), (any) one, or none of the sets of items as dimension-like (x-axis) when the data exhibit 50% percent complexity and follow a noncompensatory 2D MIRT model with 10 items per dimension.

The consistency of item classification. Figure 37 plots the classification consistencies for factorially simple items across complexity levels (x-axis) when the data follow a noncompensatory 2D MIRT model with 10 items per dimension.

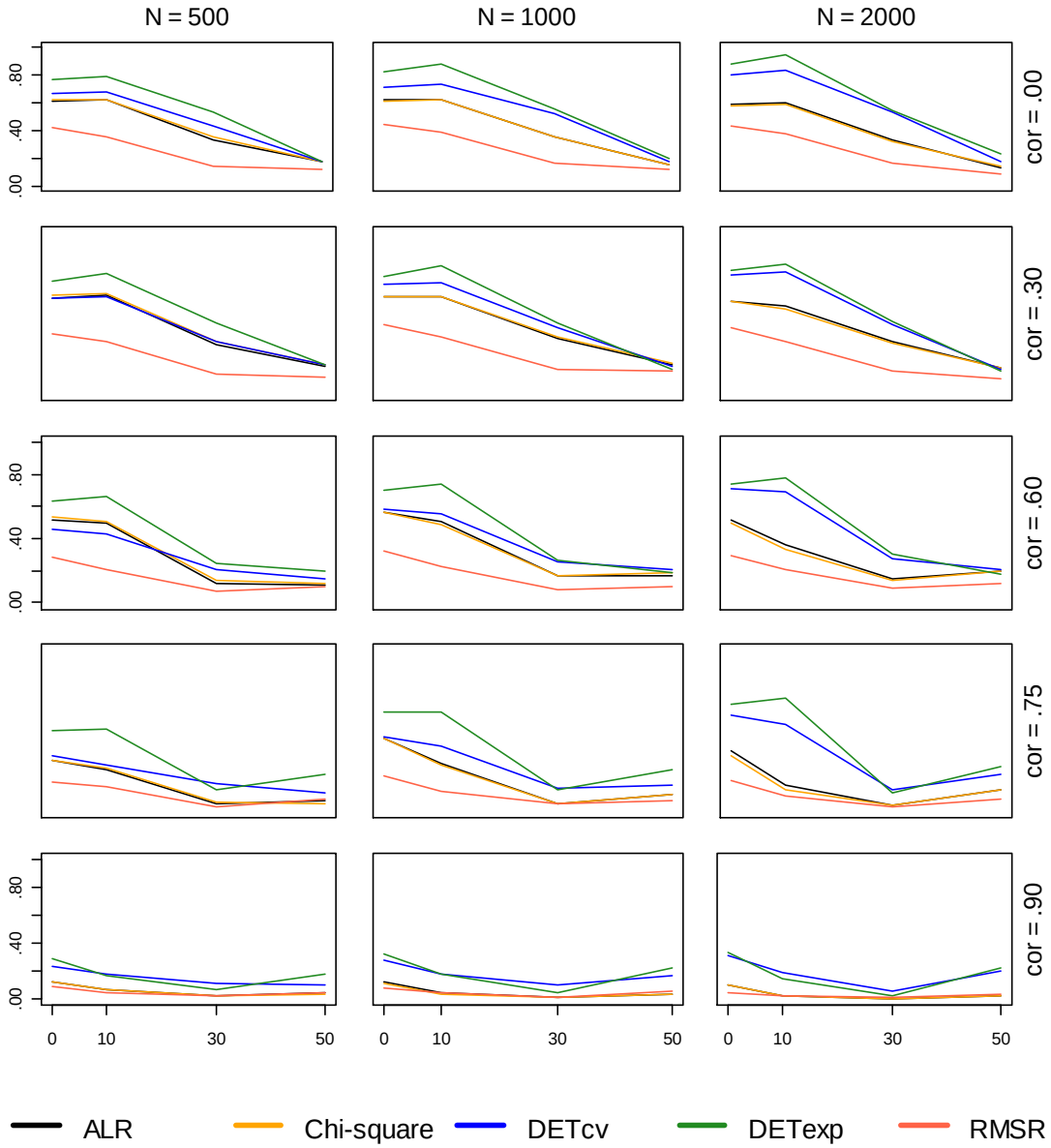

Figure 37. Consistency of factorially simple items across complexity levels when the data follow a noncompensatory 2D MIRT model with 10 items per dimension.

Overall, the DETECT-based methods tended to report higher classification rates when compared to the NOHARM-based methods. In particular, the highest rates were observed for DETECTe in conditions of 0% and 10% of complexity when correlation equaled .00 and .30. DETECTcv followed a similar pattern of

classification to DETECTe, however, larger differences were found in conditions with $N = 500$ between the two methods.

Figure 38 plots the classification consistencies for factorially complex items across complexity levels (x-axis) when the data follow a noncompensatory 2D MIRT model with 10 items per dimension. DETECT-based methods yielded higher classification rates of factorially complex items across sample size correlation levels in conditions with 30% and 50% complexity. At 10% complexity, NOHARM-based methods tended to yield higher classification rates when correlations were .60 or lower. However, as correlations increased, the DETECT-based methods tended to be as or more consistent than the NOHARM-based methods.

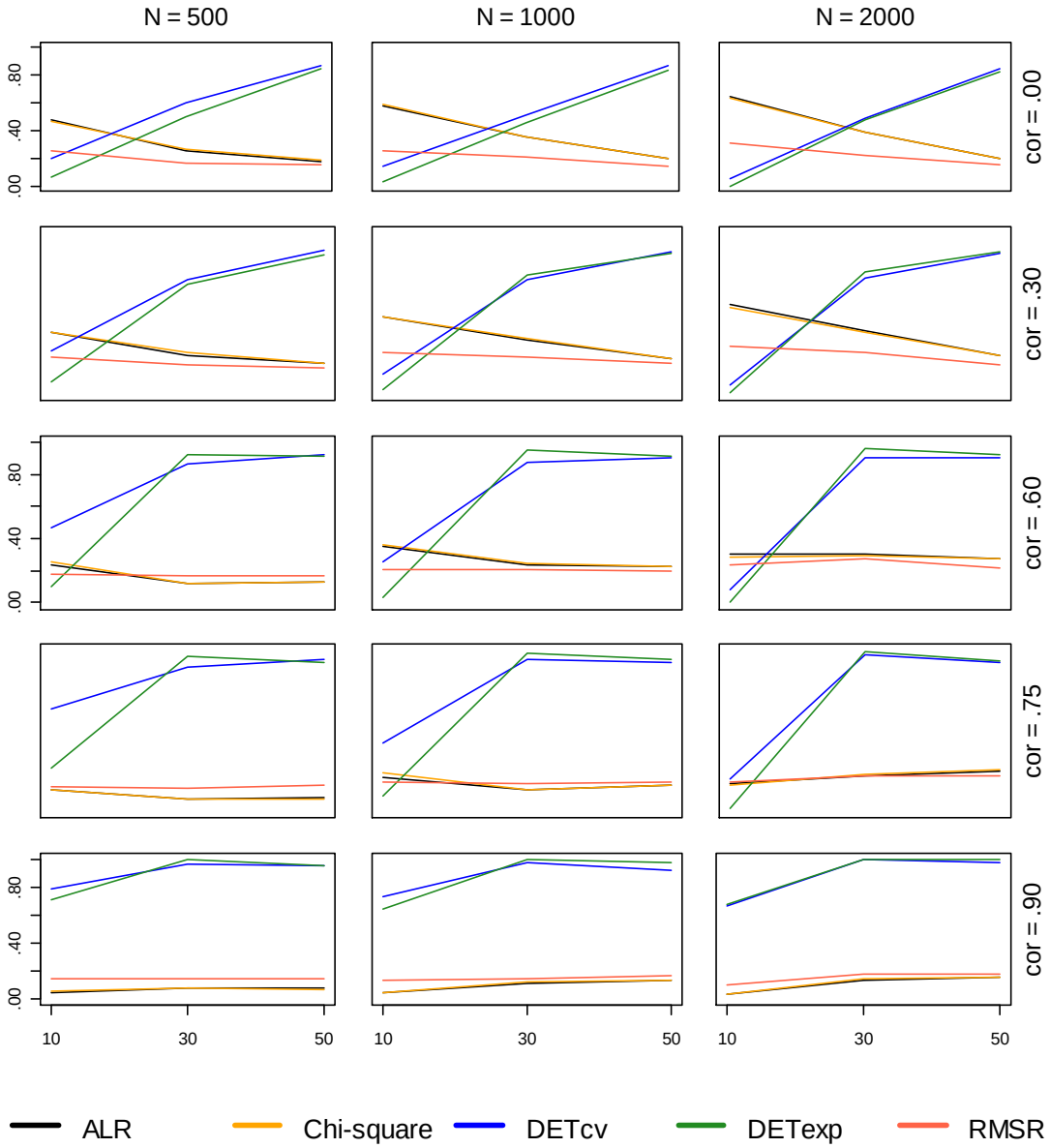

Figure 38. Consistency of factorially complex items across complexity levels when the data follow a noncompensatory 2D MIRT model with 10 items per dimension.

Tests with ten items per dimension in 3D structures.

The proportion of correct dimensional selection. Figure 39 plots proportions of times within a condition that a method selected the correct solution across different levels of complexity (x-axis) when the data follow a

noncompensatory 3D MIRT with 10 items per dimension. It was observed that *ALR* and $\chi^2_{G/D}$ tended to perform better than other methods. In 3D noncompensatory conditions, the methods generally performed better in when *N* = 2000 across different levels complexities and correlations. Generally, low proportions correct were noted for all the methods across different complexity and correlation levels, except $\chi^2_{G/D}$ and *ALR* at 0% and 10% complexity in conditions with *N* = 1000 and *N* = 2000 when correlations were .30 or smaller.

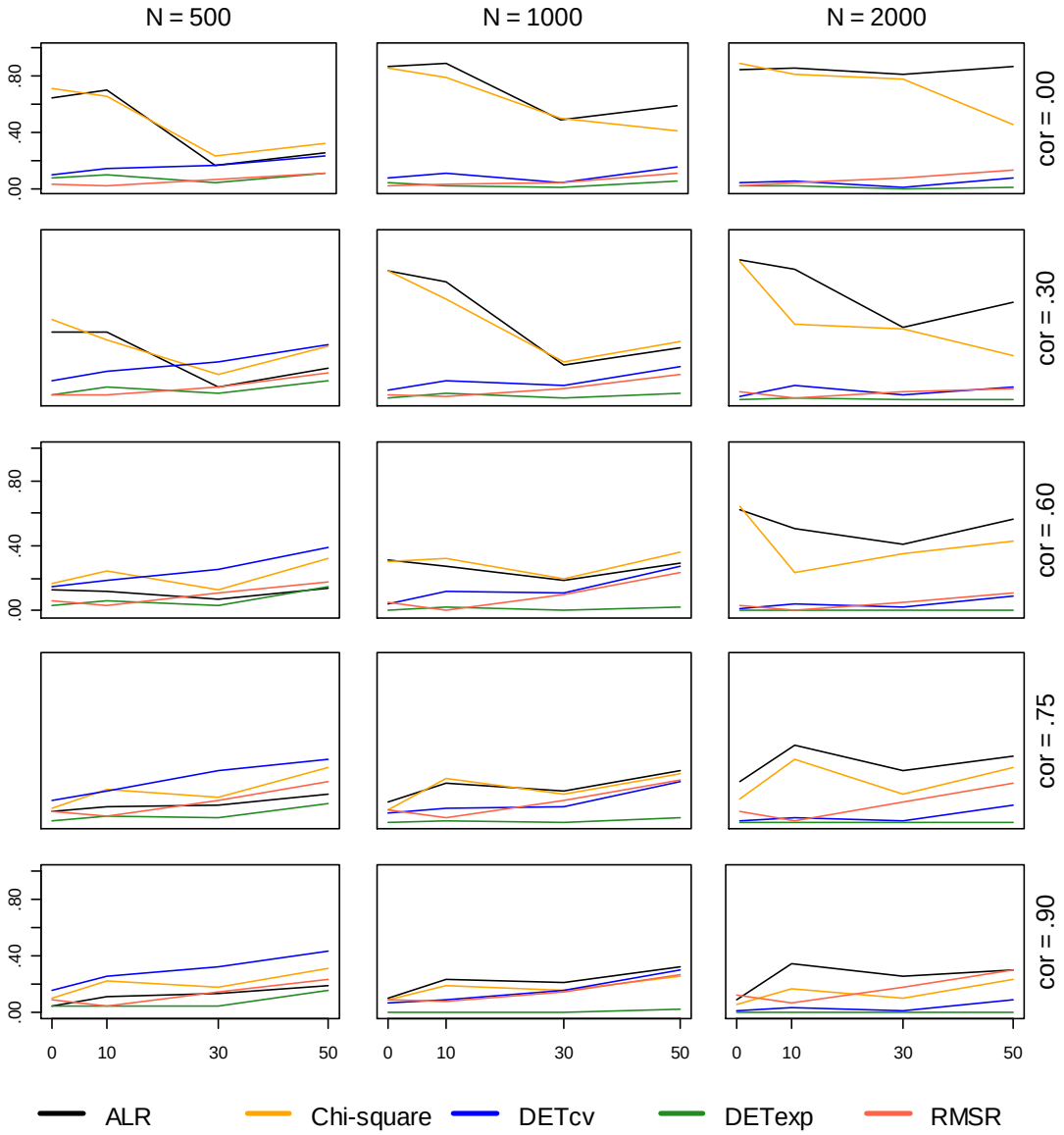

Figure 39. Proportion correct across complexity levels when the data follow a noncompensatory 3D MIRT model with 10 items per dimension.

The proportion of dimensional labeling. In conditions where data follow a 3D noncompensatory MIRT model with 10 items per dimension, complexity levels had a somewhat small effect on how the methods performed with regards to labeling sets of items as dimension-like. To illustrate the main findings in these

conditions, Figure 40 plots the marginal proportions that each method identified sets of items as dimension-like for 50% complexity levels across the sample sizes and correlations (plots for 0%, 10%, and 30% of complexity look very similar and with only a few minor deviations; thus plots for 0%, 10%, and 30% are included in Appendix B).

From Figure 40, it was observed that methods generally reported low marginal proportions for labeling three sets of items as dimension-like. This was noted across the sample sizes, although conditions with $N = 500$ generally reported lower marginal proportions. When $N = 1000$ and $N = 2000$, the highest reported marginal proportions for labeling three sets of items as dimension-like was .49 (DETECTe in a condition with correlation of .00 and $N = 2\,000$).

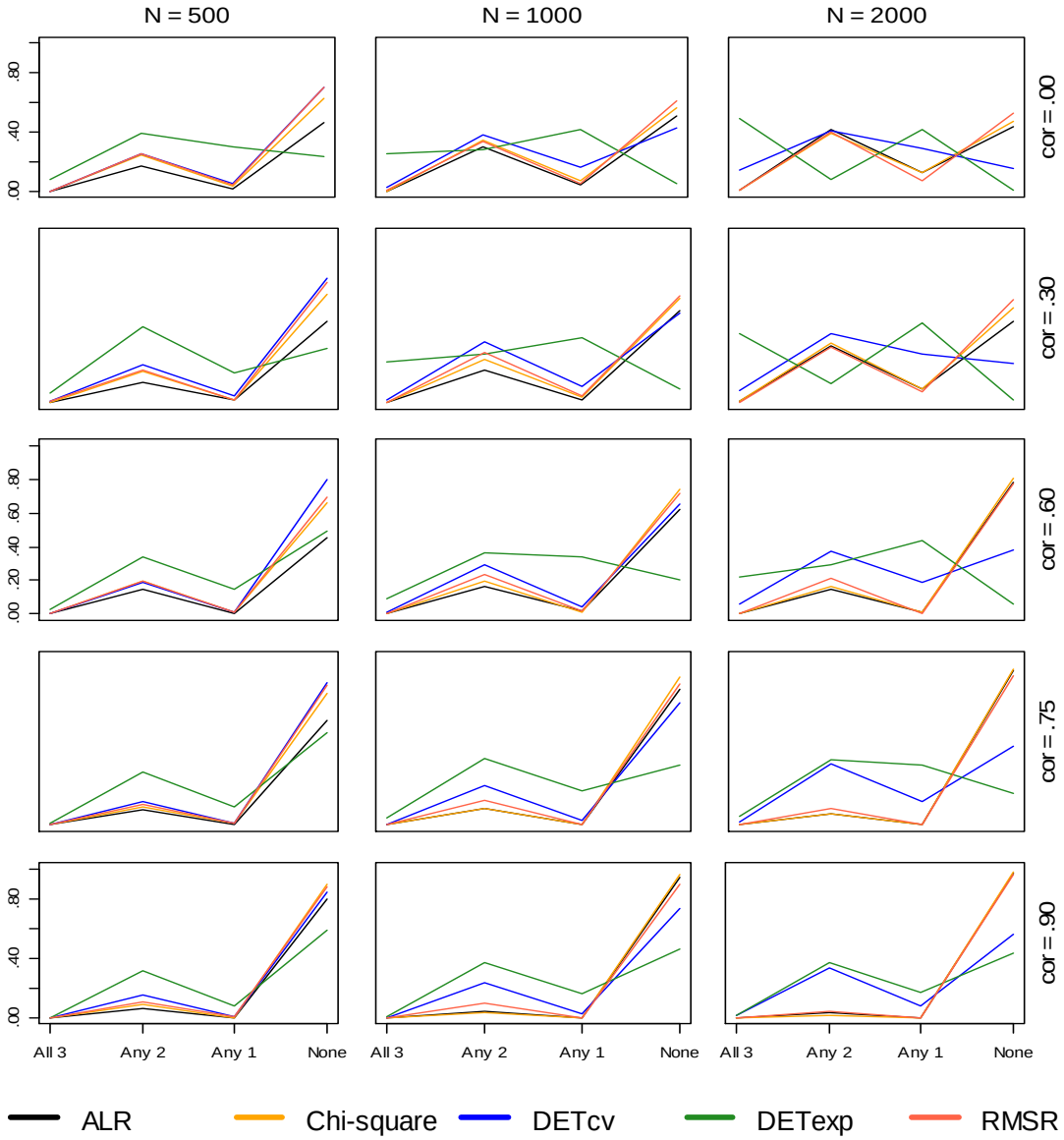

Figure 40. Marginal proportions across 500 replications that a method identified three, any two (both), (any) one, or none of the sets of items as dimension-like (x-axis) when the data exhibit 50% percent complexity and follow a noncompensatory 3D MIRT model with 10 items per dimension.

Typically, the methods tended to report higher marginal proportions for not being able to label any set of items as dimension-like. Some exceptions were

found for DETECTe, in conditions with correlations of .60 or lower and when *N* = 2000.

In conditions with *N* = 500, DETECTe reported higher marginal proportions (compared to other four methods) for labeling any two set of items as dimension-like across all correlation levels. In conditions with *N* = 1000 and *N* = 2000, DETECTe was able to label any one set as dimension-like, while other methods were most successful in identifying any two sets of items as dimension-like (up to .75 correlation). Overall, it was observed that methods generally did not report high marginal proportions for labeling sets of items as dimension-like for any level of complexity.

The consistency of item classification. Figure 41 plots the classification consistencies for factorially simple items across complexity levels (x-axis) when the data follow a noncompensatory 3D MIRT model with 10 items per dimension. For all methods, the classification rates at any level of complexity were somewhat low, particularly as the correlations increased. The two highest classification rates obtained were DETECTe rates in conditions with 0% and 10% complexity, *N* = 2000 with correlation of .00 (.73 and .76, respectively). Also, that the lines within each graph are nearly horizontal, suggests that complexity levels did not have much impact.

The DETECT-based methods reported higher classification rates than the NOHARM-based methods across all conditions, with DETECTe yielding higher rates than DETECTcv. The difference between the DETECT-based methods

decreased as the correlation increased. To some extent, the rates also increased as the sample size increased; particularly for DETECTcv.

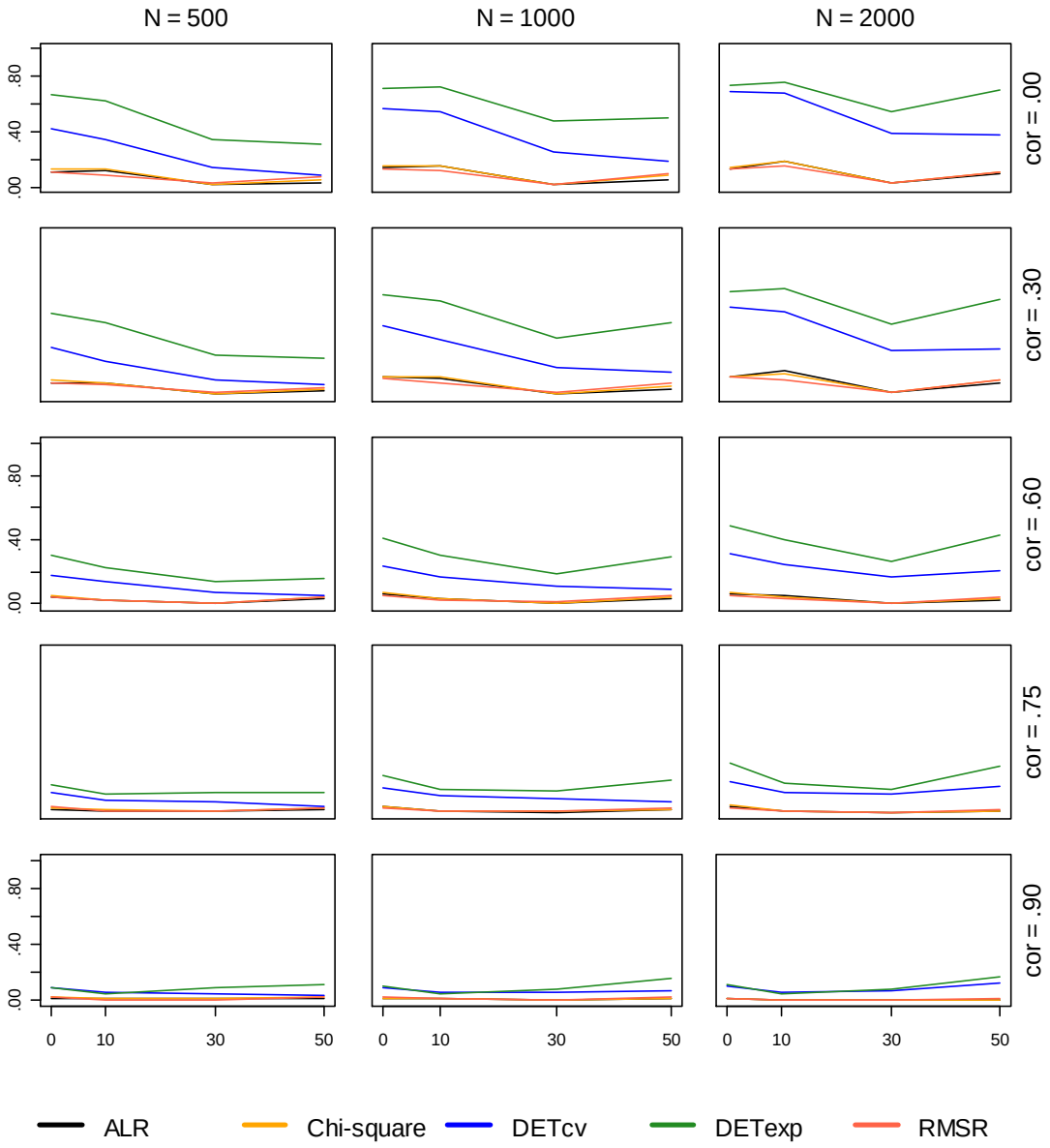

Figure 41. Consistency of factorially simple items across complexity levels when the data follow a noncompensatory 3D MIRT model with 10 items per dimension.

Figure 42 plots the classification consistencies for factorially complex items across complexity levels (x-axis) when the data follow a noncompensatory

3D MIRT model with 10 items per dimension. A distinct behavior for both types of methods was found when it came to classification of the factorially complex items. The DETECT-based methods obtained higher classification rates than the NOHARM-based methods across all sample sizes and correlation levels.

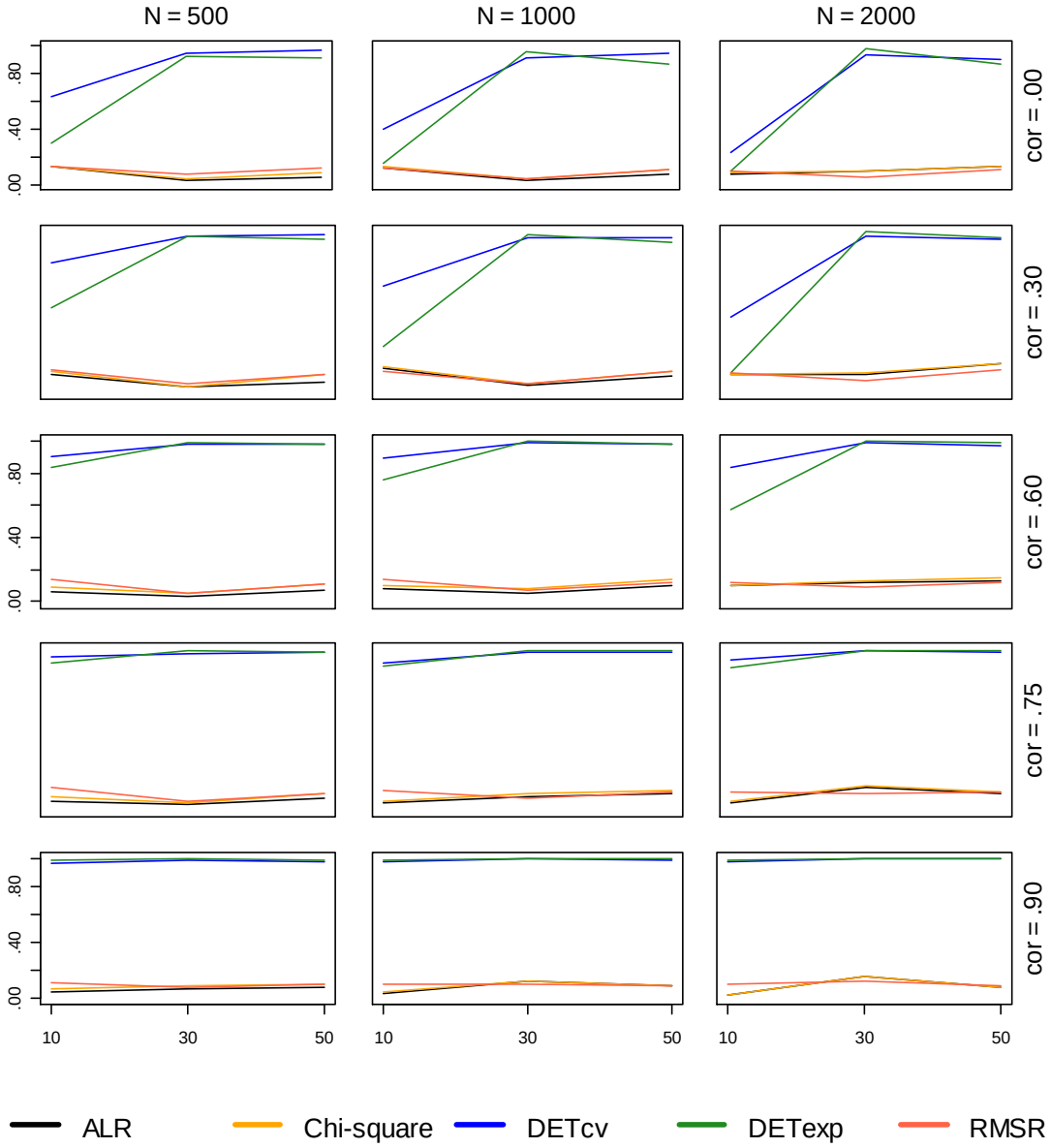

Figure 42. Consistency of factorially complex items across complexity levels when the data follow a noncompensatory 3D MIRT model with 10 items per dimension.

With a few minor exceptions at 0% complexity and lower levels of correlation (i.e., .00 and .30) for the DETECT-based methods, classification rates were stable across the levels of complexity (note the mainly horizontal lines in the graphs). *ALR*, $\chi^2_{G/D}$, and RMSR reported similar classification rates to each other; across all levels of complexity, these rates never rose above .19.

Tests with twenty items per dimension with 2D structures.

The proportion of correct dimensional selection. Figure 43 plots the proportion of times within a condition that a method selected the correct 2D solution across different complexity levels (x-axis) when the data follow a noncompensatory 2D MIRT model with 20 items per dimension.

As illustrated in Figure 43, *ALR* and RSMR had larger proportions of correct selection than either of DETECT-based methods in most conditions. Exceptions were found in conditions with correlation of .00 and N = 1000 and N = 2000, where DETECTe performed equally well or better than other methods across 0% and 10% complexity. DETECTe also had higher proportions correct than DETECTcv although in many of the conditions, both methods performed poorly. Particular poor performance was noted in conditions with increased correlation levels or when more complexity was modeled into the data.

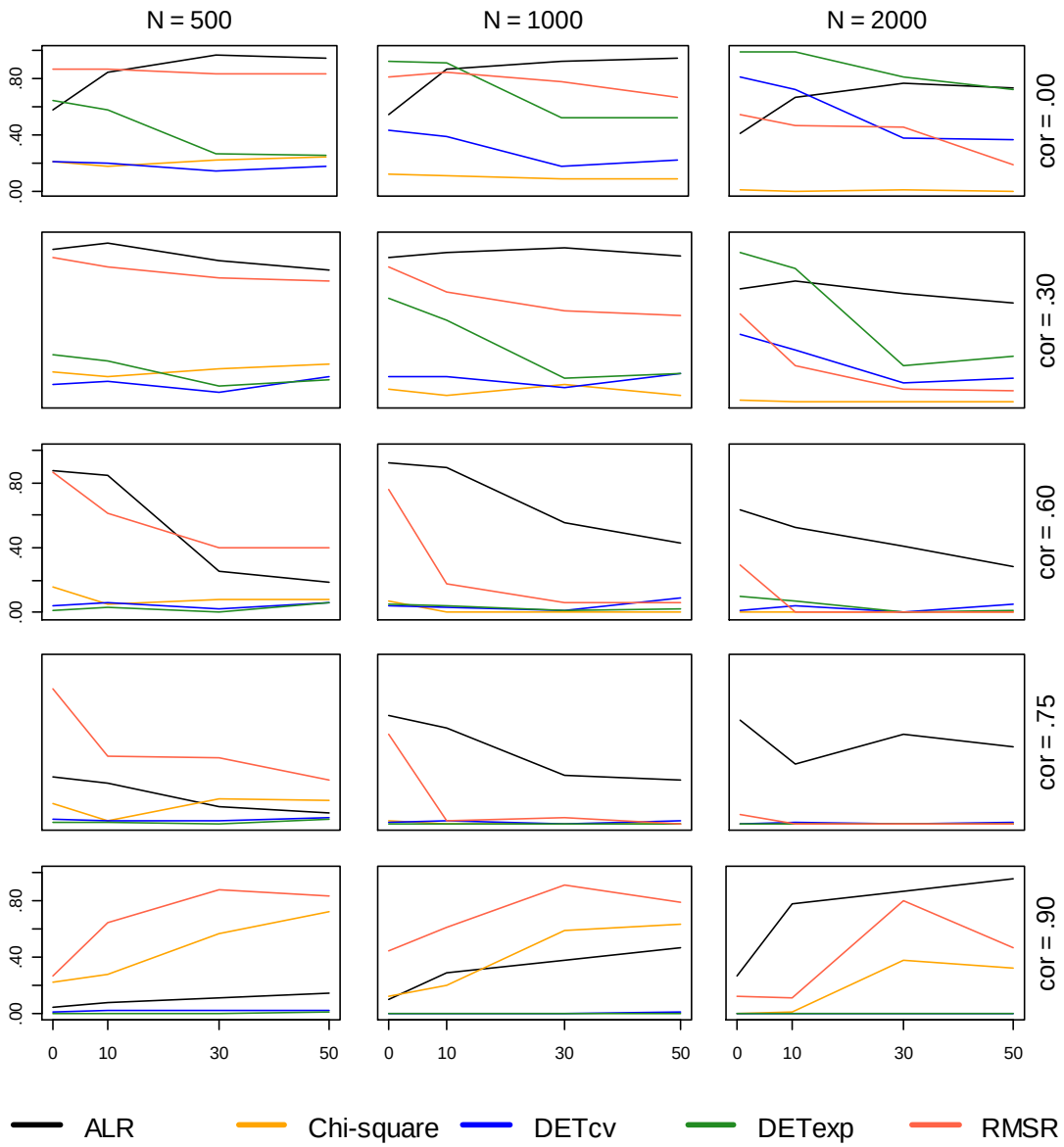

Figure 43. Proportion correct across complexity levels when the data follow a noncompensatory 2D MIRT model with 20 items per dimension.

Within a correlation level, RMSR performed better in conditions with *N* = 500 than in conditions with *N* = 1000 and *N* = 2000. Similarly, *ALR* tended to perform better in conditions with *N* = 500 when correlations were medium to low. However, the opposite was found when correlations reached .75 or higher; there

ALR's performance improved as the sample size increased. $\chi^2_{G/D}$ performed poorly across all complexity levels and the three sample size when correlations were .75 or lower. Only in the conditions with .90 correlation did $\chi^2_{G/D}$ show some improvement; the highest proportion correct observed for $\chi^2_{G/D}$ was in a condition with N = 500 and 50% complexity (72% correct).

The proportion of dimensional labeling. Figure 44 plots the marginal proportions of labeling sets of items as dimension-like for conditions where the data exhibit 30% complexity, following a true 2D noncompensatory structure with 20 items per dimension (note that figures for 0% and 10% look very similar to 30% complexity, thus only one figure is included in the text; figures associated with 0% and 10% can be found in Appendix B).

When the correlation was .00 or .30, RMSR, *ALR*, and the DETECT-based methods were generally successful in labeling two set of items as dimension-like. However, as the correlation increased, marginal proportions for labeling two sets of items as dimension-like tended to decrease for all methods across sample size. Further, it was noted that $\chi^2_{G/D}$was most successful in labeling any one set as dimension-like; particularly in conditions with N = 2000 (across all correlation levels) or across all sample size conditions when correlation was .60 or larger. Interestingly, at a correlation of .60, both DETECT-based methods tended to have higher marginal proportions for labeling two or none of the sets as dimension-like. At a correlation .75 or above, the NOHARM-based methods tended to have

higher marginal proportions for labeling any one set of items as dimension-like, a pattern noted particularly in cases with $N = 2000$ (note the inverse "V" shaped lines).

Conditions whose lines created the inverse "V" shape (i.e., $N = 1000$ and $N = 2000$ conditions with .90 correlation), suggested that high marginal proportions for labeling any one set of items as dimension-like for all methods were obtained. These types of patterns were largely observed across all conditions with 50% complexity (see Appendix B), suggesting that at 50% complexity, all methods tended to label only one set of items as dimension-like more often than either two or none of the sets as dimension-like.

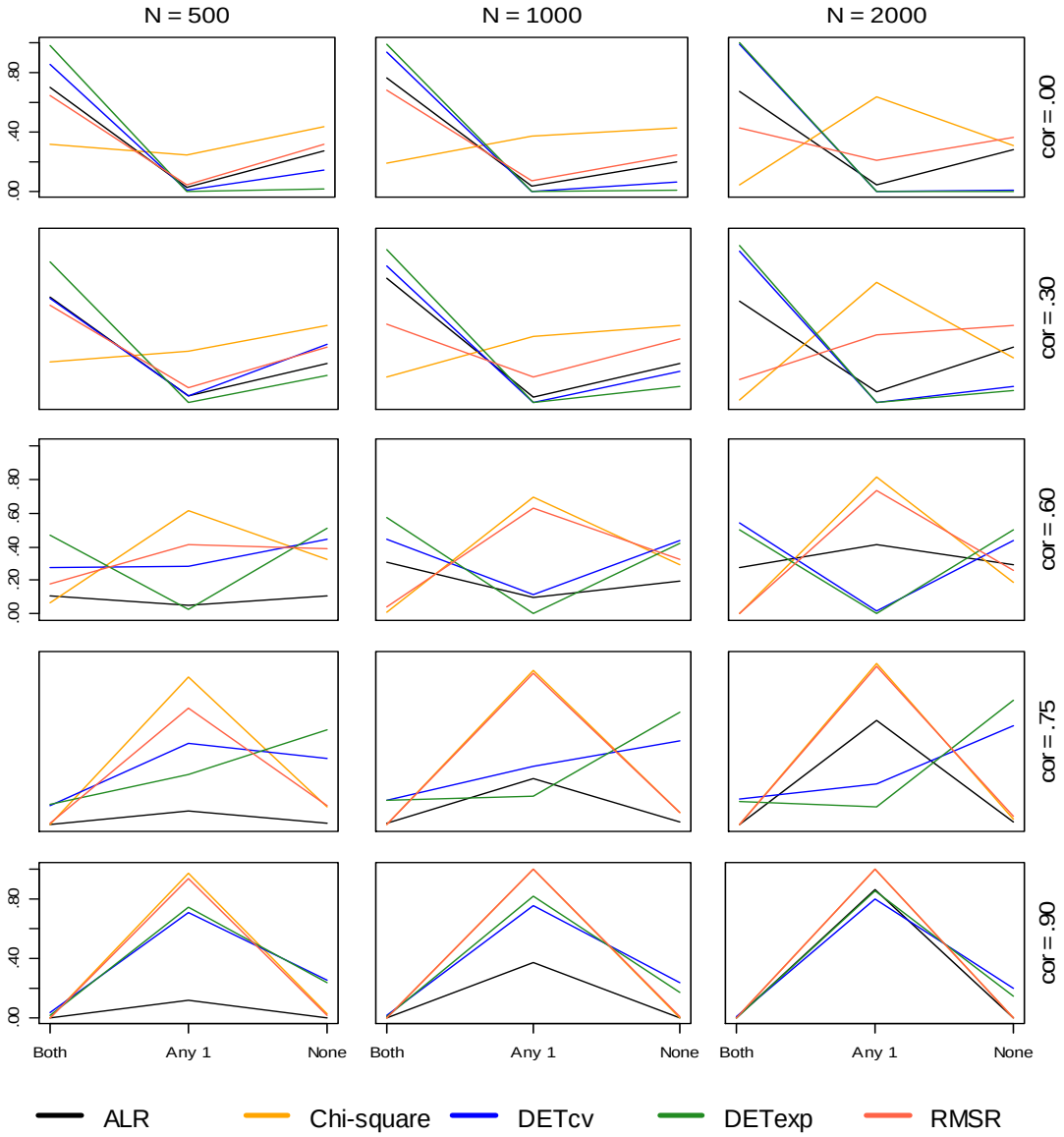

Figure 44. Marginal proportions across 500 replications that a method identified two (both), (any) one, or none of the sets of items as dimension-like (x-axis) when the data exhibit 30% percent complexity and follow a noncompensatory 2D MIRT model with 20 items per dimension.

The consistency of item classification. Figure 45 plots the classification consistencies for factorially simple items across complexity levels (x-axis) when the data follow a noncompensatory 2D MIRT model with 20 items per dimension.

Classification rates of factorially simple items in these conditions were highest for conditions with lower levels of correlations. The DETECT methods reported higher classification rates than their NOHARM counterparts, with larger differences found in conditions with smaller correlations and larger sample sizes. Classification consistency rates for all methods tended to drop as the complexity levels increased; particularly in conditions of .60 or less correlation for complexity levels of 30% and 50%.

As correlations increased to .90, none of the methods reported rates higher than .55 (DETECTe classification rate in condition with 0% complexity and N = 2000). Generally, at 50% complexity, none of the methods yielded high classification rates for any correlation level or sample size.

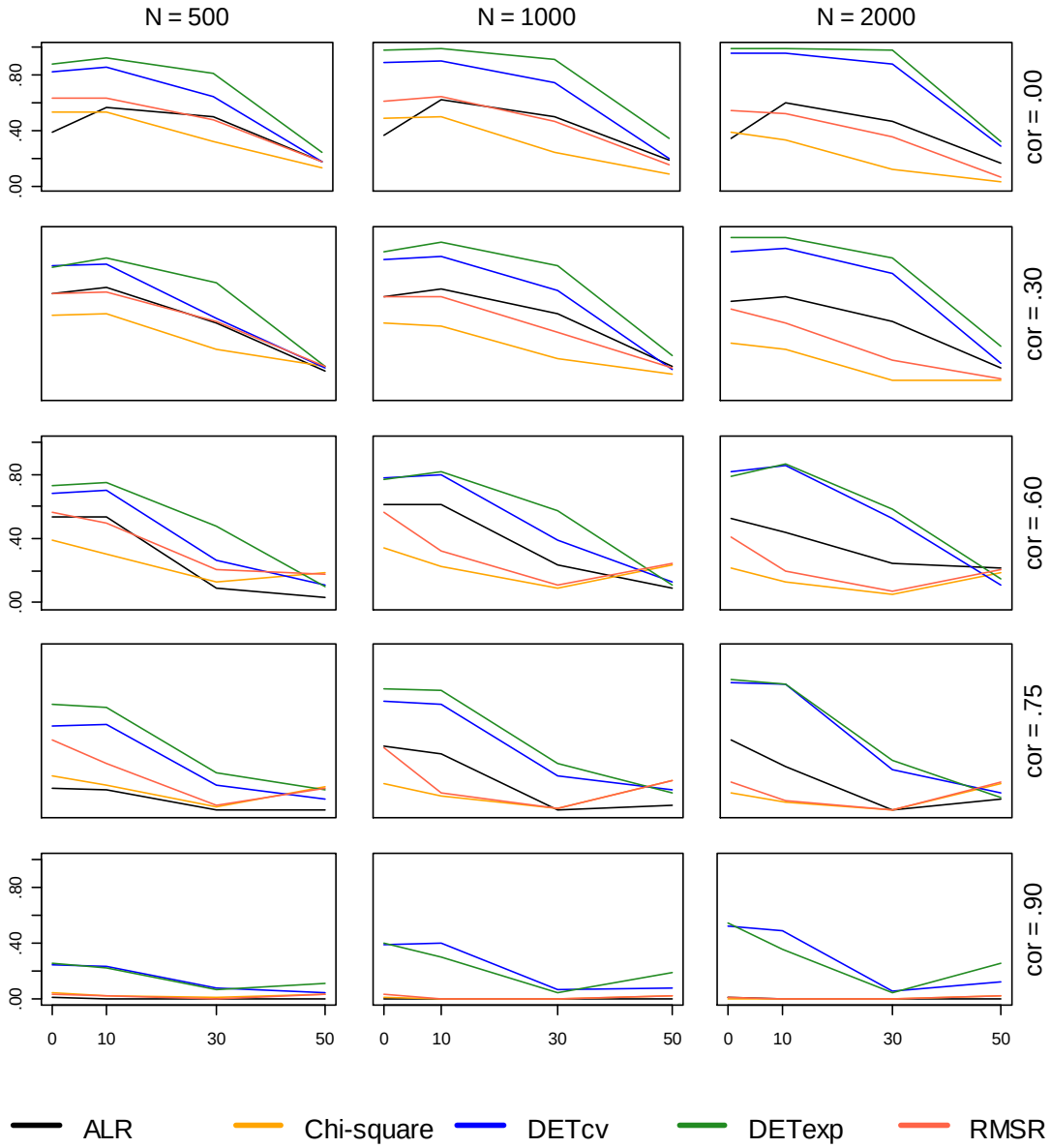

Figure 45. Consistency of factorially simple items across complexity levels when the data follow a noncompensatory 2D MIRT model with 20 items per dimension.

Figure 46 plots the classification consistencies for factorially complex items across complexity levels (x-axis) when the data follow a noncompensatory 2D MIRT model with 20 items per dimension. In conditions with 10% and 30% of complexity and correlations of .00 and .30, the NOHARM-based methods were

more consistent in classifying factorially complex items than their DETECT-based counterparts.

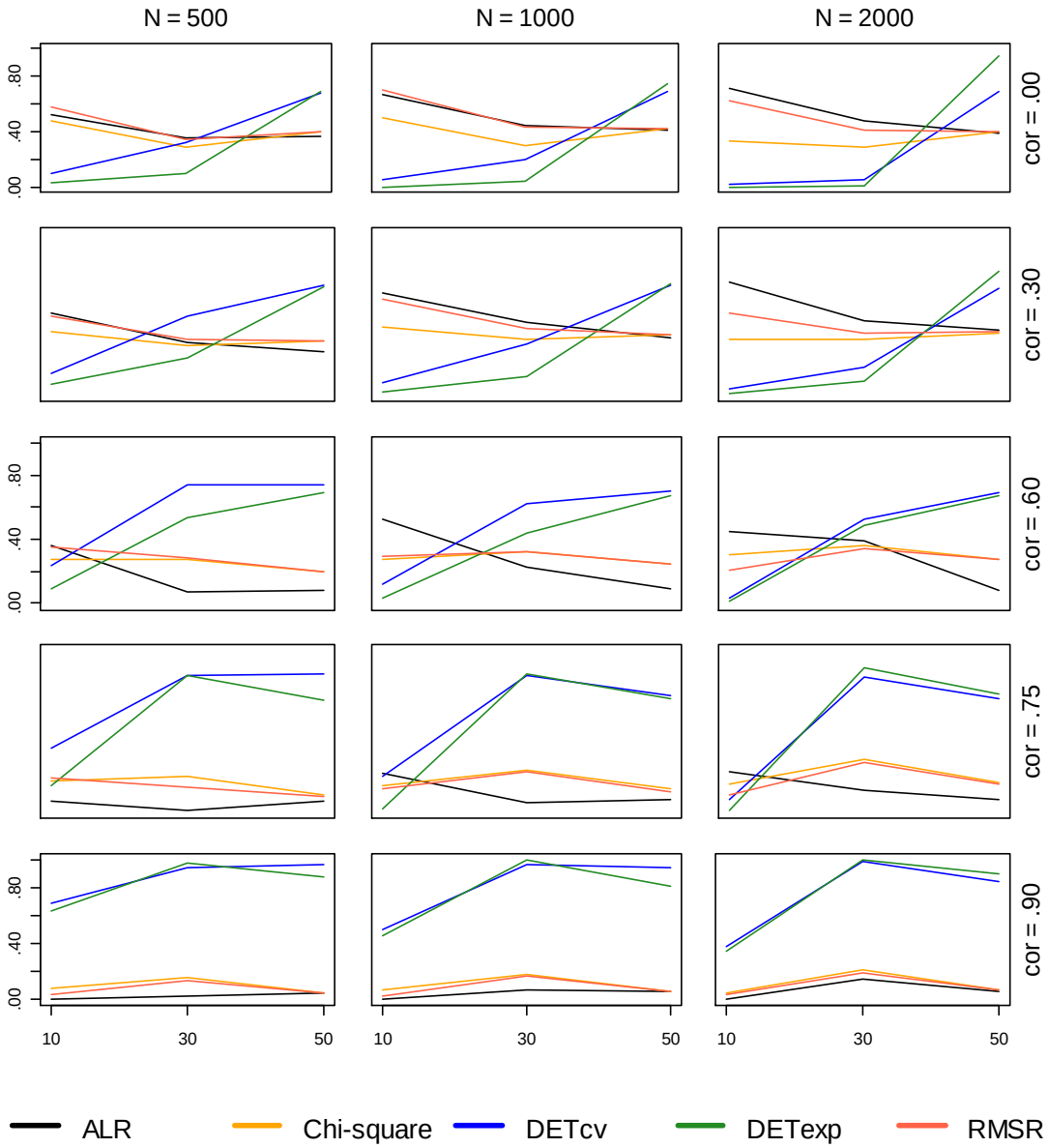

Figure 46. Consistency of factorially complex items across complexity levels when the data follow a noncompensatory 2D MIRT model with 20 items per dimension.

The opposite was found for complexity levels of 30% and 50% when the correlations increased to .75 or higher. At that point, a clear differentiation was

observed between the classification rates of the two groups of methods. DETECT methods yielded higher classification rates across all three sample sizes, while NOHARM-based methods reported rates of .25 or less.

Tests with twenty items per dimension with 3D structures.

The proportion of correct dimensional selection. Figure 47 plots proportion correct across complexity levels when the data follow a noncompensatory 3D MIRT model with 20 items per dimension. From the figure, it was observed that both NOHARM- and DETECT-based methods performed generally poorly across sample size and correlation levels at complexity levels of 10% or greater. One notable exception was the performance of RMSR, which yielded high proportion correct in a condition of correlation of .00, across all levels of complexity and sample sizes.

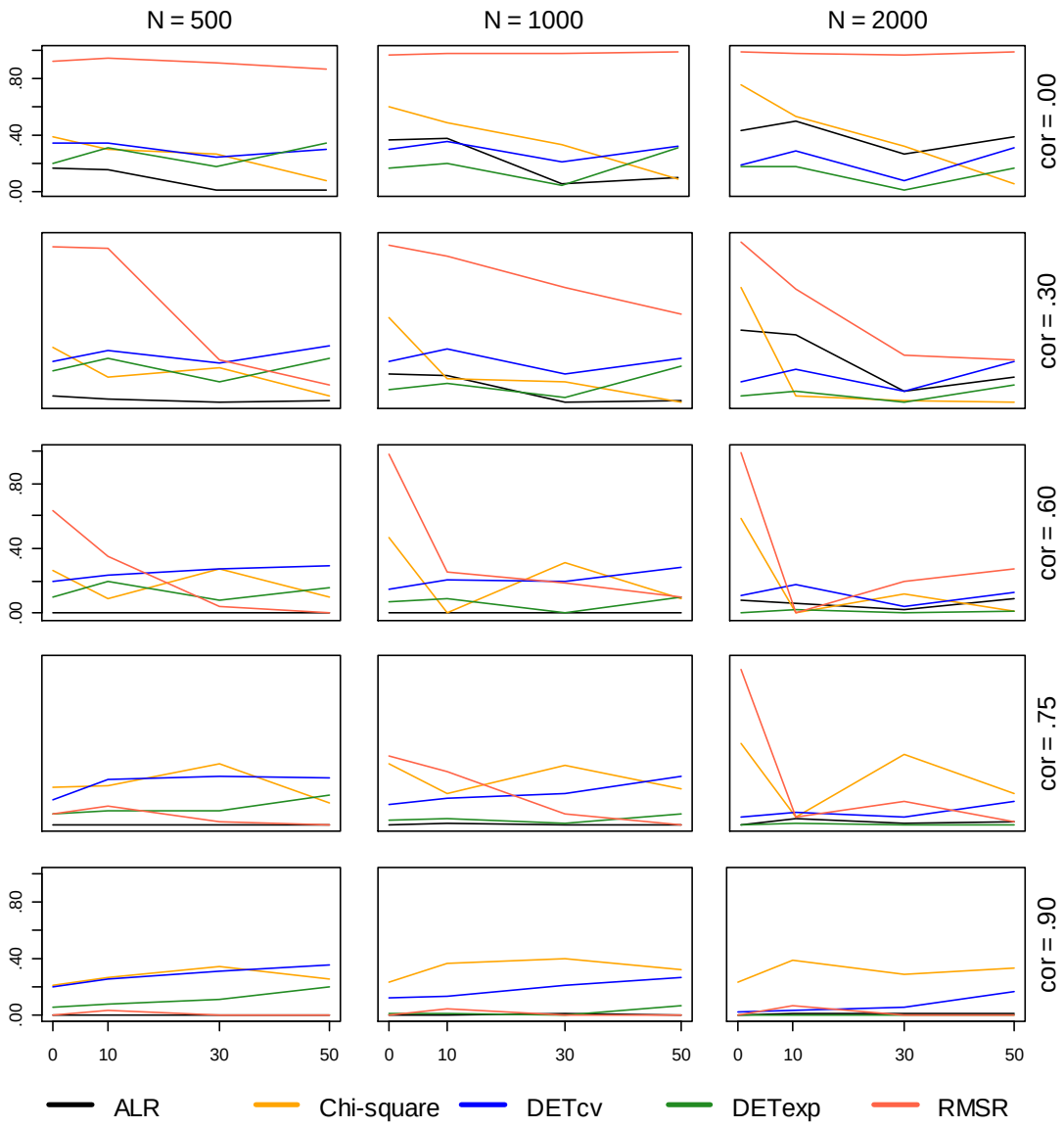

Figure 47. Proportion correct across complexity levels when the data follow a noncompensatory 3D MIRT model with 20 items per dimension.

Further, RMSR obtained high proportions correct in conditions with 0% complexity when the correlation equaled .30 across the sample sizes, or in N = 1000 and N = 2000 when the correlation was .60 and .75 respectively.

$\chi^2_{G/D}$ mostly performed poorly across all conditions when correlations were .75 or

lower. Its highest proportions correct were found in condition with N = 2000 and 0% complexity across correlation levels.

ALR and the two DETECT-based methods also tended to yield low proportions correct. Their respective proportions correct ranged between .38 and .46 across various sample size and correlation levels. Although neither of the DETECT-based methods performed well, it was observed that DETECTcv outperformed DETECTe.

The proportion of dimensional labeling. In conditions where data follow a 3D noncompensatory MIRT with 20 items per dimension, complexity levels had a somewhat small effect on how well the methods labeled sets of items as dimension-like. To illustrate the main findings in these conditions, Figure 48 plots the marginal proportions that each method identified sets of items as dimension-like for 50% complexity levels across the sample sizes and correlations (plots for 0%, 10%, and 30% of complexity looked very similar and with only a few minor deviations and are included in Appendix B). From Figure 48, it was observed that the methods generally reported low marginal proportions for labeling three set of items as dimension-like. This was noted across sample size and correlation levels.

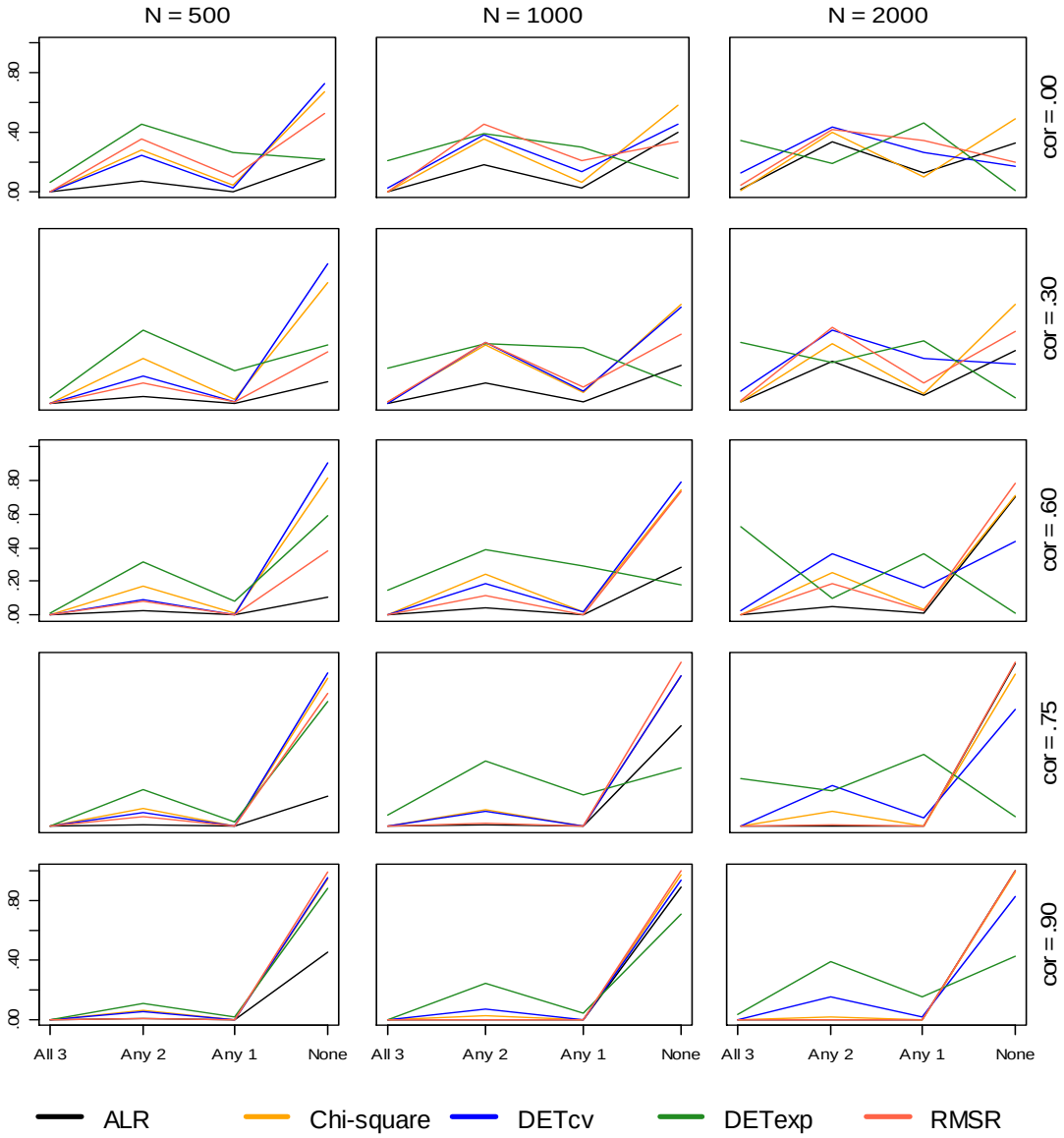

Figure 48. Marginal proportions across 500 replications that a method identified three, any two (both), (any) one, or none of the sets of items as dimension-like (x-axis) when the data exhibit 50% percent complexity and follow a noncompensatory 3D MIRT model with 20 items per dimension.

DETECTe yielded higher marginal proportions for labeling any two set of items as dimension-like in conditions with N = 500 and N = 1000; it also recorded

the lowest marginal proportions for labeling none of the set of items as dimension-like in N = 1000 and N = 2000 across all levels of correlation.

RMSR and $\chi^2_{G/D}$ yielded marginal proportions across conditions that were similar in magnitude to each other; the highest marginal proportions obtained from both methods were those that labeled none of the set of items as dimension-like. *ALR* was similar to other NOHARM methods, however, out of the three NOHARM-based methods, it tended to have the smallest marginal proportions for labeling of any one set of items as dimension-like.

The consistency of item classification. Figure 49 plots the classification consistencies for factorially simple items across complexity levels (x-axis) when the data follow a noncompensatory 3D model with 20 items per dimension. From the figure, it was observed that the DETECT-based methods reported higher classification rates than the NOHARM-based methods across all levels of correlation and sample size. These differences were noted particularly in conditions with smaller correlation levels across different sample sizes. NOHARM-based methods obtained low classification rates across conditions.

Additionally, within a correlation level (except for .90 correlation), as the sample size increased, methods reported higher classification rates. In conditions with .90 correlation, however, none of the methods yielded high classification rates, regardless of sample size.

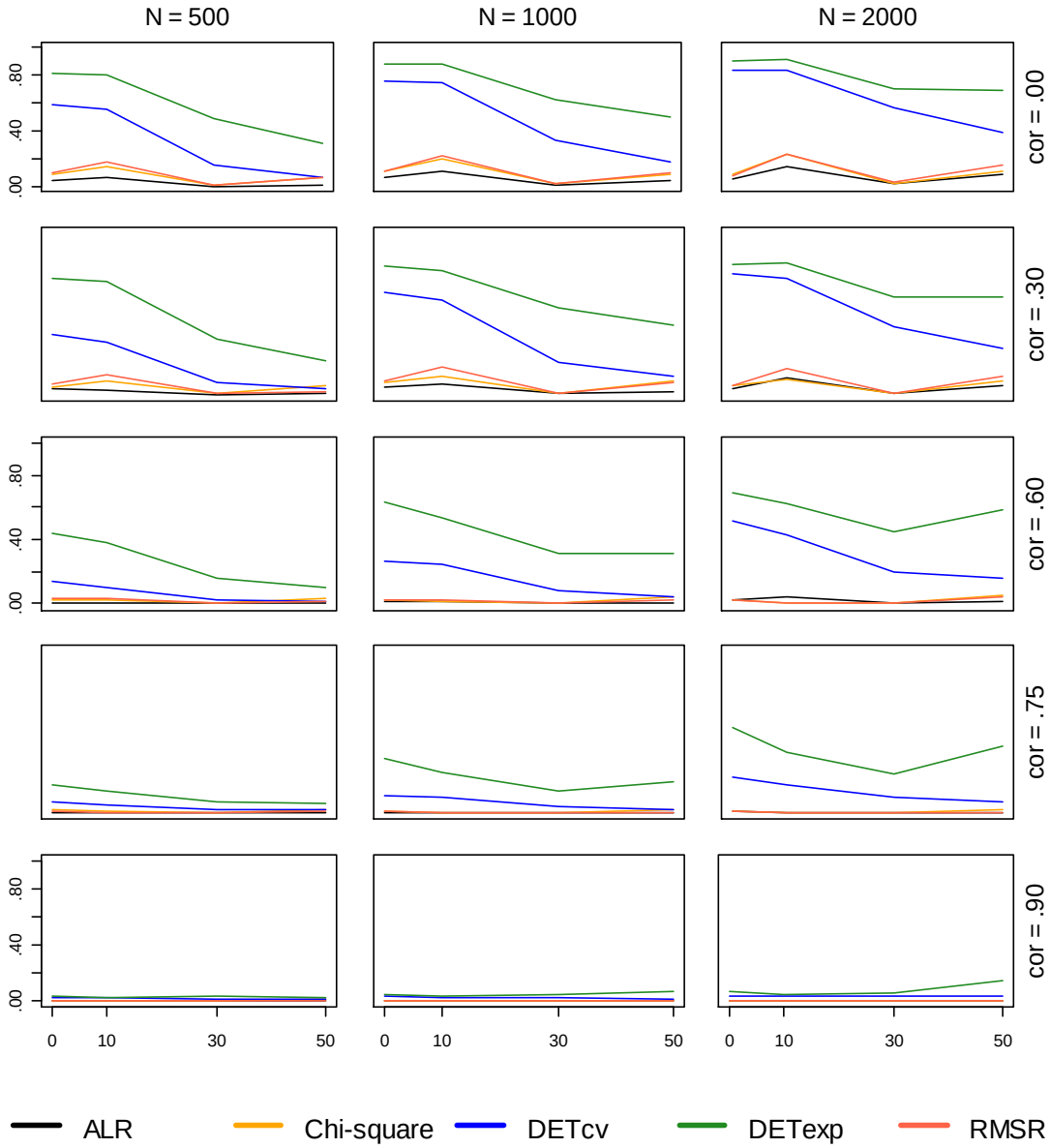

Figure 49. Consistency of factorially simple items across complexity levels when the data follow a noncompensatory 3D MIRT model with 20 items per dimension.

Figure 50 plots the classification consistencies for factorially complex items across complexity levels (x-axis) when the data follow a noncompensatory 3D model with 20 items per dimension. The DETECT-based methods were much

more consistent in classification of the factorially complex items than were NOHARM-based methods.

The classification rates for DETECTe and DETECTcv were high for conditions with 30% and 50% complexity across all sample size and correlation levels. At 10% complexity, the DETECT-based methods performed better at higher levels of the correlations. Complex item classification rates for *ALR*, $\chi^2_{G/D}$, and RMSR were very low and similar to each other (never rising above .25.)

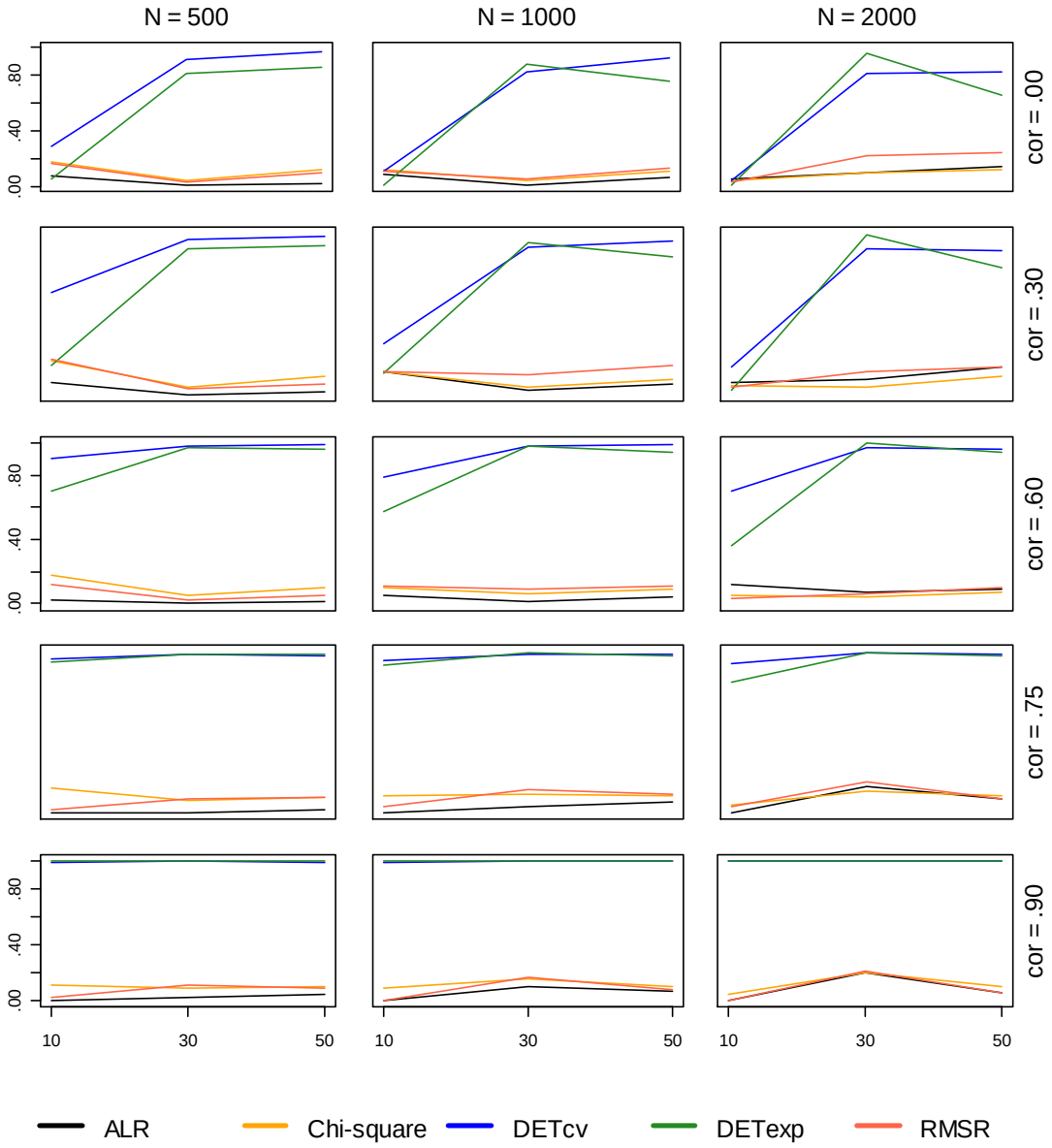

Figure 50. Consistency of factorially complex items across complexity levels when the data follow a noncompensatory 3D MIRT model with 20 items per dimension.

Effects due to the number of items on determining correct dimensionality. The preceding presentation has displayed results separately by the number of items associated with each dimension. Additional analyses were

conducted to examine the effects of the number of items associated with each dimension and sample size in conditions with noncompensatory data. Figures 51 through 56 correspond to analyses of the effects for varying the number of items for all sample size levels and dimensional structures. The figures plot the proportion of times within a condition (i.e., out of 500 replications) that each method accurately selected the correct dimensional structure in noncompensatory models. In the graphs, the y-axis ranges from 0 to 1 and represents the proportion of replications for which the method yielded the correct number of dimensions. Connected lines on the graphs (from 10 to 20 items per dimension) are drawn only for illustration purposes, not to imply any function between the two categories. Within a graph, different colors represent the five methods of interest.

Figure 51 plots the proportion correct when the data follow a noncompensatory, 2D MIRT model for 10 and 20 items per dimension for N = 500. RMSR showed improvement in proportion correct when the number of items increased in all conditions where data followed a 2D noncompensatory MIRT.

For $\chi^2_{G/D}$, however, an increase in the number of items resulted in worse performance in most conditions in terms of lower proportions of correctly identifying the true number of dimensions. The decrease in performance was noted across various levels of complexity and correlation, with most notable decreases occurring at lower levels of complexity and correlations (in two conditions, both at correlation of .90, $\chi^2_{G/D}$showed no improvement when

complexity was 30%, and showed improvement of .17 when complexity was 50%).

ALR tended not to be affected by the increase in the number of items when correlations were .00 or .30 across all levels of complexity, or at a correlation of .60 and 0% and 10% of complexity. However, *ALR*'s performance decreased as the number of items increased when the correlation was .60 and complexity was 30% and 50%, as well as at all complexity levels for correlations of .75 and .90. This suggested that across the complexity and correlation levels, increase in the number of items affected *ALR*'s performance negatively (i.e., smaller proportion correct) for only some conditions.

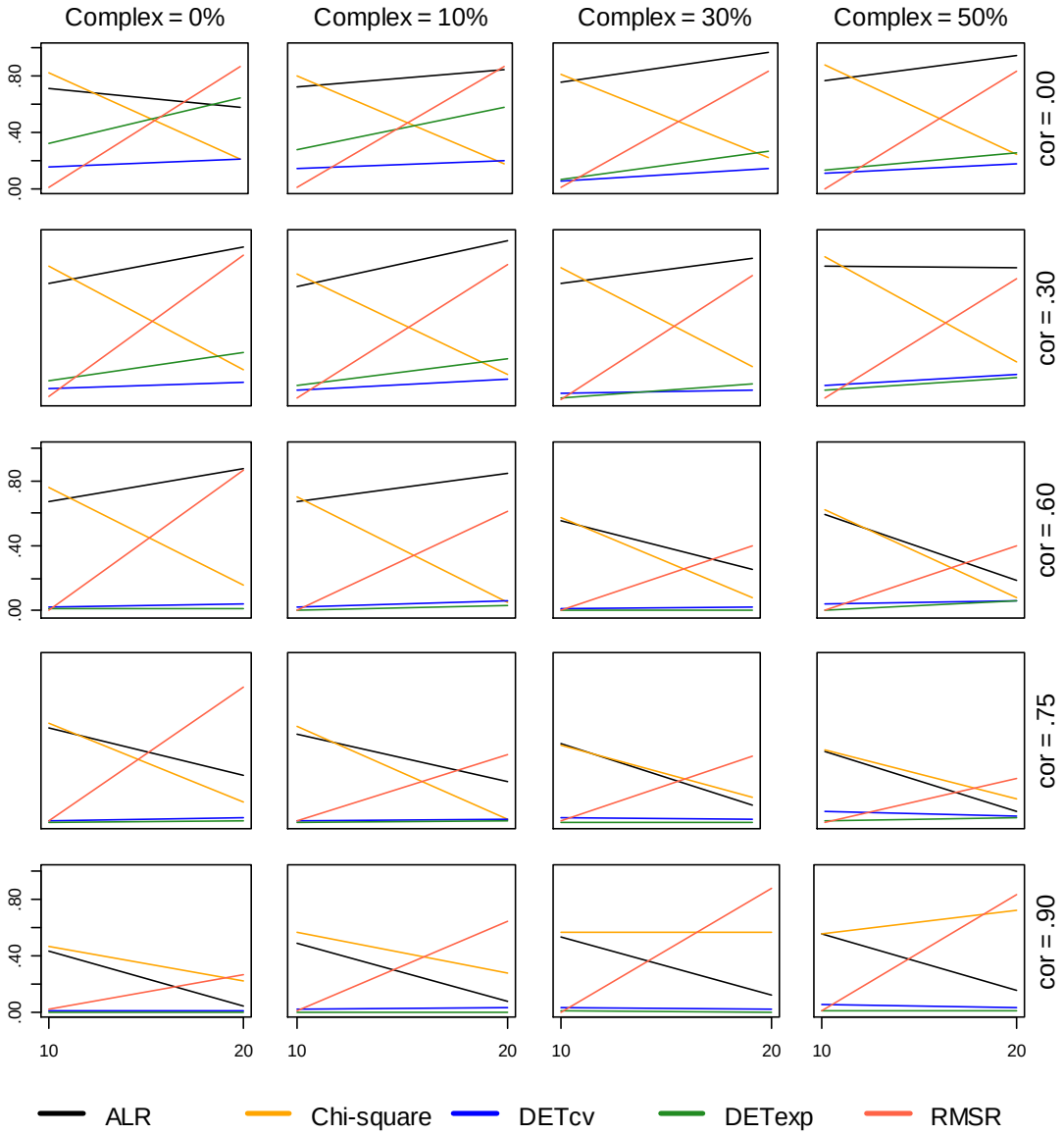

Figure 51. Proportion correct when the data follow a noncompensatory, 2D MIRT model for 10 and 20 items per dimension for N = 500.

The DETECT-based methods seemed to be less affected by the increase in the number of items for these conditions than NOHARM-based methods. A few slight increases in performance were noted for DETECTcv in lower complexity conditions with correlation of .00. However, it should be also noted that the

overall performance of the DETECT-based methods was very poor across complexity and correlation levels in the conditions with N = 500 where data followed noncompensatory 2D MIRT model.

Figure 52 plots the proportion correct when the data follow a noncompensatory, 2D MIRT model for 10 and 20 items per dimension for N = 1000. From Figure 52, it was observed that four out of five methods (all but $\chi^2_{G/D}$) yielded higher proportions correct when the number of items increased. Degrees of upward shifts however varied across the methods. The most notable upward shift in proportion correct going from 10 to 20 items per dimension was recorded by DETECTe in conditions with correlation of .00 and all levels of complexity, as well as 0% and 10% complexity with correlation of .30. It is also noteworthy that in those same conditions, DETECTe had somewhat large proportions correct.

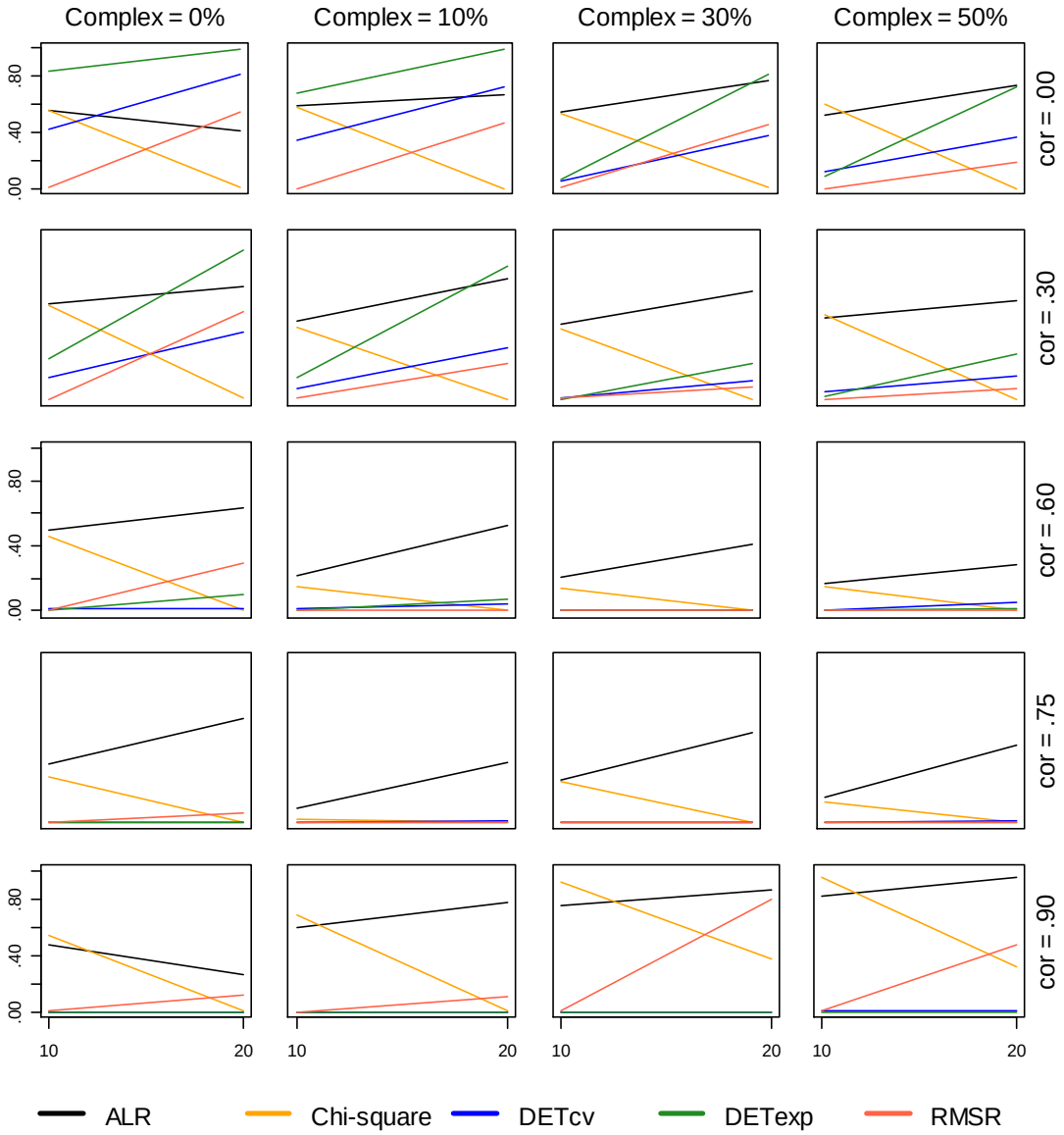

Figure 52. Proportion correct when the data follow a noncompensatory, 2D MIRT model for 10 and 20 items per dimension for N = 1000.

In the remaining conditions, an increase in the number of items did not affect DETECTe's performance, which was very poor regardless of the number of items. DETECTcv and RMSR followed the same pattern as DETECTe (same conditions yielded some improvement, although the magnitude of improvement

was smaller than in DETECTe). For *ALR*, only slight shifts upward or downward were noted as the number of items increased; most notable shifts occurred in higher complexity conditions with correlations of .75.

Figure 53 plots the proportion correct when the data follow a noncompensatory, 2D MIRT model for 10 and 20 items per dimension for N = 2000. It can be noted that the methods tended to maintain the same relationship between the increase of items and their performance when N = 2000 as they did when sample size was 1000.

Four out of five methods (all but $\chi^2_{G/D}$) tended to be positively affected by the increase in the number of items when correlations were at .30 or lower. At correlations of .60 or higher, generally the methods' performances stayed the same or decreased in moving from 10 to 20 items per dimension. Exceptions were found in *ALR*, which tended to benefit from the increase in the number of items at high correlations across complexity levels, and RMSR, which showed some improvement for complexity levels of 30% and 50% when correlation was .90. The DETECT-based methods once again showed an upward shift in moving from 10 to 20 items per dimension only in conditions with lower correlation levels. Although, as noted earlier, at .60 or higher correlation, the DETECT methods performed suboptimal across any complexity level regardless of the number of items.

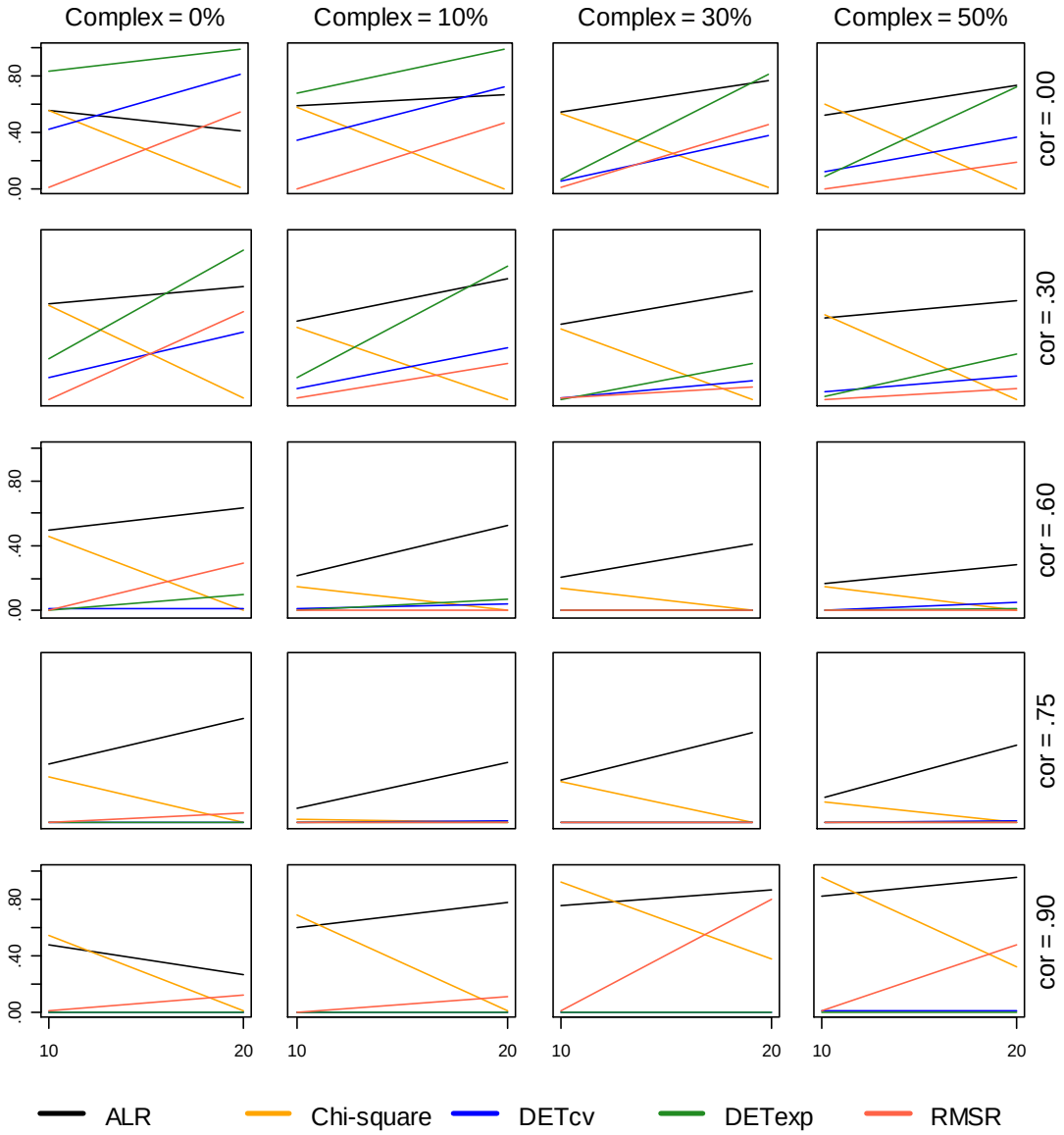

Figure 53. Proportion correct when the data follow a noncompensatory, 2DMIRT model for 10 and 20 items per dimension for $N = 2000$.

The impact of an increase in the number of items per dimension for each sample size was also investigated for conditions where data follow a 3D noncompensatory MIRT. Figure 54 plots the proportion correct when the data follow a noncompensatory, 3D MIRT model for 10 and 20 items per dimension

for N = 500. As suggested by Figure 54, RMSR was the only method that largely improved as the number of items increased; and that was not the case for all conditions.

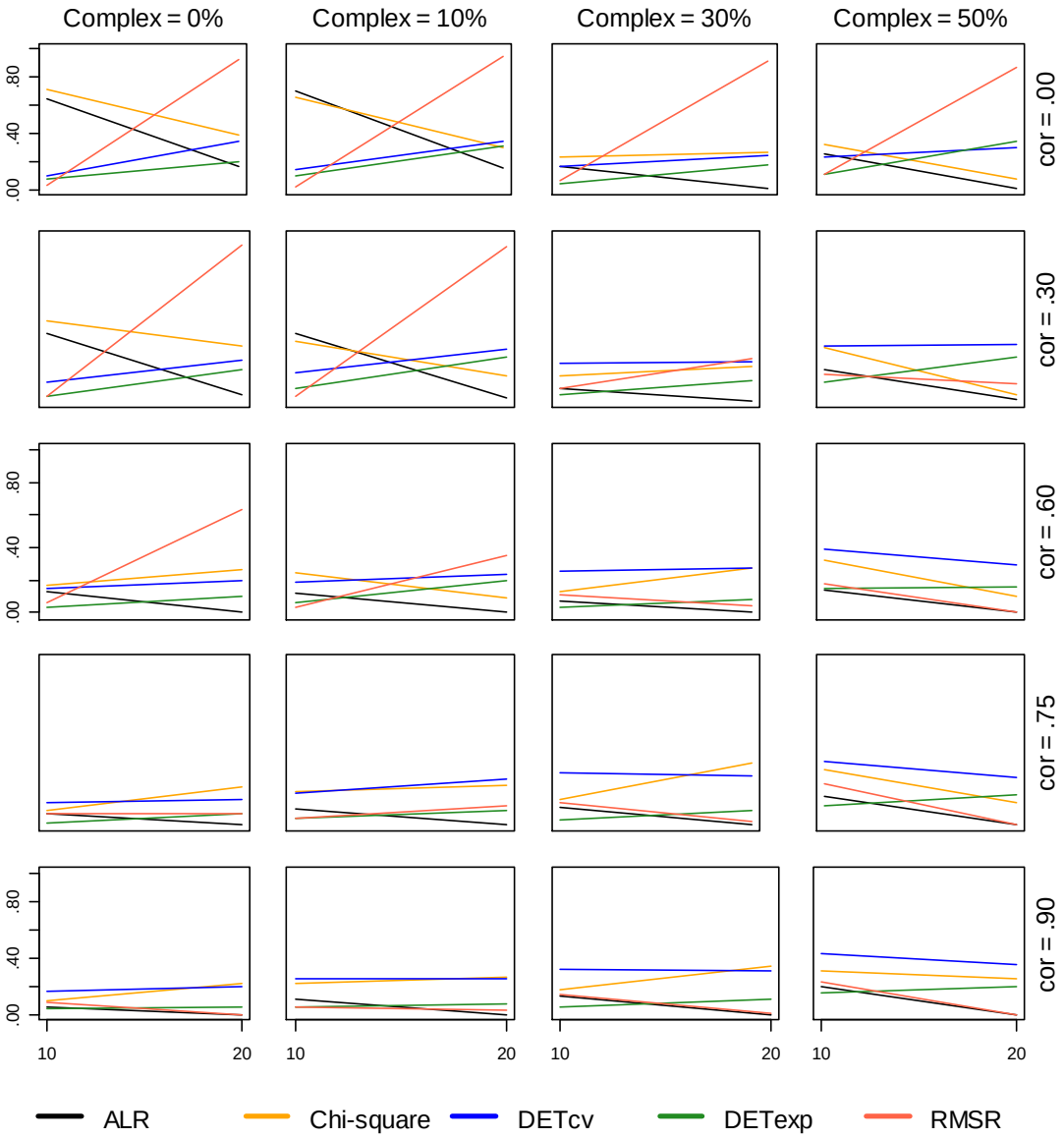

Figure 54. Proportion correct when the data follow a noncompensatory, 3D MIRT model for 10 and 20 items per dimension for N = 500.

The improvement in RMSR was only noted in conditions with correlation of .00 (across the levels of complexity) and in conditions with .30 or .60 correlation and 0% or 10% complexity. $\chi^2_{G/D}$ and *ALR* generally performed worse as the number of items increased for conditions of low correlation. The DETECT-based methods showed a slight upward shift in proportion correct in some conditions, however, as noted in the previous discussion of noncompensatory MIRT models, the DETECT-based methods yielded low proportions correct across conditions.

Figure 55 plots the proportion correct when the data follow a noncompensatory, 3D MIRT model for 10 and 20 items per dimension for N = 1000. It was observed that RMSR performed better when the number of items increased for conditions with correlation of .30 or less, across all levels of complexity. While the DETECT-based methods also yielded an upward shift from 10 to 20 items per dimension for the same set of conditions (correlations of .30 or less and all complexity levels), the increase in proportion correct was much less pronounced compared to the RMSR.

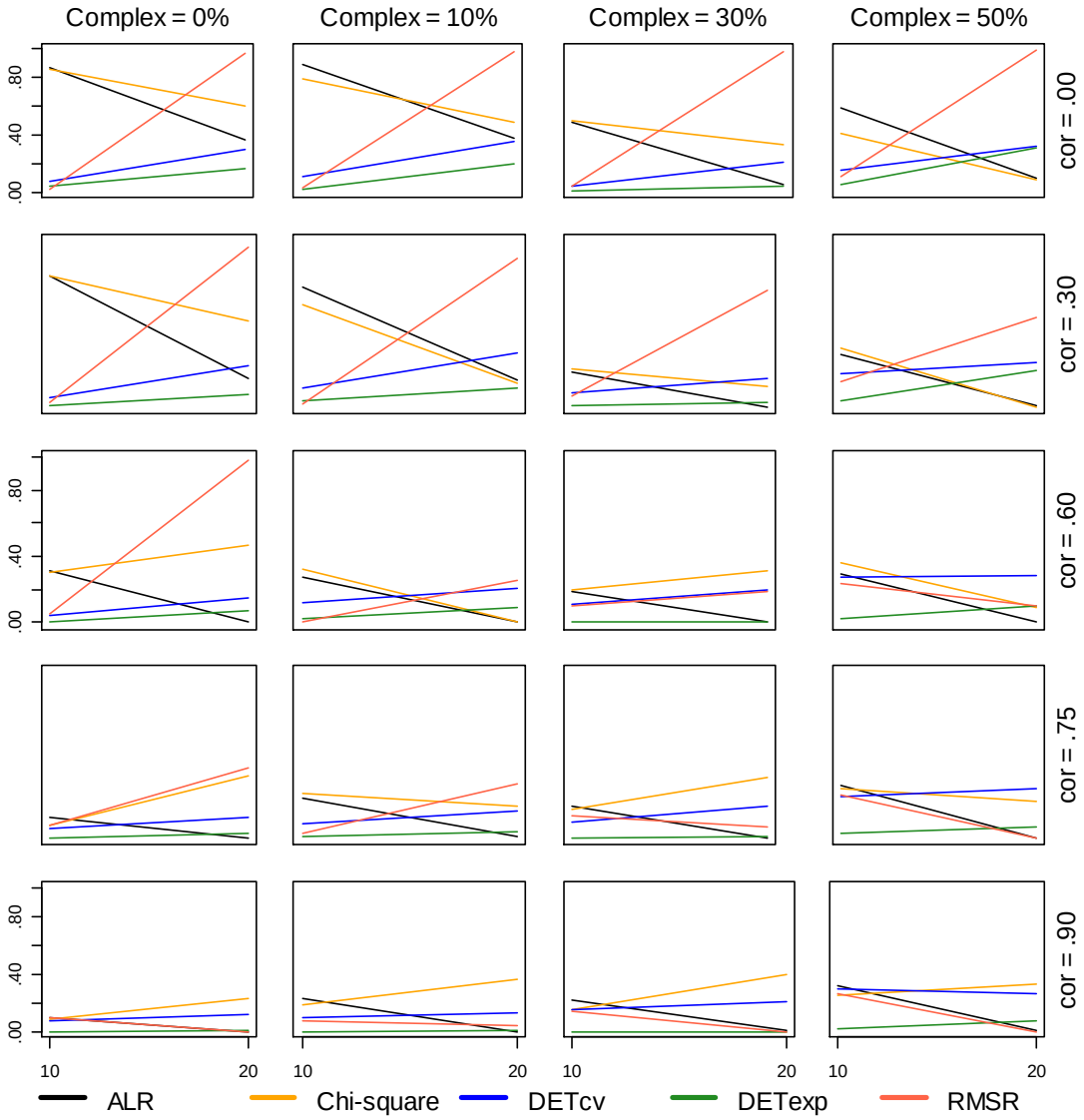

Figure 55. Proportion correct when the data follow a noncompensatory, 3D MIRT model for 10 and 20 items per dimension for N = 1000.

$\chi^2_{G/D}$ and *ALR* tended to decrease in performance as the number of items increased, particularly when correlation levels were .30 or less. Overall, the methods seemed to perform similarly for conditions when correlations were .75 or larger across the complexity levels. In those cases, the proportion correct for either 10 or 20 items per dimension was not very high.

The proportions of correct identification of dimensionality when the data follow a noncompensatory 3D MIRT model for 10 and 20 items per dimension for $N = 2000$ are plotted in Figure 56. General conclusions made about the impact of increase of the number of items echoed those previously discussed $N = 500$ and $N = 1000$. Most often, the increase in the number of items helped the RMSR method to obtain higher proportions correct in conditions with small correlations across complexity levels, and in conditions with 0% of complexity and correlations of .75 or smaller.

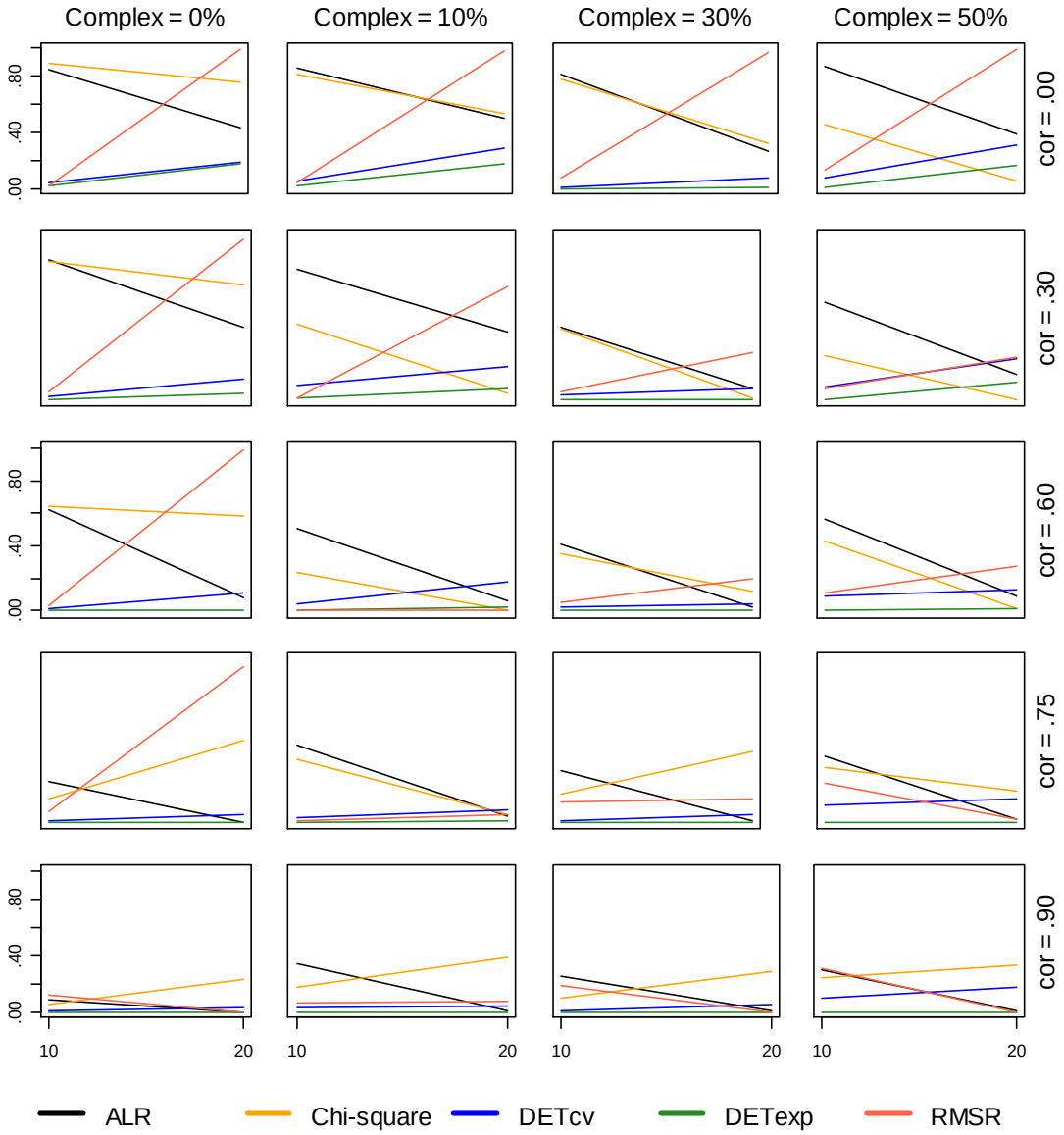

Figure 56. Proportion correct when the data follow a noncompensatory, 3D MIRT model for 10 and 20 items per dimension for N = 2000.

The other four methods performed equally well or better with 10 items per dimension than with 20 items per dimension, although there were a few exceptions. In the conditions with 3D noncompensatory MIRT, regardless of the

sample size or the number of items, it was observed that as complexity and correlation levels increased, performances for the methods generally worsened.

Effects due to the number of items on methods' ability to label sets of items as dimension-like. A comparison of results for 2D conditions where data follow noncompensatory MIRT model suggests that the number of items per dimension meaningfully affected RMSR proportions of labeling two sets of items as dimension-like when complexity was at 0% (Figure 33 and Figure B12 in Appendix B). In those conditions, RMSR increased in proportion of labeling two sets of items as dimension-like across the levels of correlation and sample size. The other methods remained somewhat unaffected by the increase in items in conditions with .60 or smaller correlation across sample sizes. Most notably, *ALR* decreased in proportion of labeling two or any one sets of items as dimension-like when the number of items increased, but the DETECT-based methods tended to improve in labeling sets of items as dimension-like as the number of items increased.

As complexity increased to 10% (comparing Figure 34 and Figure B13 in Appendix B), an increase in number of items negatively affected $\chi^2_{G/D}$ to label two sets of items as dimension like in conditions with correlation of .60 or lower (i.e., smaller proportions of labeling two sets of dimensions were observed). However, at correlations of > .60, $\chi^2_{G/D}$ seemed to be positively affected by the increase in number of items, yielding larger proportions of labeling two or any one sets of items as dimension-like. Other methods tended to be only slightly

affected by the increase in number of items; when affected, methods did not yield always positive or negative shifts in proportions of labeling of sets of items. Often it was dependent on correlation level.

Generally, in conditions with 2D and complexity of 30% when data follow a noncompensatory MIRT model, an increase in number of items positively affected methods in labeling sets of items as dimension-like when correlations were ≤ .60. However, opposite effect was found when correlations were > .60 (comparing Figures 35 and 44). Comparison of results for 50% complexity when data follow a 2D noncompensatory MIRT model suggested that the number of items per dimension did not meaningfully affect the proportion of labeling sets of items as dimension-like for any of the methods (comparison of Figure 36 and Figure B14 in Appendix B).

In conditions where data follow a 3D noncompensatory MIRT model, an increase in number of items per dimension meaningfully affected only the DETECT-based methods across sample sizes and correlation levels of .60 or lower. NOHARM-based methods did not seem to be meaningfully affected by the increase in number of items per dimension. These behaviors were noted across all complexity levels, however, as complexity increased, the positive effect (i.e., higher proportions of labeling sets of items as dimension-like) diminished. These comparisons were made based on Figures B9 and B15 in Appendix B for 0%, Figures B10 and B16 in Appendix B for 10%, Figures B11 and B17 in Appendix B for 30%, and Figures 40 and 48 for 50% complexity.

Effects due to the number of items on methods' ability to consistently classify items. Comparison of results for conditions where data follow a 2D noncompensatory MIRT model suggests that the number of items per dimension somewhat affected the DETECT-based methods. Consistency rates for factorially simple items of in the DETECT-based methods were higher in conditions with 20 items per dimension. These comparisons were based on a visual comparison of Figures 37 and 45. Effects of the increase in number of items on consistency rates for factorially complex items in 2D conditions were again somewhat meaningful for DETECT-based methods; however, the effects for factorially complex items were in downward direction. In other words, the increase in number of items per dimension in conditions with 2D noncompensatory MIRT model yielded lower classification rates of factorially complex items in the DETECT-based methods. As with factorially simple items, the NOHARM-based methods tended to be less affected by the increase in number of items in classification of factorially complex items (these comparisons were based on visual inspection of Figures 38 and 46).

A comparison of results for conditions where data follow a 3D noncompensatory MIRT model suggest that an increase in number of items only affected the DETECT-based methods in their ability to classify factorially simple items. An increase in the number of items per dimension led to higher classification rates of factorially simple items from the DETECT-based methods across sample sizes and correlation levels of .75 or lower. The NOHARM-based methods' classification rates were not meaningfully affected by the increase of

items per dimension (these conclusions were based on comparisons of Figures 41 and 49). Comparison of classification results for factorially complex suggested that an increase in the number of items had no meaningful effect on classification rates of any of the methods. Only a slight decrease in consistency rates of factorially complex items was noted in conditions with 10% complexity and low correlations for DETECT-based methods (these comparisons were based on visual inspection of Figures 42 and 50).

Chapter 5

DISCUSSION

The primary purpose of this study was to investigate the performance of current, popular methods in determining test dimensionality when the data exhibit complex structure. Specifically, this study examined the performance of methods rooted in conditional covariance theory implemented in DETECT (exploratory and cross-validated), and methods based on the output from NOHARM ($\chi^2_{G/D}$, *ALR*, and RMSR), a nonlinear factor analytic procedure. The data were generated such that varying degrees of complexity were introduced.

General Discussion of Methods' Performances

This research sought to answer the question of how well the methods perform in assessing dimensionality of the tests when the data exhibit complexity. The performance of five methods under consideration was evaluated using three main outcomes. A number of design factors were manipulated, including data-generating model, sample size, true number of dimensions, correlation(s) between the dimensions, number of items per dimension, and the amount of complex items. The effects of these, broadly speaking, were as follows.

A main effect for data-generating model was observed in this study. In compensatory conditions, the DETECT-based methods tended to outperform the NOHARM-based methods in correctly identifying the true dimensionality. In compensatory cases, the DETECT-based methods also tended to be more

consistent (than the NOHARM-based methods) in classifying the factorially simple items for lower levels of complexity and in classifying the factorially complex items for the highest level of complexity.

In noncompensatory conditions, however, the NOHARM-based methods of $\chi^2_{G/D}$and *ALR* were more successful in correct identification of dimensional structure than DETECT-based methods. Classification of factorially simple and factorially complex items suffered greatly for the NOHARM-based methods in noncompensatory conditions.

As complexity levels increased, the NOHARM-based methods decreased in their accuracy to select correct dimensionality structure more so than the DETECT-based methods. An increase in complexity also affected the methods' ability to label sets as dimension-like and item classifications. Methods tended to label more sets of items as dimension-like when complexity levels were 30% or lower, particularly in compensatory conditions.

Sample size had somewhat divergent effect for the two types of methods. For the DETECT-based methods, generally, an increase in sample size either improved the performance of the methods, or did not affect it much. For the NOHARM-based method $\chi^2_{G/D}$, increases in sample size tended to hinder its performance more often than to improve it. For NOHARM-based *ALR*, increase in sample size contributed to better performance in conditions with higher dimensionality and 10 items per dimension (i.e., when the number of items per dimension was smaller and true dimensionality larger, increase in sample size

positively affected *ALR's* performance). However, increase in sample size in other conditions tended to result in poorer performance of *ALR* (similar to what was observed with $\chi^2_{G/D}$).

The magnitude of the correlation(s) between dimensions particularly affected the performance of the methods in noncompensatory conditions, where increases in correlation generally yielded lower proportions correct and classification rates of the methods. It noteworthy that, as was the case with other design factors, the effects of manipulating the correlation effect on the method were not equal or consistent.

Lastly, $\chi^2_{G/D}$, *ALR*, and DETECTe methods tended to perform about the same under 3D as they did under 2D structures, while RMSR and DETECTcv performed better as the true dimensionality increased (particularly when complexity and correlations increased).

While the above summaries concern broad summaries of main effects, the following subsections provide syntheses and recommendations for the compensatory and noncompensatory contexts.

Data Following Compensatory Structures

The DETECT-based methods typically outperformed the NOHARM-based methods in terms of identifying the correct number of dimensions, especially when the correlations were .60 or smaller, and the sample size was larger. These findings are consistent with previous research on DETECT (e.g.,

Girl et al., 2006, Zhang & Stout, 1999b). Particularly good performance of the DETECT-based methods was noted in conditions with complexity levels of 30% or less. As the complexity levels increased and the sample size decreased, the performance typically diminished. Between the two DETECT methods, DETECTe often outperformed DETECTcv, mostly when N = 500 and N = 1000 and in conditions with longer tests (i.e., 20 items per dimension).

The latter result was, however, not surprising. When conducting exploratory DETECT using a cross-validated mode (i.e., DETECTcv), a researcher decides how much of the whole sample is to be used as the training sample. In the current study, 50% of the sample was dedicated to the training sample. The amount of information for any one analysis of DETECT in the cross-validated mode was less than in the exploratory mode. Therefore, it comes as no surprise that the largest differences in performance between DETECTe and DETECTcv were found in conditions with smaller sample sizes and longer tests (i.e., conditions with 20 items per dimension). Nonetheless, DETECT methods tended to perform better than their NOHARM-based counterparts in correctly identifying a true dimensionality in conditions with the compensatory MIRT across all complexity levels for various sample sizes and correlation levels.

Of the three NOHARM-based methods, $\chi^2_{G/D}$ and *ALR* generally outperformed RMSR. $\chi^2_{G/D}$ was generally found to be most accurate in conditions with shorter tests, particularly when the sample size and/or complexity were

small. The performance of $\chi^2_{G/D}$ diminished as the correlation levels increased to .75 and the sample size was large, a finding consistent with the research on $\chi^2_{G/D}$ when 0% complexity conditions were considered (e.g., De Champlain & Gessaroli, 1998).

The performance of *ALR* improved in conditions with the correlation of .60 and 30% or less complexity as the number of items increased; an opposite effect was found for $\chi^2_{G/D}$, particularly when N = 2000. This finding was consistent with the previous research on *ALR*, which suggested that an increase in the number of items improved the accuracy of *ALR* (Finch & Habing, 2005).[7] However, this finding was somewhat inconsistent with the research on $\chi^2_{G/D}$ (De Champlain & Gessaroli, 1998), which suggested that the effects of the number of items as well as the correlation level had little or no effect on $\chi^2_{G/D}$. As in the current study, the performance was negatively affected by the increase in the test length. It should be noted, however, that De Champlain and Gessaroli (1998) acknowledged limitations of their findings, particularly with respect to considering more complex multidimensional models, such as those investigated in the current study.

[7] It is noteworthy that in the same study, Finch and Habing (2005) found DETECT to perform worse with an increase in items. In the current study, it was observed that, typically, an increase in items did not affect DETECTe, but it did slightly affect DETECTcv.

In conditions where the data followed a compensatory MIRT model, all methods seemed to successfully label two sets of items as dimension-like for true 2D structures when the data exhibited 30% or less complexity and the correlation was .75 or smaller. When the complexity increased to 50%, the DETECT-based methods tended to have success labeling only one dimension-like set in conditions with a small correlation and typically had high marginal proportions for not being able to label any sets of items as dimension-like. The NOHARM-based methods tended to label either two or none sets of items as dimension-like when correlation was .75.

In true 3D conditions, the methods tended to label two sets as dimension-like well up to 30% complexity as well; however, the effect of the correlation level was more notable in 3D than in 2D compensatory conditions. As the correlations rose above .30 in 3D conditions, larger sample sizes were needed to successfully label three dimension-like sets.

All methods yielded high consistency rates of factorially simple items when the complexity levels were 30% or less and the correlation levels were .75 or lower. $\chi^2_{G/D}$, *ALR*, and DETECTe tended to have higher rates than the DETECTcv and RMSR; however, in the low correlations and when $N = 2000$, those differences were only slight. An increase in true dimensionality (from 2D to 3D) resulted in minor differences in classification rates for individual methods; the DETECT-based methods were most notable in improving classification rates

for factorially simple items when true dimensionality increased. For compensatory conditions in either 2D or 3D cases, an increase in the number of associated items per dimension did not seem to affect any of the methods' classification rates.

For factorially complex items from 2D compensatory conditions, the NOHARM-based methods yielded classification consistency rates around .50 across all levels of complexity, sample sizes, and correlation. The DETECT-based methods were less consistent in situations with complexity levels below 50%, but more consistent with complexity levels of at 50%. An increase to 3D structures did not affect the NOHARM-based methods and their classification rates. The DETECT-based methods, however, yielded higher classification rates of factorially complex items at 30% complexity (compared to 50% in 2D). These DETECT results are somewhat similar to those found in Gierl et al. (2006) study. However, an exact comparison cannot be made due to different strategies for computing classification rates.

Synthesizing the preceding discussion, the following recommendations can be drawn in compensatory MIRT situations. The DETECT-based methods, particularly DETECTe, performed the best in terms of identifying the number of dimensions. This was true even for high levels of complexity, a somewhat surprising result given that DETECT assumes simple structure. However, as the complexity increases, it becomes more difficult to label the resulting sets of items

from DETECT in terms of the dimensions. Moreover, DETECT is fairly inconsistent in its classification of complex items. These difficulties occur because DETECT assigns all the items to non-overlapping clusters, and so in situations where the correct number of clusters is supported, the complex items wind up being inconsistently assigned to the clusters, complicating the interpretations of the clusters.

Thus, DETECTe can be recommended for determining the number of dimensions, when the MIRT models are compensatory in nature. There appears to be little difference between the exploratory and cross-validated DETECT methods. Where differences exist, the exploratory approach generally performed better. However, researchers should have caution when interpreting the clusters when simple structure does not hold. DETECT provides indices meant to indicate when approximate simple structure does not hold (e.g., *r* ratio or IDN index; Roussos & Ozbek, 2006, Zhang & Stout, 1999b). More research on DETECT's utility for identifying the presence of complex structure—and therefore alerting the researcher to have caution in interpreting the resulting clusters—is needed.

Data Following Noncompensatory Structures

The NOHARM-based methods $\chi^2_{G/D}$ and *ALR* most often correctly identified the true dimensional structure in 2D conditions with 10 items per dimension across all complexity levels. In 2D conditions where the number of items increased to 20 per dimension, *ALR* remained to be one of the most accurate

methods but $\chi^2_{G/D}$ performance diminished. Increase in the number of items helped RMSR and DETECTe methods to improve in accuracy, although DETECTe method was only accurate in *N* = 1000 and *N* = 2000 conditions with 0% or 10% complexity and small correlations (.00 and .30).

An increase in true underlying dimensionality (from 2D to 3D) resulted in *ALR* and $\chi^2_{G/D}$ performing best, in particular with *N* = 2000 and lower correlations. An increase in the number of items in 3D conditions led to decreased accuracy in all methods except RMSR across all complexity levels. RMSR performed well in conditions with correlation of .00, as well as conditions with 0% of complexity and .60 correlation (especially when *N* = 2000).

Thus, recommendations for determining the number of dimensions in noncompensatory situations are somewhat dependent on the number of dimensions as well as number of items associated with dimensions. *ALR* and $\chi^2_{G/D}$ tended to be the most accurate methods in conditions that had 10 items per dimension and where true dimensionality was 2D rather than 3D. RMSR tended to benefit from the increase in both items and dimensions; however, given that RMSR method generally performed suboptimally, it is not recommended to use for most situations examined in this study. RMSR outperformed other methods only in a small number of conditions – conditions in which the data followed a noncompensatory 3D MIRT model with 20 items per dimension, 0% of complexity with low to moderate correlations, and across complexity levels when

correlations were .30 or less. The DETECT-based methods did not perform as well in noncompensatory condition, and therefore might also not be optimal methods to determine dimensionality.

The marginal proportions for labeling sets of items as dimension-like were typically low, suggesting that the methods generally failed to label two (three) sets of items as dimension-like in 2D (3D) noncompensatory situations. In 2D conditions with 10 items per dimension, an increase in complexity resulted in the methods labeling two sets of items as dimension-like less often, and labeling one or none of the sets as dimension-like more often. Similar observation was made when the number of items increased to 20 per dimension, where RMSR and $\chi^2_{G/D}$ had the most success in labeling one set of items as dimension-like (compared to the rest of the methods which yielded low marginal proportions for labeling any set of items as dimension-like). When true dimensionality increased to 3D, all methods failed to label three sets of items as dimension-like across the sample size and correlations.

The DETECT-based methods were more consistent in classifying factorially simple items across complexity levels, sample sizes, and correlations. However, as complexity and correlation levels increased the classification rates for all methods decreased. An increase in the number of items did not affect the classification rates too much and patterns of behaviors of the methods remained consistent (i.e., the DETECT-based methods yielded higher consistency rates for

factorially simple items than NOHARM-based methods).

In most conditions, the DETECT-based methods classified factorially complex items equally or more consistent than the NOHARM-based methods. In particular, as complexity, the number of items, and the true dimensionality increased, the DETECT-based methods were notably more consistent than any NOHARM-based method.

Given the results of the noncompensatory conditions, if the researcher hypothesizes that the nature of the relationship between the constructs is indeed conjunctive, the methods of DETECT may not be appropriate. In those cases, the researcher should adopt other methods. As these results suggest, for noncompensatory situations, the NOHARM-based methods *ALR* or $\chi^2_{G/D}$should almost always be employed. For the most part they were comparable, with a slight edge to *ALR* in some cases. However, it should be noted that neither *ARL* nor $\chi^2_{G/D}$ yielded high proportions of labeling the sets of items as dimension-like, and classification rates for both factorially simple and factorially complex items were low across conditions. Therefore, despite the recommendation to use the NOHARM-based methods of either *ALR* or $\chi^2_{G/D}$, the results of the current study should be taken as initial understanding of noncompensatory MIRT in dimensionality assessment.

Where do we go from here? An exploratory approach to understanding the test dimensionality can be particularly useful in applications of newly

developed instruments, or in tests that measure a construct that invokes complex relationships between the examinees and the items where little is known about that complexity. Assumptions related to the nature of the relationship between the constructs also need to be determined by the researcher, because they may be important in the choice of the dimensionality method to assess the number of dimensions.

The current study has shed some light onto the performance of the methods in assessing multidimensional item responses. It is suggested that the selection of tools by the researcher may have an impact on what optimal solution is obtained given a variety of factors. For example, RMSR is not recommended for assessing dimensionality in general. For other methods, given that the methods examined showed to be stronger in some conditions and weaker in others, the selection of the dimensionality assessment method is not simple. Rather, it might depend on a number of factors or characteristics of the data.

Given the differences in the results for the compensatory and noncompensatory conditions, perhaps the most consideration should be given to understanding how the constructs combine in the item response process. If the researcher believes compensatory relationships hold, DETECTe should be used for assessing the number of dimensions, but should be used cautiously in interpreting the clusters of items if simple structure does not hold. If the researcher believes noncompensatory relationships hold, *ALR* should be used for assessing the number of dimensions. However, neither *ALR* nor any other method

is likely to yield groupings of items that can be accurately interpreted in terms of their true dimensional structure. Further, *ALR* resulted in selecting a unidimensional solution most often of the three NOHARM-based methods (recall that DETECT yielded no solutions that favored one factor). These results speak to *ALR*'s tendency to under-factor more often, particularly in situations where the correlation levels increased.[8]

Thus, a general recommendation is that multiple sources be used in evaluating dimensionality of an assessment, particularly when the complexity in the item responses is present. Using multiple sources and triangulation of results might provide a firmer support for appropriate score interpretation.

Although this work builds on the existing literature in dimensionality assessment for compensatory MIRT, it also presents first insights into performances of the studied methods in dimensionality assessment of noncompensatory data. As suggested in many of the conditions with a noncompensatory model, the investigated methods may have limits in their suitability. This may have larger implications, particularly with an increase in cognitive diagnostic assessments. These types of testing scenarios call for a need in better understanding of the procedures for noncompensatory data. Specifically,

[8] As noted in the final section of *future directions*, further examination about performances of all methods when erring is warranted.

we need better tools to evaluate internal structures of instruments which we observe and measure that may assume relationships that are not compensatory.

Impact and Contributions

The results of this study contribute to a better understanding of how the exploratory approaches of methods based on DETECT and NOHARM perform in the evaluation of test dimensionality, specifically when the data exhibits complexity. The current study brings both methodological and practical contributions to the area of dimensionality assessment. Methodologically, there are two main contributions. One, the impact of complexity in dimensionality assessment is a relatively unexplored area. Two, there is a general lack of research on the NOHARM and The DETECT-based methods when the underlying MIRT model is noncompensatory, an issue addressed to some extent in the current study.

In practice, this study's results are meaningful in several ways. The topic of dimensionality assessment has explicit connections to the issues in practical assessment, such as design, scoring, and interpretation. In test design, a researcher may be concerned with specifications of the content domain, item format, as well as the process of item construction (Tate, 2002). In all of these processes, being aware of test dimensionality is important because potential consequences might arise if wrongful assumptions about test dimensionality are made.

For example, let us assume that a new science test is developed where the assessment is viewed as multidimensional and complexities in the data responses

are expected (see Leighton, Gokiert, & Cui, 2007 for a detailed example of science assessment and multidimensional complexities within). During the item design process for a science reasoning assessment, an item writer may create a science item such that it taps into multiple aspects of proficiency in scientific reasoning (e.g., selective encoding and comparison processes in inductive reasoning and selective combination processes in deductive reasoning).

For such an item, evidence supporting its complexity could be gathered by utilizing factor analytic techniques to dimensionality assessment. Using a nonlinear factor analytic procedure, such as NOHARM, may indeed be appropriate to investigate the (intended) item's relationship with latent factors. However, before a technique is used to examine the item's relationship to the constructs of interest (and thus providing evidence or lack thereof in the validation process), the method itself ought to be shown to perform well.

This study's results alert us to some circumstances where the methods performed suboptimally in selecting the correct dimensional structure. This, in turn, may implicate how the item's relationship to the constructs is interpreted. If the methods err in identifying the true underlying number of constructs, the associations of items to those constructs may be questionable.

The results of the current study relate to scoring and interpretation processes of the test in a more straightforward way. Scoring and interpretation of the scores of an assessment are both tied to the process of comprehensive

validation. According to the AERA, APA, and NCME standards, if a test provides more than one score (e.g., subscores), “the interrelationships of those scores should be shown to be consistent with the construct(s) being assessed” (p. 20). This calls for providing evidence for the internal structure of the test, and the dimensionality assessment is precisely tasked to do so. In other words, as researchers and test developers, we seek to find evidence and support for a particular score interpretation of an assessment.

One aspect of that is to examine and evaluate whether the internal structure of the test reflects the intended construct(s), which in turn informs how the test scores are reported. If a test is scored and reported using subscales, the interpretation of the multiple scores implies that multiple constructs are measured by the test. These interpretations are only meaningful if the internal structure of the assessment and intended construct(s) align (i.e., support for multidimensionality is gathered).

The current study evaluated five popular methods currently used in dimensionality assessment that can provide support for this alignment. More specifically, the evaluation of the methods was conducted for situations that involve possibly factorially complex multidimensional assessments, the type of assessments that are becoming more popular in current educational settings.

The results suggest that the methods of NOHARM and DETECT indeed may be useful and appropriate tools for dimensionality assessment in some of

these complex testing scenarios. In particular, the conditions that yield data with a simple structure and have assumed compensatory relationships among the constructs and items are well suited for the application of the procedures studied here. However, the results of the current study also suggest that these methods (as they currently operate) may not serve well in dimensionality assessment as our assessments become more complex and multi-layered.

Limitations of the Study

The current study has several limitations, some of which are related to the procedures themselves, while others are reflective of the design of the study.

The limitation of the procedures mainly points to the estimation and nonconvergence issues related to NOHARM. As the number of items and/or examinees increased, the estimation time for NOHARM became rather lengthy, and it resulted in more occurrences of failure to obtain the reliable estimates (i.e., nonconvergence).

Nonconvergence was also observed in particular datasets. As discussed in detail in the previous chapter, the estimator implemented in NOHARM cannot handle perfect item response vectors. This can be problematic in several testing scenarios, particularly, when the tests are short or the sample size is small. For instance, if a measure is short (e.g., as a screen test or in a pilot setting) and/or the population of interest is particular (e.g., a special population of severely depressed individuals), an endorsement of all (or none) of the items may be a plausible

event. For those cases, the NOHARM procedure cannot be used to assess dimensionality as the model estimates cannot be obtained. The DETECT-based methods do not suffer from the presence of perfect response vectors

As with other factor analytical approaches and procedures, the application of NOHARM requires a researcher to determine the optimal number of factors to be extracted. Although three methods based on NOHARM output were investigated in the current study ($\chi^2_{G/D}$, *ALR*, and RMSR), more research is needed to arrive to a consensus which of the three, if any, is most suitable. While the current results are consistent with previous research to some extent by finding support for *ALR* or $\chi^2_{G/D}$, both of these methods performed suboptimally in some scenarios (e.g., $\chi^2_{G/D}$ tended to identify the true number of dimensions less often in conditions with a correlation between dimension of .75 and when N = 2000).

The data characteristics contribute to the limitations of the procedures on two other fronts: completeness of the data and binary scoring of the item responses. Both DETECT and NOHARM, as standalone procedures applied in the current study, can only accommodate complete data. In other words, cases with missing item responses cannot be used in estimation. The effects of missing data techniques on the performance for either method are largely unknown.

A more general limitation of the study pertains to the choice of item response scoring. Only dichotomously data were considered in the study, as is assumed in both NOHARM and DETECT. (An extension of DETECT for

polytomous data exists, but is currently not commercially available.) Current assessments, however, more frequently utilize different item formats, supporting dichotomous and polytomous scoring.

Several other limitations pertain to the current study. In the current study, only the 2PL MIRT models for data generation were considered. Although a rationale to model data without the pseudo-guessing parameter present may be justified, omitting it limits the generalizability of the results.[9]

Similarly, only one set of item parameters was chosen for all conditions; previous literature found differences in DETECT and NOHARM's performance when different sets of item parameters were used (Finch & Habing, 2005). This implies that the generalizability of the results to other tests that pose different item parameter characteristics may be limited.

Furthermore, in the current study, simple and complex structures were considered. One could argue that the approximate simple structure would be a more realistic choice, thus suggesting a limitation of the baseline use of the exact simple structure. In addition, only 2D and 3D structures were considered.

[9] One justification is that not much is known about noncompensatory MIRT and dimensionality assessment of complex data. Thus, it was vital to first understand the performance of the methods in conditions that were "more" ideal, before introducing additional sources of complexity such as a pseudo-guessing parameter.

Although results for these dimensionality structures allow for some comparison with previous research, performance of the methods under in higher conditions when complexity in the data is present remains unknown (e.g., Finch & Habing, 2005, study compared two- and six-dimensional structures when evaluating DETECT and the NOHARM-based methods).

The choices for sample sizes and test lengths were largely based on the previous literature; however, they limit the generalizability of the results. For example, the situations with tests shorter than 20 total items or sample sizes of less than 500 were not considered, and thus, the conclusions for such testing scenarios cannot be provided, although such testing scenarios are very plausible in some settings (e.g., pilot studies, attitude measures, etc.).

Future Directions and Conclusion

Given the limitations of this study (and general constraints of the methods themselves), future research in dimensionality assessment is warranted. In addition to understanding how the methods performed, it will be important to further understand their performance through an investigation of errors they made. Thus, future directions of this line of research would involve examining over- and under-factoring of the methods when they erred.

Additionally, future work may involve inclusion of the pseudo-guessing parameter often modeled in multiple-choice items. Furthermore, different sets of item parameters may influence the performance of the methods, thus for

generalizability purposes, it would be beneficial to compare the current results with the results based on a different set of item parameters.

Future research should also focus on how to deal with different data, including polytomously scored and missing data, which are often found in educational assessments. A better understanding of how the current methods are impacted by various applications of missing data techniques may allow for more inclusion of data that are not complete when assessing multidimensionality.

Although the scenarios considered in this study included only those when the researcher has no *a priori* hypotheses of test dimensionality, a confirmatory approach to examine the methods performance should also be considered, as both NOHARM and DETECT have confirmatory capabilities.

The final and perhaps most important step forward is to continue research on how NOHARM, DETECT, and other methods used in dimensionality assessment perform under noncompensatory conditions. Given that most if not all methods are aligned with a compensatory nature of the relationship, it might be important to continue to investigate better options for dimensionality assessment in those conditions. Further developments in current and newly developed procedures that better align with the principles of noncompensatory relationships may be imperative as new complex assessments that assume such relationships get implemented. This issue needs to be addressed further, as we utilize dimensionality assessment as part of the comprehensive validation process that leads to appropriate score interpretations.

REFERENCES

Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement, 13,* 113-127.

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education, 7,* 255-278.

Ackerman, T. (1996). Graphical representation of multidimensional item response theory model. *Applied Psychological Measurement, 20,* 311-329.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Bock, R. D., Gibbons, R., Schilling, S. G., Muraki, E., Wilson, D. T., & Wood, R. (1999). *TESTFACT 3: Test scoring, items statistics, and full-information item factor analysis*. Chicago: Scientific Software International.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov Chain Monte Carlo. *Applied Psychological Measurement, 27,* 395-414.

Camilli, G., Wang, M., & Fesq, J. (1995). The effects of dimensionality on equating the law school admission test. *Journal of Educational Measurement, 32,* 79-96.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1,* 245-276.

Cattell, R. B. (1978). *The scientific use of factor analysis in behavioral and life sciences.* New York: Plenum.

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22,* 265-289.

Cheng, C.-P., & Weng, L.-J. (2005). Parallel analysis with unidimensional binary data. *Educational and Psychological Measurement, 65,* 697-716.

Childs, R. A. & Oppler, S. H. (2000). Implications of test dimensionality for unidimensional IRT scoring: An investigation of a high-stakes testing program. *Educational and Psychological Measurement, 60* (6), 939-955.

Cliff, N. (1988). The eigenvalues-greater-than-one rule and the reliability of components. *Psychological Bulletin, 103,* 276-279.

Cook, L. L., Dorans, N. J., & Eignor, D. R. (1988). An assessment of the dimensionality of three SAT-verbal test editions. *Journal of Educational Statistics, 13,* 19-43.

Crawford, A. V., Green, S. B., Levy, R., Lo, W. L., Scott, L., Svetina, D., & Thompson, M. S. (April, 2009). *Evaluation of Parallel Analysis Methods for Determining the Number of Factors.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA. Note: authors ordered alphabetically.

De Champlain, A. F. (1996). The effect of multidimensionality on IRT true-score equating for subgroups of examinees. *Journal of Educational Measurement, 33,* 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.

DiBello, L. V., Roussos, L. A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao and S. Sinharay (Eds.), *Handbook of statistics, Volume 26* (pp. 979-1030). North-Holland: Elsevier.

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

Embretson, S. E. (1997). Multicomponent response models. In W. van der Linden & R. Hambleton (Eds), *Handbook of Modern Item Response Theory* (pp.305-321). New York, NY: Springler-Verlag.

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*, 272-299.

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

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

Finch, H., & Habing, B. (April, 2003). *Performance of DIMTEST and NOHARM-based statistics for testing unidimensionality.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

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

Finch, H., & Habing, B. (2007). Performance of DIMTEST- and NOHARM-based statistics for testing unidimensionality. *Applied Psychological Measurement, 31,* 292-307.

Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. Multivariate Behavioral Research, 23, 267-269.

Fraser, C., & McDonald, R. P. (2003). User's guide NOHARM. A Windows program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory. Retrieved from NOHARM software.

Froelich, A. G., & Habing, B. (2008). Conditional covariance-based subtest selection for DIMTEST. *Applied Psychological Measurement, 32,* 138-155.

Froelich, A. G., & Stout, W. (2003). *A new bias correction method for the DIMTEST procedure.* Manuscript submitted for publication.

Gessaroli, M. E., & De Champlain, A. F. (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, 33,*157-192.

Gessaroli, M. E., De Champlain, A. F., & Folske. (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.

Gierl, M. J., Leighton, J. P., & Tan, X. (2006). Evaluating DETECT classification accuracy and consistency when data display complex structure. *Journal of Educational Measurement, 43*, 265-289.

Goldstein, H. (1980). Dimensionality, bias, independence and measurement scale problems in latent trait test score models. *British Journal of Mathematical and Statistical Psychology, 33*, 234-246.

Gorsuch, R. (1991). UniMult guide. Altadena, CA: UniMult.

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guide: Review of current practices and suggestions for future research. *Applied Measurement in Education, 17*(2), 145-220.

Green, S. B. (1983). Identifiability of spurious factors using linear factor analysis with binary items. *Applied Psychological Measurement, 7,* 139-147.

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

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26,* 333-352.

Harwell, M., Stone, C. A., Hsu, T.-C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20,* 101-125.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9,* 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*, 1-14.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30,* 179-185.

Hubbard, R., & Allen, S. J. (1987). A cautionary note on the use of principal components analysis: Supportive empirical evidence. *Sociological Methods and Research, 16,* 301-308.

Ip, E. H. (2001). Testing for local dependency in dichotomous and polytomous item response models. *Psychometrika, 66*, 109-132.

Jang, E. E., & Roussos, L. (2007). An investigation into the dimensionality of TOEFL using conditional covariance-based nonparametric approach. *Journal of Educational Measurement, 44*, 1-22.

Junker, B. W. (1993). Conditional association, essential independence and monotone unidimensional item response models. *The Annals of Statistics, 21, 1*359-1378.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed). *Educational measurement* (4th edition). Washington, DC: American Council on Education/Praeger.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20,* 141-151.

Kim, H. R. (1994). New techniques for the dimensionality assessment of standardized test data (Doctoral dissertation, University of Illinois at Urbana-Champaign). Dissertation Abstracts International, 55-12B, 5598.

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

Leighton, J. P., Gokiert, R. J., & Cui, Y. (2007). Using exploratory and confirmatory methods to identify the cognitive dimension in a large-scale science assessment. *International Journal of Testing, 7,* 141-189.

Levy, R., Mislevy, R. J., & Sinharay, S. (2009). Posterior predictive model checking for multidimensionality in item response theory. *Applied Psychological Measurement, 33,* 519-537.

Levy, R. & Svetina, D. (2011). A Generalized dimensionality discrepancy measure for dimensionality assessment in multidimensional item response theory. *British Journal of Mathematical and Statistical Psychology, 64,* 208-232.

Levy, R., & Svetina, D. (May, 2010). *A Framework for Characterizing Dimensionality Assessment and Overview of Current Approaches.* Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.

Li, X. (2008). An Investigation of the item parameter drift in the examination for the certificate of proficiency in English (ECPE). *English Language Institute, 6,* 1-29.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading MA: Addison-Welsley Publishing Company.

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,* 51-71.

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

McDonald, R. P. (1994). Testing for approximate dimensionality. In D. Laveault, B. D. Zumbo, M. E. Gessaroli, & M. W. Boss (Eds.*), Modern theories in measurement: Problems and issues* (pp. 31-61). Ottawa, Canada: Edumetrics Research Group, University of Ottawa.

McDonald, R. P. (1997). Normal-ogive multidimensional model. In W. J. van der Linden, & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 257-269). New York: Springer-Verlag.

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

McDonald, R. P. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement, 24,* 99-114.

McDonald, R. P., & Ahlawat, K. S. (1974). Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology, 27,* 82-99.

McKinley, R., & Reckase, M. D. (1982). The use of the general Rasch model with multidimensional item response data (RR ONR82-1). Iowa City: American College Testing Program.

McLeod, D., Swygert, K. A., & Thissen, D. (2001). Factor analysis for item scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 189-215). Hillsdale, NJ: Lawrence Erlbaum.

Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics, 11,* 3-31.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A Brief Introduction to Evidence Centered Design*. (Technical Report RR-03-16). Princeton, NJ: Educational Testing Service.

Mislevy, R. J., & Bock, R. D. (1982). BILOG: Item analysis and test scoring with binary logistic models. Mooresville IN: Scientific Software.

Monahan, P. O., Stump, T. E., Finch, H., & Hambleton, R. K. (2007). Bias of exploratory and cross-validated DETECT index under unidimensionality. *Applied Psychological Measurement, 31,* 483-503.

Muthén, L. K., & Muthén, B. O. (1998-2006). Mplus User's Guide. Fourth Editions. Los Angeles, CA: Muthén & Muthén.

Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality. *Journal of Educational Measurement, 28*, 99-117.

Nandakumar, R. (1993). Assessing essential unidimensionality of real data. *Applied Psychological Measurement, 17*, 29-38.

Nandakumar, R., & Stout, W. (1993). Refinement of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics, 18*(1), 41–68.

Nandakumar, R., & Yu, F. (1996). Empirical validation of DIMTEST on nonnormal ability distributions. *Journal of Educational Measurement, 33*, 355-368.

Nandakumar, R., Yu, F., Li, H., & Stout, W. (1998). Assessing unidimensionality of polytomous data. *Applied Psychological Measurement, 22*, 99-115.

No Child Left Behind Act (NCLBA) of 2001, Pub L. No. 107-110, §_1111, 115 Stat. 1449-1453 (2002).

Preacher, K. J., & MacCallum, R. C. (2003). Repairing Tom Swift's electric factor analysis machine. *Understanding Statistics, 2,* 13-43.

R Development Core Team (2006). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

Rindskopf, D., & Rose, T. (1988). Some theory and applications of confirmatory second-order factor analysis. *Multivariate Behavioral Research, 23,* 51-67.

Reckase, M. D. (1985). Models for multidimensional tests and hierarchically structured training materials (research report No. ONR85-1).

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271-286). New York: Springer.

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

Reckase, M. D., Carlson, J. E., Ackerman, T. A., & Spray, J. A. (1986, June). *The interpretation of unidimensional IRT parameters when estimated from multidimensional data.* Paper presented at the Psychometric Society annual meeting, Toronto.

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

Reckase, M., Thompson, T., & Nering, M. (1997, June). *Identifying similar item content clusters on multiple test forms.* In T. Miller (Chair), *High-dimensional simulation of item response data for CAT research.* Symposium conducted at the annual meeting of the Psychometric Society, Gatlingburg, TN.

Reddon, J. R. (1985). MAPF and MAPS: subroutines for the number of principal components. *Applied Psychological Measurement, 9*, 97.

Revelle, W., & Rocklin, T. (1979). Very simple structure: An alternative procedure for estimating the number of interpretable 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*, 215-243.

Roussos, L. A., Stout, W., & Marden, J. (1998). Using new proximity measures with hierarchical cluster analysis to detect multidimensionality. *Journal of Educational Measurement, 35*, 1–30.

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research and Perspectives, 6*, 219-262.

Seraphine, A. E. (2000). The performance of DIMTEST when latent trait and item difficulty distributions differ. *Applied Psychological Measurement, 24*, 82-94.

Spray, J. A., Davey, T. C., Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1990). Comparison of two logistic multidimensional item response theory models (research report series No. ONR90-8): ACT.

Stone, C. A. & Yeh, C.-C. (2006). Assessing the dimensionality and factor structure of multiple-choice exams: An empirical comparison of methods using the Multistate Bar Examination. *Educational and Psychology Measurement, 66*(2), 193-214.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*, 589-617.

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

Stout, W., Froelich, A. G., & Gao, F. (2001). Using resampling methods to produce an improved DIMTEST procedure. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 357-375). New York: Springer-Verlag.

Stout, W., Habing, B., Douglas, J., Kim, H. R., Roussos, L., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement, 20*, 331-354.

Sympson, J. B. (1978). A model for testing with multidimensional items. In D. J. Weiss (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference* (pp. 82-89). Minneapolis: University of Minneapolis, Department of Psychology, Psychometric Methods Program.

Takane, Y. and De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*(3), 393-408.

Tate, R. L. (2002). Test dimensionality. In J. Tindal & T. M. Haladyna (Eds.), *Large scale assessment programs for all students: Development, implementation, and analysis*. Mahwah, NJ: Lawrence Erlbaum.

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

Tran, U. S., & Formann, A. K. (2009). Performance of Parallel Analysis in Retrieving Unidimensionality in the Presence of Binary Data. *Educational and Psychological Measurement, 69,* 50-61.

van Abswoude, A. A., van der Ark, L., & Sijtsma, K. (2004). A comparative study of test data dimensionality assessment procedures under nonparametric IRT models. *Applied Psychological Measurement, 28,* 3–24.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika, 41,* 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, and E. Helmes (Eds.), *Problems and Solutions in Human Assessment: Honoring Douglas Jackson at Seventy* (pp. 41-71). Boston: Kluwer.

Walker, C. M., & Beretvas, S. N. (2003). Comparing multidimensional and unidimensional proficiency classifications: Multidimensional IRT as a diagnostic aid. *Journal of Educational Measurement, 40,* 255-275.

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,* 239-252.

Wilson, D., Wood, R. L., & Gibbons, R. (1991). TESTFACT 2 [Computer software.] Chicago: Scientific Software International.

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,* 354-365.

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

Yang, X., &Zhang, J. (April, 2001). Construction and evaluation of bias-corrected estimators of DETECT dimensionality index. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Yao, L., & Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement, 31,* 83-105.

Yeh, C.-C. (2007). The effect of guessing on assessing dimensionality in multiple-choice tests: A Monte Carlo study with application. Unpublished dissertation. University of Pittsburg.

Yen, W. M. (1985). Increasing item complexity: A possible cause of scale shrinkage for unidimensional item response theory. *Psychometrika, 50*, 399-410.

Zhang, J. (2007). Conditional covariance theory and detect for polytomous items. *Psychometrika, 72,* 69-91.

Zhang, J., & Stout, W. (1999a). Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika, 64*, 129-152.

Zhang, J., & Stout, W. (1999b). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64*, 213-249.

Zwick, W. R., & Velicer, W. F. (1982). Variables 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,* 432-442.

APPENDIX A

TABULAR RESULTS FOR PROPORTION CORRECT

Table A1

Tabulated Results of Proportion Correct for Conditions where Data Follow 2D Compensatory MIRT with 10 Items per Dimension

		0% Complexity					10% Complexity					30% Complexity					50% Complexity				
N	ρ	*ALR*	$\chi^2_{G/D}$	R	De	Dcv	*ALR*	$\chi^2_{G/D}$	R	De	Dcv	*ALR*	$\chi^2_{G/D}$	R	De	Dcv	*ALR*	$\chi^2_{G/D}$	R	De	Dcv
500	.00	.93	1.00	.04	1.00	1.00	.93	1.00	.03	1.00	1.00	.93	1.00	.04	1.00	.99	.71	1.00	.02	1.00	.91
	.30	.89	1.00	.02	1.00	1.00	.91	1.00	.03	1.00	.98	.90	1.00	.02	1.00	.94	.41	.99	**	.99	.78
	.60	.87	1.00	.03	1.00	.92	.90	1.00	.03	1.00	.87	.84	1.00	.03	.98	.68	.22	.94	-	.85	.50
	.75	.88	1.00	.03	.97	.71	.88	1.00	.02	.93	.59	.85	1.00	.01	.84	.37	.28	.84	**	.53	.23
	.90	.80	.73	.02	.50	.14	.70	.53	.01	.42	.18	.44	.11	.01	.12	.13	.63	.45	**	.08	.12
1000	.00	.91	1.00	.03	1.00	1.00	.91	1.00	.02	1.00	1.00	.95	1.00	.04	1.00	1.00	.25	.97	**	1.00	.99
	.30	.89	1.00	.04	1.00	1.00	.92	1.00	.04	1.00	1.00	.93	1.00	.03	1.00	1.00	.04	.75	-	1.00	.95
	.60	.89	1.00	.03	1.00	.99	.91	1.00	.03	1.00	.98	.83	1.00	.01	1.00	.92	.03	.26	-	.96	.70
	.75	.89	1.00	.02	1.00	.90	.89	1.00	.02	.99	.84	.75	.99	-	.96	.59	.05	.11	-	.70	.36
	.90	.86	.99	.01	.67	.33	.85	.95	.01	.59	.29	.49	.40	-	.23	.12	.58	.97	**	.06	.11
2000	.00	.83	1.00	.04	1.00	1.00	.85	.99	.04	1.00	1.00	.94	1.00	.03	1.00	1.00	**	.59	-	1.00	1.00
	.30	.82	1.00	.04	1.00	1.00	.82	1.00	.03	1.00	1.00	.81	1.00	.01	1.00	1.00	-	.03	-	1.00	1.00
	.60	.82	.99	.02	1.00	1.00	.83	.99	.03	1.00	1.00	.48	.96	-	1.00	.99	-	-	-	1.00	.91
	.75	.78	.99	.04	1.00	.99	.79	.99	.01	1.00	.96	.23	.91	-	1.00	.89	**	-	-	.85	.55
	.90	.56	.95	**	.88	.57	.64	.93	.01	.77	.47	.14	.68	-	.30	.14	.24	.75	-	.02	.08

Note: R stands for RMSR; a method based on NOHARM output. De stands for DETECT exploratory; a method based on DETECT procedure. Dcv stands for DETECT cross-validated; a method based on DETECT procedure. . "-" indicates actual zero correct; "**" indicates < .01 proportion correct.

Table A2

Tabulated Results of Proportion Correct for Conditions where Data Follow 3D Compensatory MIRT with 10 Items per Dimension

		0% Complexity					10% Complexity					30% Complexity					50% Complexity				
N	ρ	*ALR*	$\chi^2_{G/D}$	R	De	Dcv	*ALR*	$\chi^2_{G/D}$	R	De	Dcv	*ALR*	$\chi^2_{G/D}$	R	De	Dcv	*ALR*	$\chi^2_{G/D}$	R	De	Dcv
500	.00	.94	.99	.64	1.00	1.00	.96	1.00	.69	1.00	1.00	.96	1.00	.56	1.00	.99	.58	.73	.02	1.00	.98
	.30	.96	1.00	.68	1.00	1.00	.96	1.00	.64	1.00	1.00	.74	1.00	.47	1.00	.98	.60	.16	-	.99	.95
	.60	.99	1.00	.59	1.00	.98	.98	1.00	.61	1.00	.97	.47	1.00	.48	.99	.88	.24	.12	**	.94	.73
	.75	.97	1.00	.58	1.00	.90	.97	1.00	.56	.99	.90	.43	.89	.46	.97	.67	.40	.54	.25	.78	.55
	.90	.60	.21	.47	.80	.59	.44	.07	.37	.73	.59	.24	**	.12	.43	.54	.18	.02	.17	.34	.51
1000	.00	.95	.99	.75	1.00	1.00	.98	.99	.70	1.00	1.00	.97	1.00	.46	1.00	1.00	.62	.05	-	1.00	1.00
	.30	.99	1.00	.70	1.00	1.00	.97	1.00	.72	1.00	1.00	.85	1.00	.26	1.00	1.00	.65	-	-	1.00	.99
	.60	.99	1.00	.69	1.00	1.00	.99	1.00	.65	1.00	1.00	.39	.97	.29	1.00	.98	.23	-	-	.97	.89
	.75	.99	.99	.61	1.00	.99	.98	.99	.52	1.00	.98	.45	.89	.30	.98	.87	.28	.26	.09	.80	.69
	.90	.87	.87	.46	.93	.68	.78	.64	.29	.87	.61	.45	.09	.16	.45	.44	.33	.14	.24	.23	.45
2000	.00	.95	1.00	.72	1.00	1.00	.98	.99	.74	1.00	1.00	.95	1.00	.19	1.00	1.00	.71	-	-	1.00	1.00
	.30	.97	.99	.71	1.00	1.00	.98	1.00	.71	1.00	1.00	.88	.98	.02	1.00	1.00	.59	-	-	1.00	1.00
	.60	.99	1.00	.71	1.00	1.00	.98	1.00	.56	1.00	1.00	.31	.71	.07	1.00	1.00	.22	-	-	.98	.95
	.75	.99	.97	.58	1.00	1.00	.99	.95	.32	1.00	1.00	.43	.38	.18	1.00	.96	.20	.02	.03	.83	.78
	.90	.84	.80	.18	.99	.85	.85	.66	.05	.95	.78	.58	.43	.26	.67	.42	.48	.32	.23	.09	.37

Note: R stands for RMSR; a method based on NOHARM output. De stands for DETECT exploratory; a method based on DETECT procedure. Dcv stands for DETECT cross-validated; a method based on DETECT procedure. "-" indicates actual zero correct; "**" indicates < .01 proportion correct.

Table A3

Tabulated Results of Proportion Correct for Conditions where Data Follow 2D Compensatory MIRT with 20 Items per Dimension

		0% Complexity					10% Complexity					30% Complexity					50% Complexity				
N	ρ	*ALR*	$\chi^2_{G/D}$	R	De	Dcv	*ALR*	$\chi^2_{G/D}$	R	De	Dcv	*ALR*	$\chi^2_{G/D}$	R	De	Dcv	*ALR*	$\chi^2_{G/D}$	R	De	Dcv
500	.00	.81	.91	.98	1.00	1.00	.99	.91	.97	1.00	.98	.99	.97	.98	1.00	.92	.96	.61	.54	1.00	.71
	.30	.99	.93	.98	1.00	.99	.99	.93	.97	1.00	.96	1.00	.97	.93	1.00	.84	.85	.14	.08	.99	.61
	.60	.99	.96	.98	1.00	.90	.99	.96	.98	1.00	.81	.98	.91	.81	.99	.65	.55	.01	-	.91	.39
	.75	.98	.94	.97	.99	.63	.99	.96	.97	.97	.57	.97	.82	.63	.90	.38	.42	**	**	.64	.19
	.90	.79	.91	.95	.52	.15	.64	.87	.91	.46	.13	.22	.73	.46	.22	.06	.63	.80	.40	.11	.05
1000	.00	.80	.88	.98	1.00	1.00	.98	.87	.95	1.00	1.00	1.00	.95	.94	1.00	.99	.75	.05	.06	1.00	.96
	.30	.99	.89	.98	1.00	1.00	.97	.90	.97	1.00	1.00	.99	.89	.76	1.00	.97	.13	-	-	1.00	.88
	.60	.98	.91	.97	1.00	.99	.98	.90	.97	1.00	.98	.98	.66	.44	1.00	.91	.02	-	-	.99	.72
	.75	.99	.85	.95	1.00	.94	.98	.82	.92	1.00	.88	.95	.36	.16	.99	.73	.03	-	-	.91	.41
	.90	.97	.59	.81	.79	.34	.93	.55	.69	.73	.30	.52	.09	.02	.45	.10	.75	.20	.05	.15	.05
2000	.00	.65	.54	.87	1.00	1.00	.83	.60	.87	1.00	1.00	1.00	.82	.71	1.00	1.00	.06	-	-	1.00	1.00
	.30	.81	.60	.92	1.00	1.00	.84	.64	.90	1.00	1.00	.98	.46	.22	1.00	1.00	**	-	-	1.00	1.00
	.60	.85	.66	.90	1.00	1.00	.88	.60	.84	1.00	1.00	.81	.03	**	1.00	1.00	-	-	-	1.00	.95
	.75	.82	.48	.77	1.00	1.00	.81	.34	.61	1.00	.99	.57	**	-	1.00	.95	**	-	-	.99	.77
	.90	.77	.07	.29	.98	.65	.80	.06	.14	.95	.53	.36	-	-	.69	.31	.37	-	-	.26	.12

Note: R stands for RMSR; a method based on NOHARM output. De stands for DETECT exploratory; a method based on DETECT procedure. Dcv stands for DETECT cross-validated; a method based on DETECT procedure. "-" indicates actual zero correct; "**" indicates < .01 proportion correct.

Table A4

Tabulated Results of Proportion Correct for Conditions where Data Follow 3D Compensatory MIRT with 20 Items per Dimension

		0% Complexity					10% Complexity					30% Complexity					50% Complexity				
N	ρ	*ALR*	$\chi^2_{G/D}$	R	De	Dcv	*ALR*	$\chi^2_{G/D}$	R	De	Dcv	*ALR*	$\chi^2_{G/D}$	R	De	Dcv	*ALR*	$\chi^2_{G/D}$	R	De	Dcv
500	.00	.79	.83	1.00	1.00	1.00	.86	.87	1.00	1.00	.99	.91	.92	1.00	1.00	.97	.80	-	.06	1.00	.88
	.30	.96	.85	1.00	1.00	1.00	.96	.87	1.00	1.00	.99	.74	.85	1.00	1.00	.93	.47	-	-	1.00	.83
	.60	.98	.85	1.00	1.00	.97	.96	.87	1.00	1.00	.96	.41	.79	.99	1.00	.84	.35	-	**	.98	.70
	.75	.83	.84	1.00	1.00	.88	.76	.79	1.00	1.00	.81	.35	.47	.93	.98	.62	.55	.11	.44	.88	.49
	.90	.08	.63	.42	.88	.46	.07	.58	.17	.82	.47	.16	.44	-	.64	.43	.10	.13	**	.42	.39
1000	.00	.80	.77	1.00	1.00	1.00	.87	.79	1.00	1.00	1.00	.93	.80	.99	1.00	1.00	.70	-	-	1.00	1.00
	.30	.98	.79	1.00	1.00	1.00	.97	.81	1.00	1.00	1.00	.85	.70	.95	1.00	.99	.34	-	-	1.00	.96
	.60	1.00	.81	1.00	1.00	1.00	.99	.75	1.00	1.00	1.00	.45	.61	.91	1.00	.97	.33	-	-	1.00	.92
	.75	.97	.72	1.00	1.00	.99	.95	.53	1.00	1.00	.99	.51	.35	.91	1.00	.91	.65	.01	.08	.97	.75
	.90	.31	.18	.99	.98	.74	.22	.10	.87	.96	.67	.44	.51	.02	.81	.50	.34	.28	**	.53	.39
2000	.00	.80	.48	1.00	1.00	1.00	.89	.51	1.00	1.00	1.00	.95	.11	.60	1.00	1.00	.76	-	-	1.00	1.00
	.30	.99	.54	1.00	1.00	1.00	.99	.53	1.00	1.00	1.00	.92	.06	.20	1.00	1.00	.26	-	-	1.00	1.00
	.60	.99	.53	1.00	1.00	1.00	.99	.45	1.00	1.00	1.00	.45	.49	.81	1.00	1.00	.28	-	-	1.00	.99
	.75	.99	.21	1.00	1.00	1.00	1.00	.10	.95	1.00	1.00	.49	.04	.49	1.00	1.00	.70	-	.01	.96	.91
	.90	.45	.00	.71	1.00	.92	.44	-	.24	1.00	.89	.64	.15	.08	.92	.69	.62	.18	.02	.43	.46

Note: R stands for RMSR; a method based on NOHARM output. De stands for DETECT exploratory; a method based on DETECT procedure. Dcv stands for DETECT cross-validated; a method based on DETECT procedure. "-" indicates actual zero correct; "**" indicates < .01 proportion correct.

Table A5

Tabulated Results of Proportion Correct for Conditions where Data Follow 2D Noncompensatory MIRT with 10 Items per Dimension

		0% Complexity					10% Complexity					30% Complexity					50% Complexity				
N	ρ	*ALR*	$\chi^2_{G/D}$	R	De	Dcv	*ALR*	$\chi^2_{G/D}$	R	De	Dcv	*ALR*	$\chi^2_{G/D}$	R	De	Dcv	*ALR*	$\chi^2_{G/D}$	R	De	Dcv
500	.00	.71	.83	**	.32	.16	.72	.80	**	.28	.14	.76	.81	**	.06	.05	.77	.88	**	.13	.11
	.30	.72	.82	.02	.11	.07	.69	.78	**	.08	.06	.72	.81	**	.01	.04	.82	.88	**	.06	.08
	.60	.67	.76	**	.01	.02	.67	.70	**	**	.02	.55	.58	**	**	.01	.59	.62	**	**	.05
	.75	.58	.61	**	**	**	.54	.59	**	-	.01	.49	.47	**	-	.03	.44	.44	**	**	.07
	.90	.43	.46	.02	-	.01	.48	.57	**	-	.02	.53	.56	**	**	.03	.55	.55	.01	.01	.05
1000	.00	.72	.80	**	.61	.22	.67	.74	**	.49	.18	.67	.74	.01	.07	.09	.71	.81	**	.11	.14
	.30	.71	.76	**	.16	.08	.65	.71	**	.11	.08	.67	.74	**	.01	.02	.70	.81	**	.03	.09
	.60	.64	.72	.01	**	**	.54	.50	-	**	**	.56	.58	**	-	**	.53	.58	**	.01	.02
	.75	.65	.65	.01	-	**	.40	.37	**	**	**	.52	.61	**	**	**	.49	.57	-	-	.03
	.90	.53	.56	.02	-	**	.64	.78	**	-	**	.68	.88	.01	-	**	.75	.90	.01	**	.03
2000	.00	.56	.56	**	.84	.43	.59	.58	**	.68	.35	.55	.53	.01	.06	.06	.53	.61	**	.09	.12
	.30	.59	.58	-	.25	.13	.48	.44	**	.14	.07	.46	.44	**	-	.01	.50	.52	-	.01	.05
	.60	.50	.45	-	**	.01	.22	.15	**	-	.01	.21	.14	-	-	**	.17	.15	-	-	.01
	.75	.36	.28	**	-	-	.08	.02	-	**	-	.27	.25	**	-	**	.15	.13	-	-	**
	.90	.48	.55	**	-	-	.59	.68	**	-	-	.75	.91	**	-	-	.81	.96	**	-	.01

Note: R stands for RMSR; a method based on NOHARM output. De stands for DETECT exploratory; a method based on DETECT procedure. Dcv stands for DETECT cross-validated; a method based on DETECT procedure. "-" indicates actual zero correct; "**" indicates < .01 proportion correct.

Table A6

Tabulated Results of Proportion Correct for Conditions where Data Follow 3D Noncompensatory MIRT with 10 Items per Dimension

		0% Complexity					10% Complexity					30% Complexity					50% Complexity				
N	ρ	*ALR*	$\chi^2_{G/D}$	R	De	Dcv	*ALR*	$\chi^2_{G/D}$	R	De	Dcv	*ALR*	$\chi^2_{G/D}$	R	De	Dcv	*ALR*	$\chi^2_{G/D}$	R	De	Dcv
500	.00	.65	.71	.03	.07	.10	.70	.66	.02	.10	.15	.17	.23	.06	.04	.16	.26	.33	.11	.11	.23
	.30	.42	.49	.03	.03	.11	.41	.36	.03	.07	.17	.08	.15	.08	.03	.23	.19	.32	.16	.11	.34
	.60	.13	.17	.06	.03	.15	.12	.24	.03	.06	.19	.07	.13	.11	.04	.25	.14	.33	.18	.15	.39
	.75	.07	.09	.07	.01	.13	.10	.20	.04	.04	.19	.11	.16	.13	.03	.32	.18	.34	.25	.12	.39
	.90	.05	.10	.08	.04	.16	.11	.22	.05	.05	.26	.13	.18	.14	.05	.32	.19	.31	.23	.15	.43
1000	.00	.87	.86	.02	.04	.08	.89	.79	.03	.02	.11	.49	.50	.04	**	.05	.59	.41	.11	.05	.16
	.30	.79	.79	.02	**	.06	.72	.62	.01	.04	.11	.21	.23	.06	**	.09	.32	.35	.16	.04	.20
	.60	.32	.30	.05	.00	.04	.28	.33	.00	.02	.12	.19	.20	.10	.00	.11	.30	.36	.24	.02	.27
	.75	.12	.08	.08	**	.06	.24	.27	.03	**	.09	.19	.18	.14	**	.10	.32	.30	.27	.03	.25
	.90	.10	.08	.09	.00	.07	.24	.19	.08	**	.09	.21	.16	.15	.00	.15	.32	.25	.26	.02	.30
2000	.00	.85	.89	.02	.02	.05	.86	.81	.04	.02	.05	.81	.78	.08	.00	.01	.87	.46	.13	.00	.07
	.30	.86	.85	.05	**	.02	.80	.47	**	**	.09	.44	.43	.04	.00	.03	.60	.27	.07	.00	.08
	.60	.62	.64	.03	.00	.02	.51	.24	.00	**	.04	.41	.35	.05	.00	.02	.56	.43	.11	.00	.09
	.75	.25	.14	.07	.00	.01	.48	.39	**	.00	.03	.32	.17	.13	.00	.01	.40	.34	.25	.00	.11
	.90	.09	.05	.12	.00	**	.34	.17	.07	**	.03	.25	.10	.18	.00	.01	.29	.24	.30	.00	.09

Note: R stands for RMSR; a method based on NOHARM output. De stands for DETECT exploratory; a method based on DETECT procedure. Dcv stands for DETECT cross-validated; a method based on DETECT procedure. "-" indicates actual zero correct; "**" indicates < .01 proportion correct.

Table A7

Tabulated Results of Proportion Correct for Conditions where Data Follow 2D Noncompensatory MIRT with 20 Items per Dimension

		0% Complexity					10% Complexity					30% Complexity					50% Complexity				
N	ρ	*ALR*	$\chi^2_{G/D}$	R	De	Dcv	*ALR*	$\chi^2_{G/D}$	R	De	Dcv	*ALR*	$\chi^2_{G/D}$	R	De	Dcv	*ALR*	$\chi^2_{G/D}$	R	De	Dcv
500	.00	.58	.21	.87	.65	.21	.85	.18	.87	.58	.20	.97	.23	.83	.26	.14	.95	.24	.84	.25	.17
	.30	.94	.19	.89	.29	.11	.97	.16	.83	.25	.13	.87	.20	.76	.10	.06	.81	.23	.74	.13	.15
	.60	.87	.16	.86	.02	.04	.85	.05	.62	.04	.06	.26	.09	.40	.01	.02	.19	.08	.40	.06	.06
	.75	.29	.13	.83	**	.03	.25	.02	.42	.01	.02	.11	.16	.41	**	.02	.07	.15	.27	.03	.04
	.90	.05	.22	.26	-	.01	.08	.28	.64	**	.03	.12	.56	.87	**	.02	.15	.72	.83	**	.03
1000	.00	.54	.12	.81	.92	.44	.87	.11	.85	.92	.39	.92	.08	.78	.53	.18	.95	.09	.67	.52	.22
	.30	.89	.08	.83	.64	.15	.91	.04	.67	.50	.15	.94	.11	.56	.14	.09	.90	.04	.53	.17	.17
	.60	.92	.08	.76	.05	.05	.89	**	.18	.04	.04	.56	**	.06	**	.02	.43	-	.06	.03	.09
	.75	.67	.02	.55	**	.01	.59	-	.02	**	.02	.30	-	.04	-	**	.27	-	**	**	.02
	.90	.10	.13	.45	-	-	.29	.20	.61	-	**	.37	.58	.90	-	**	.47	.63	.79	-	**
2000	.00	.41	.01	.55	1-	.81	.67	**	.47	.99	.73	.77	**	.46	.81	.38	.73	-	.19	.73	.37
	.30	.70	**	.54	.92	.42	.75	-	.22	.82	.32	.66	-	.07	.22	.11	.61	-	.06	.28	.14
	.60	.64	**	.29	.10	.02	.52	-	-	.07	.05	.41	-	-	**	**	.29	-	-	.01	.05
	.75	.64	-	.06	-	**	.37	-	-	-	**	.55	-	-	-	-	.48	-	-	-	.01
	.90	.27	**	.12	-	-	.77	.01	.11	**	**	.86	.37	.79	-	-	.95	.32	.47	-	**

Note: R stands for RMSR; a method based on NOHARM output. De stands for DETECT exploratory; a method based on DETECT procedure. Dcv stands for DETECT cross-validated; a method based on DETECT procedure. "-" indicates actual zero correct; "**" indicates < .01 proportion correct.

Table A8

Tabulated Results of Proportion Correct for Conditions where Data Follow 3D Noncompensatory MIRT with 20 Items per Dimension

		0% Complexity					10% Complexity					30% Complexity					50% Complexity				
N	ρ	*ALR*	$\chi^2_{G/D}$	R	De	Dcv	*ALR*	$\chi^2_{G/D}$	R	De	Dcv	*ALR*	$\chi^2_{G/D}$	R	De	Dcv	*ALR*	$\chi^2_{G/D}$	R	De	Dcv
500	.00	.16	.39	.93	.20	.34	.16	.30	.95	.31	.34	**	.27	.92	.18	.25	.01	.08	.87	.34	.30
	.30	.04	.33	.96	.19	.25	.01	.15	.94	.27	.31	-	.21	.26	.12	.24	**	.04	.11	.27	.34
	.60	-	.27	.63	.10	.19	-	.09	.35	.19	.24	-	.28	.04	.08	.27	-	.10	**	.16	.29
	.75	-	.23	.07	.07	.15	-	.24	.12	.08	.28	-	.38	.02	.09	.30	**	.14	**	.18	.29
	.90	-	.22	-	.05	.20	**	.27	.03	.07	.26	-	.34	**	.11	.31	**	.25	-	.20	.35
1000	.00	.36	.60	.97	.16	.30	.37	.49	.98	.20	.36	.05	.34	.98	.05	.21	.10	.09	.99	.31	.32
	.30	.17	.52	.97	.07	.25	.16	.15	.90	.12	.33	-	.12	.70	.03	.17	.01	**	.54	.22	.27
	.60	**	.47	.98	.08	.15	**	**	.25	.09	.21	-	.32	.18	**	.20	**	.09	.10	.10	.28
	.75	**	.38	.43	.03	.13	**	.20	.33	.04	.17	**	.37	.07	**	.19	**	.23	**	.07	.30
	.90	-	.23	**	.01	.12	-	.36	.04	.01	.13	**	.40	-	**	.21	**	.33	-	.07	.27
2000	.00	.44	.76	.99	.17	.19	.50	.54	.98	.18	.29	.26	.33	.98	**	.07	.39	.06	.99	.16	.32
	.30	.44	.71	.99	.04	.13	.42	.03	.70	.07	.20	.06	.01	.29	-	.07	.15	**	.26	.11	.25
	.60	.09	.58	.99	**	.11	.06	-	-	.03	.18	.02	.12	.20	-	.04	.09	.01	.28	.01	.13
	.75	**	.51	.95	**	.05	.04	.05	.05	.01	.08	**	.43	.14	-	.05	.02	.19	.02	**	.14
	.90	-	.23	-	-	.03	.01	.39	.07	**	.04	.01	.29	**	-	.05	.01	.33	-	**	.17

Note: R stands for RMSR; a method based on NOHARM output. De stands for DETECT exploratory; a method based on DETECT procedure. Dcv stands for DETECT cross-validated; a method based on DETECT procedure. "-" indicates actual zero correct; "**" indicates < .01 proportion correct.

APPENDIX B

GRAPHICAL RESULTS FOR LABELING SETS OF ITEMS AS DIMENSION-LIKE

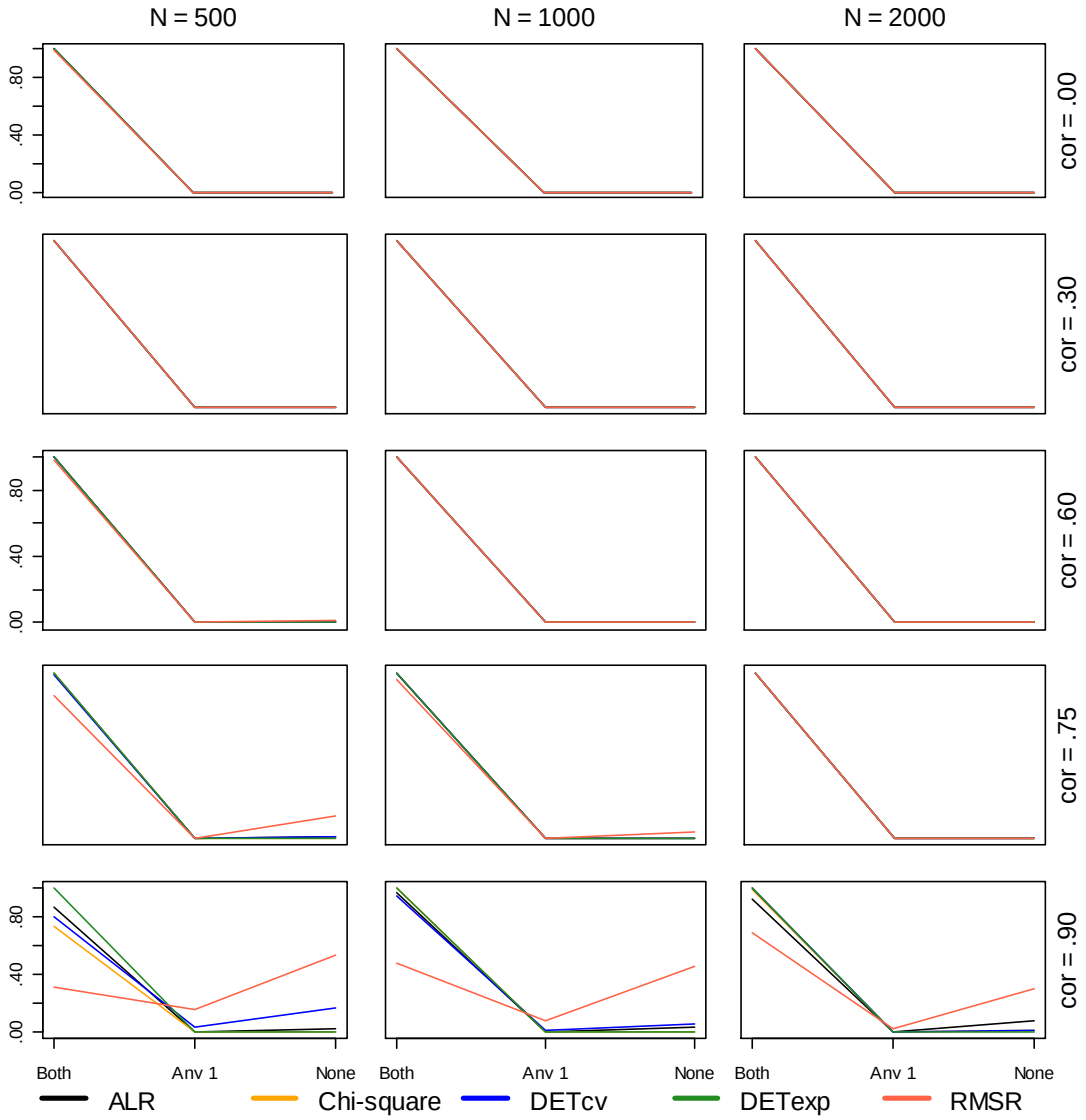

Figure B1. Marginal proportions across 500 replications that a method identified two (both), (any) one, or none of the sets of items as dimension-like (x-axis) when the data exhibit 0% complexity and follow a compensatory 2D MIRT model with 10 items per dimension.

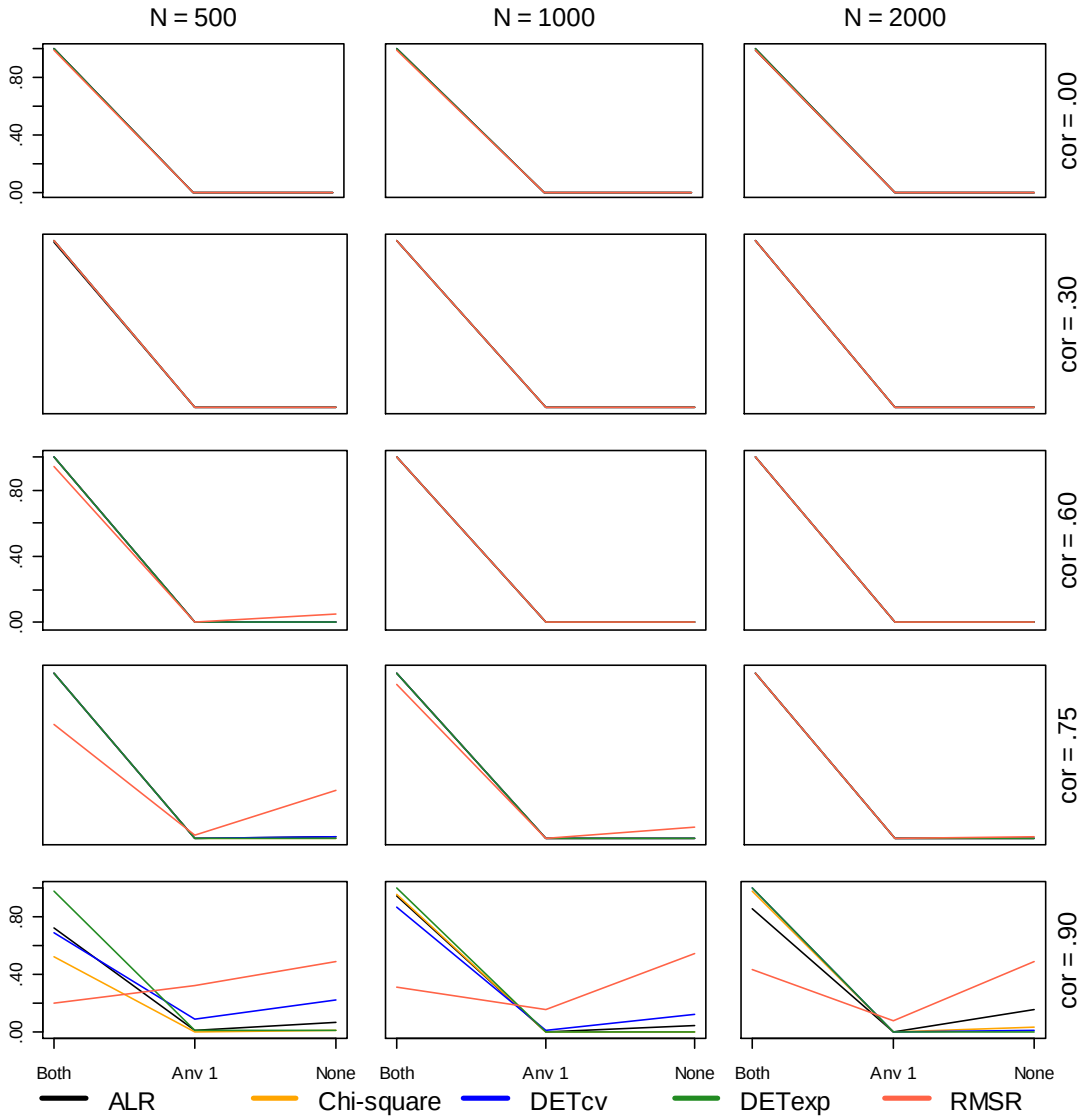

Figure B2. Marginal proportions across 500 replications that a method identified two (both), (any) one, or none of the sets of items as dimension-like (x-axis) when the data exhibit 10% complexity and follow a compensatory 2D MIRT model with 10 items per dimension.

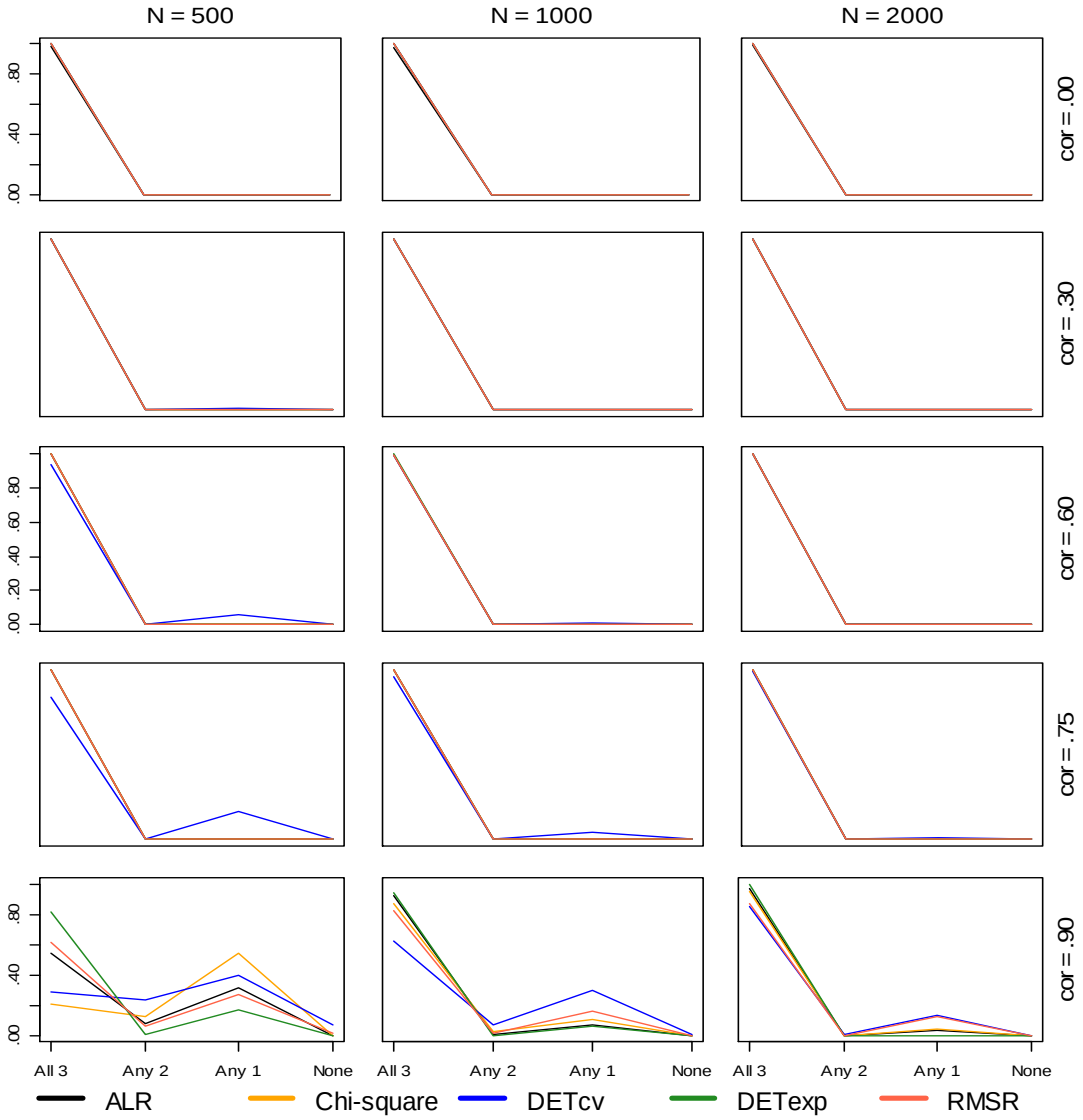

Figure B3. Marginal proportions across 500 replications that a method identified three, any two (both), (any) one, or none of the sets of items as dimension-like (x-axis) when the data exhibit 0% complexity and follow a compensatory 3D MIRT model with 10 items per dimension.

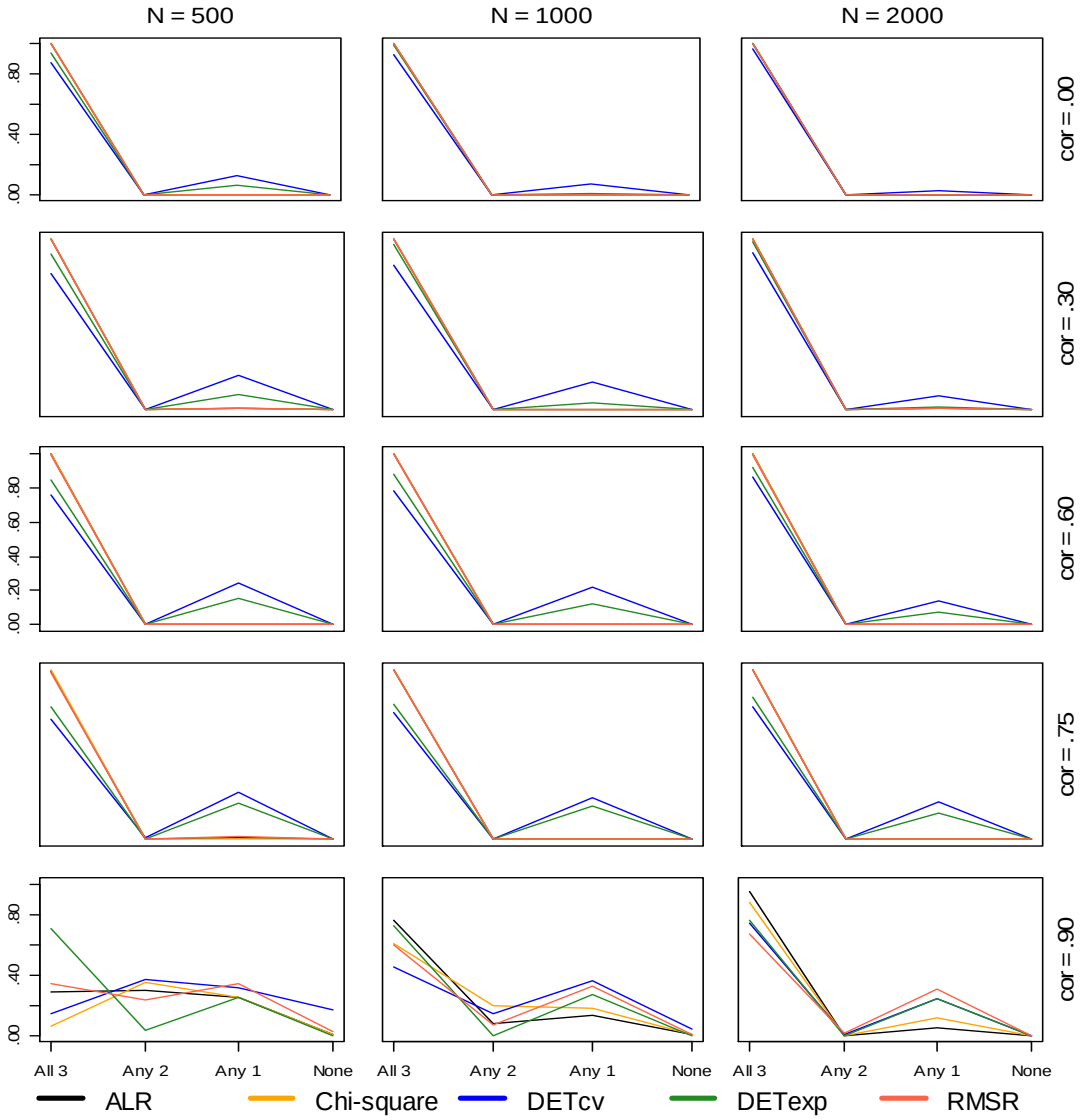

Figure B4. Marginal proportions across 500 replications that a method identified three, any two (both), (any) one, or none of the sets of items as dimension-like (x-axis) when the data exhibit 10% complexity and follow a compensatory 3D MIRT model with 10 items per dimension.

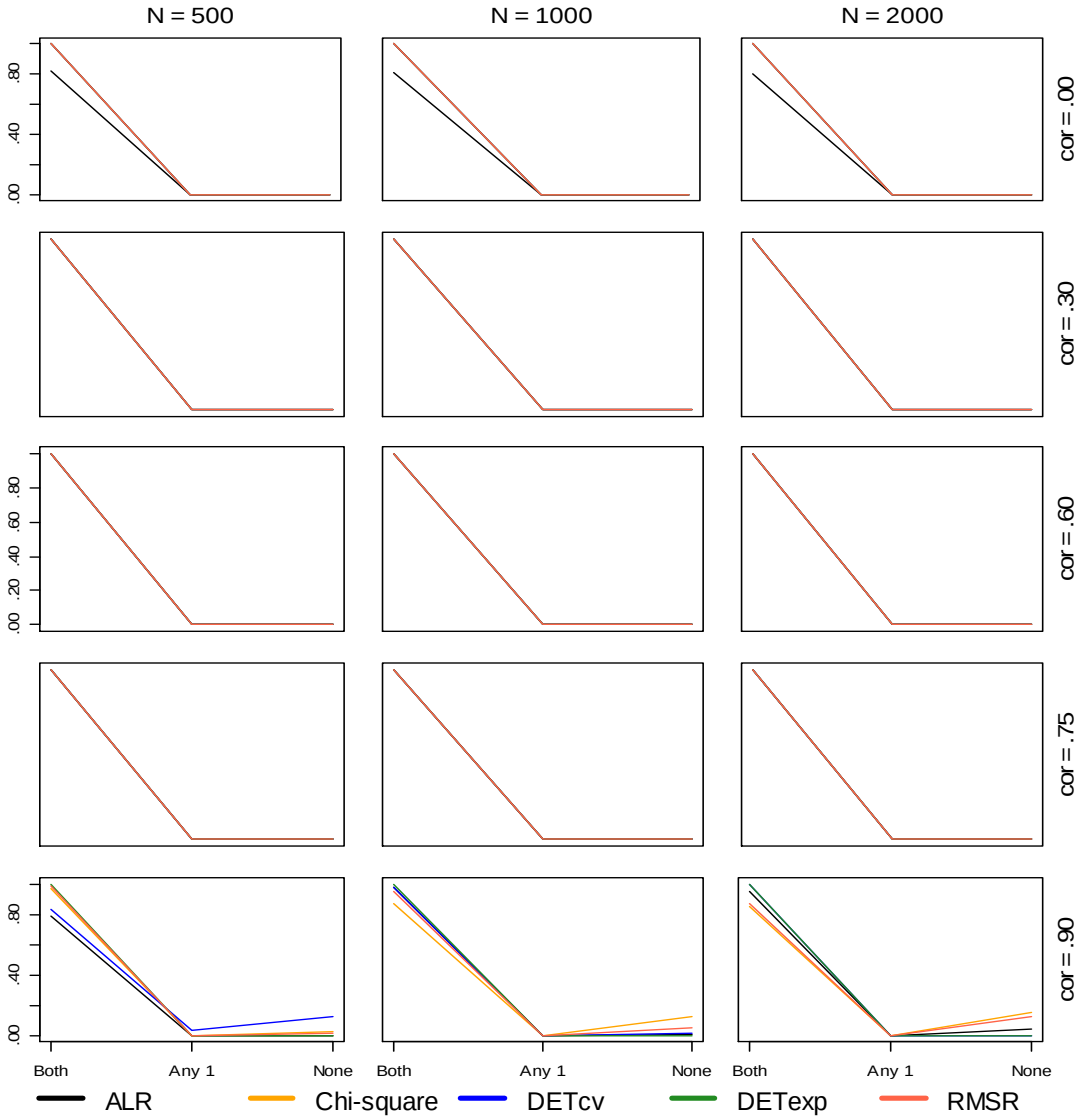

Figure B5. Marginal proportions across 500 replications that a method identified two (both), (any) one, or none of the sets of items as dimension-like (x-axis) when the data exhibit 0% complexity and follow a compensatory 2D MIRT model with 20 items per dimension.

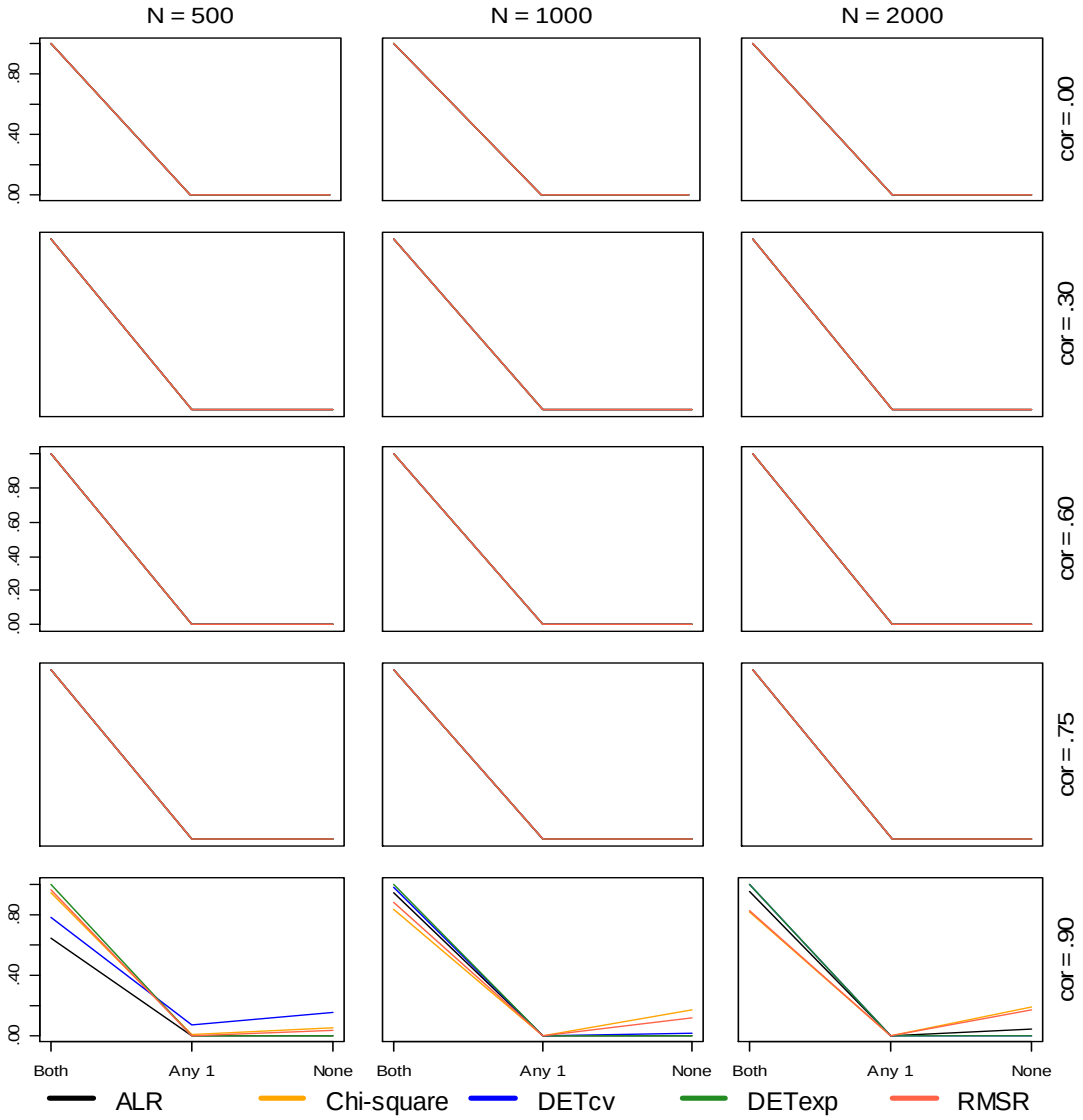

Figure B6. Marginal proportions across 500 replications that a method identified two (both), (any) one, or none of the sets of items as dimension-like (x-axis) when the data exhibit 10% complexity and follow a compensatory 2D MIRT model with 20 items per dimension.

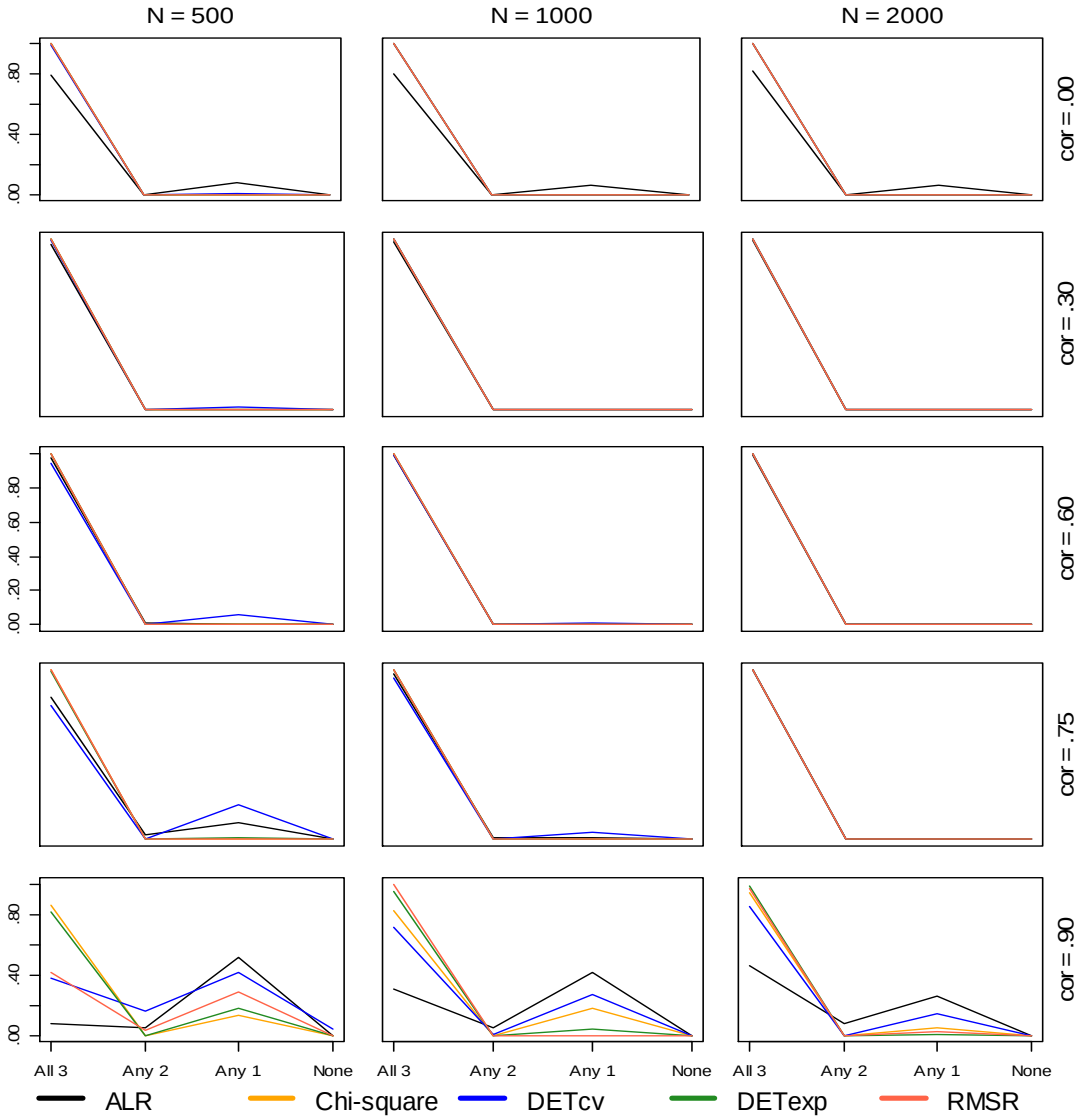

Figure B7. Marginal proportions across 500 replications that a method identified three, any two (both), (any) one, or none of the sets of items as dimension-like (x-axis) when the data exhibit 0% complexity and follow a compensatory 3D MIRT model with 20 items per dimension.

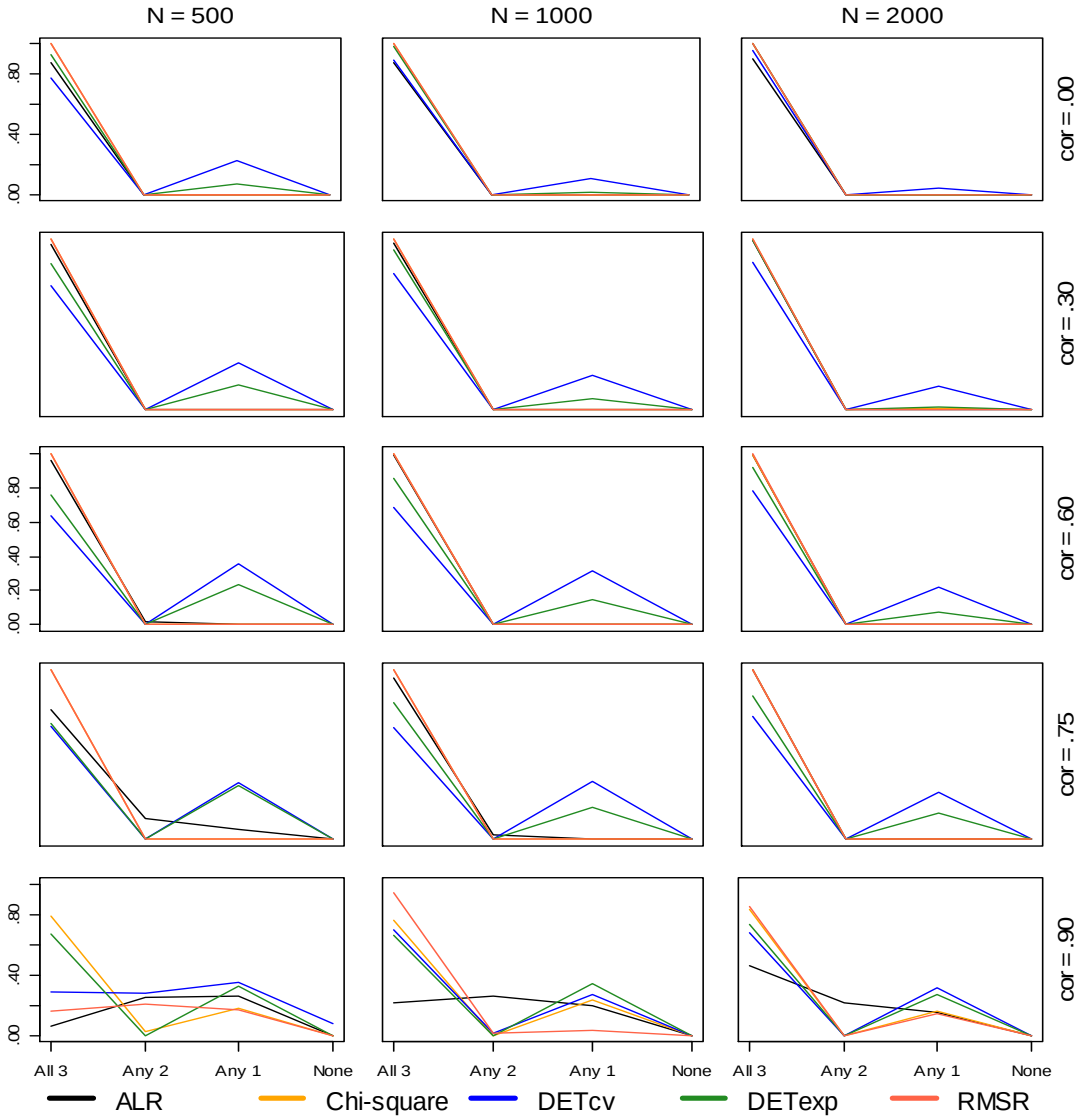

Figure B8. Marginal proportions across 500 replications that a method identified three, any two (both), (any) one, or none of the sets of items as dimension-like (x-axis) when the data exhibit 10% complexity and follow a compensatory 3D MIRT model with 20 items per dimension.

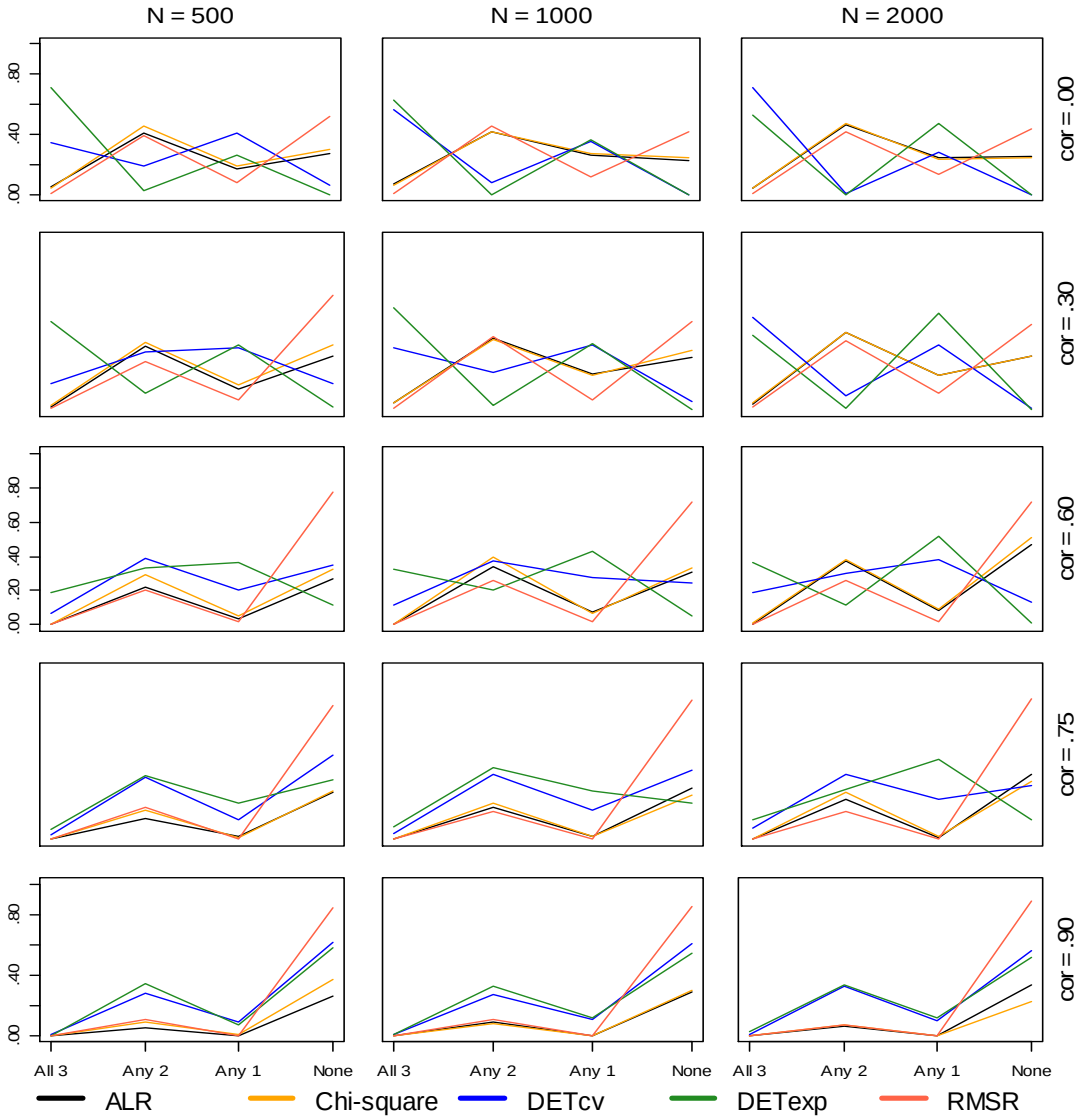

Figure B9. Marginal proportions across 500 replications that a method identified three, any two (both), (any) one, or none of the sets of items as dimension-like (x-axis) when the data exhibit 0% complexity and follow a noncompensatory 3D MIRT model with 10 items per dimension.

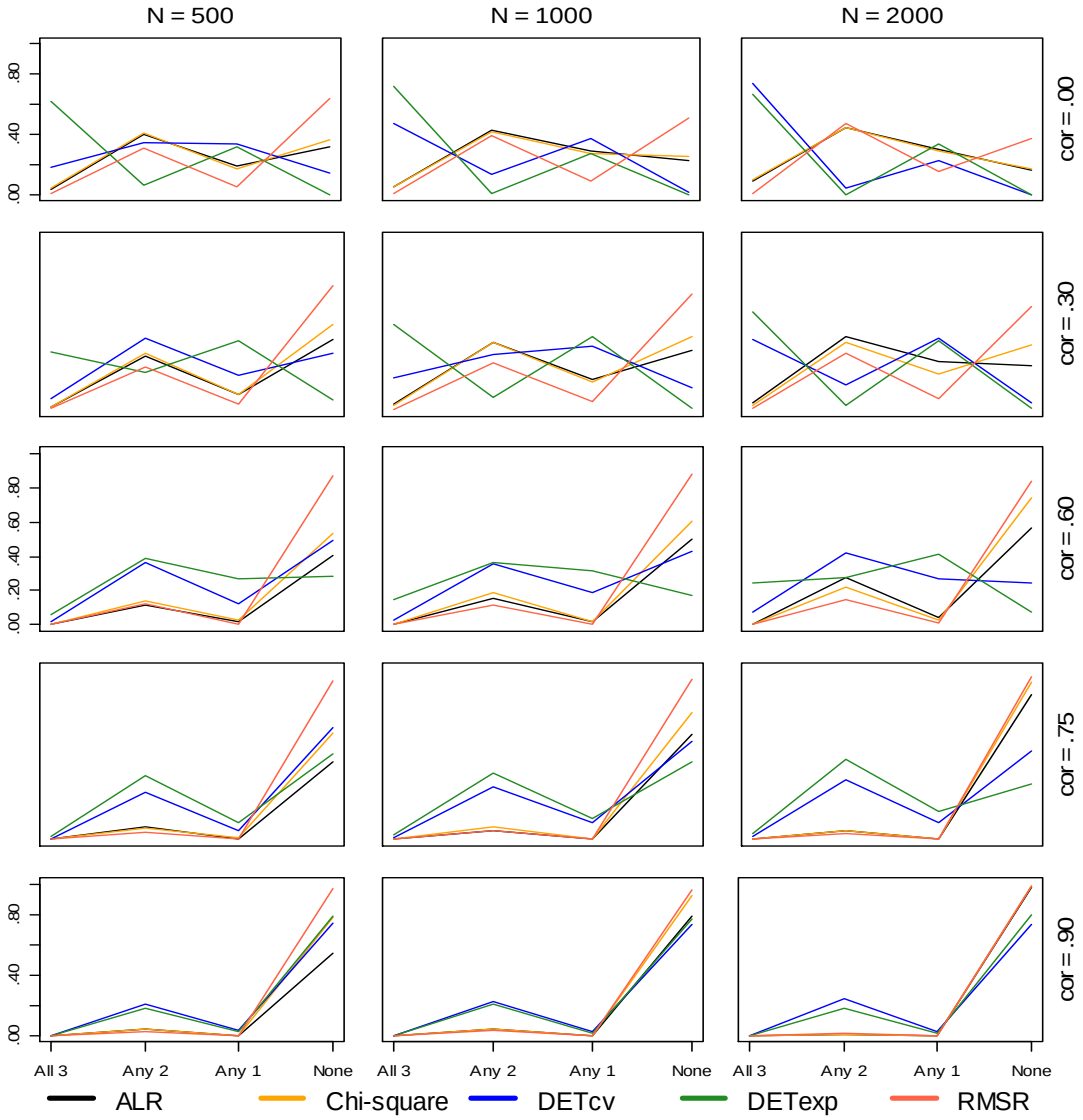

Figure B10. Marginal proportions across 500 replications that a method identified three, any two (both), (any) one, or none of the sets of items as dimension-like (x-axis) when the data exhibit 10% complexity and follow a noncompensatory 3D MIRT model with 10 items per dimension.

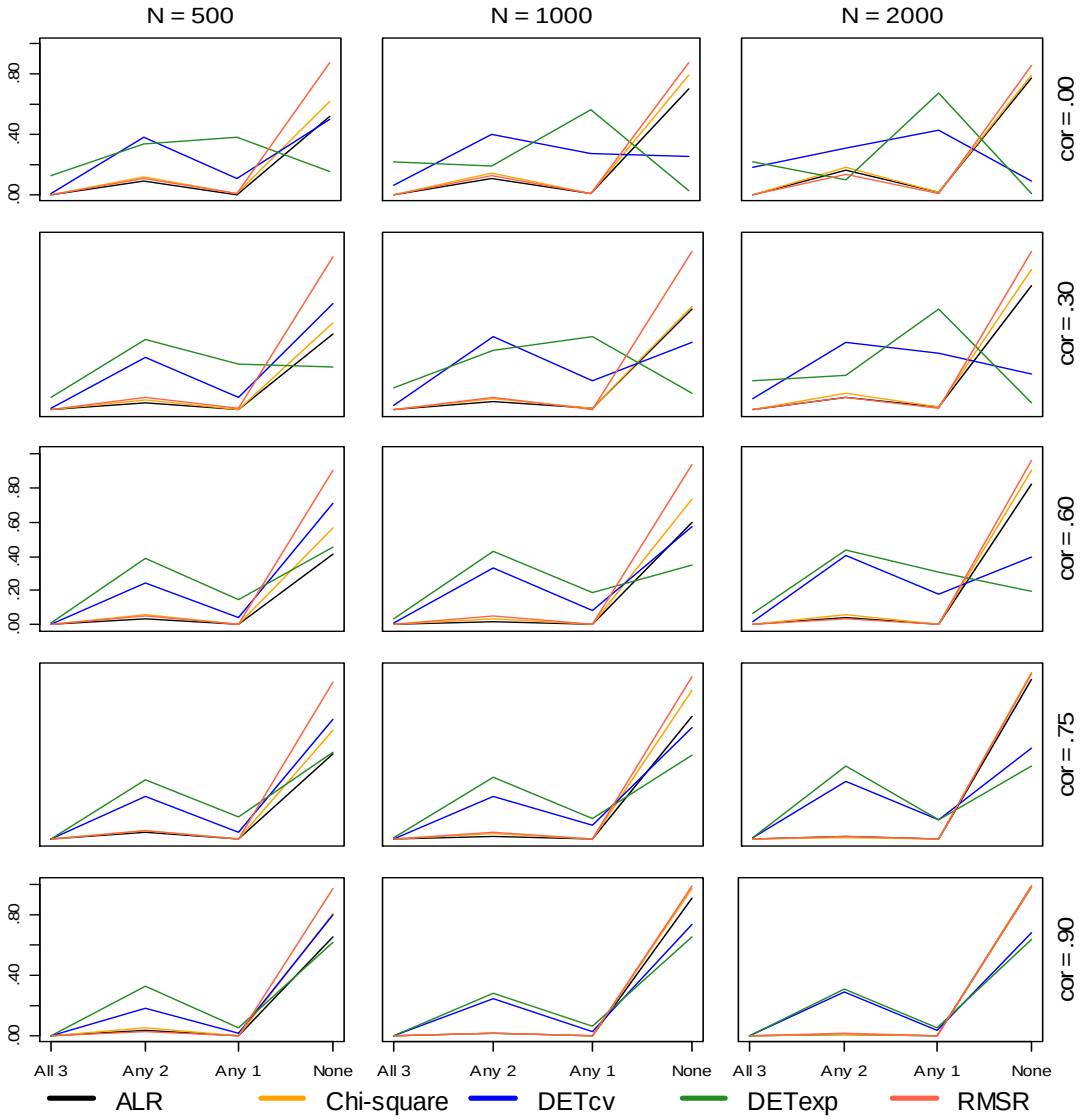

Figure B11. Marginal proportions across 500 replications that a method identified three, any two (both), (any) one, or none of the sets of items as dimension-like (x-axis) when the data exhibit 30% complexity and follow a noncompensatory 3D MIRT model with 10 items per dimension.

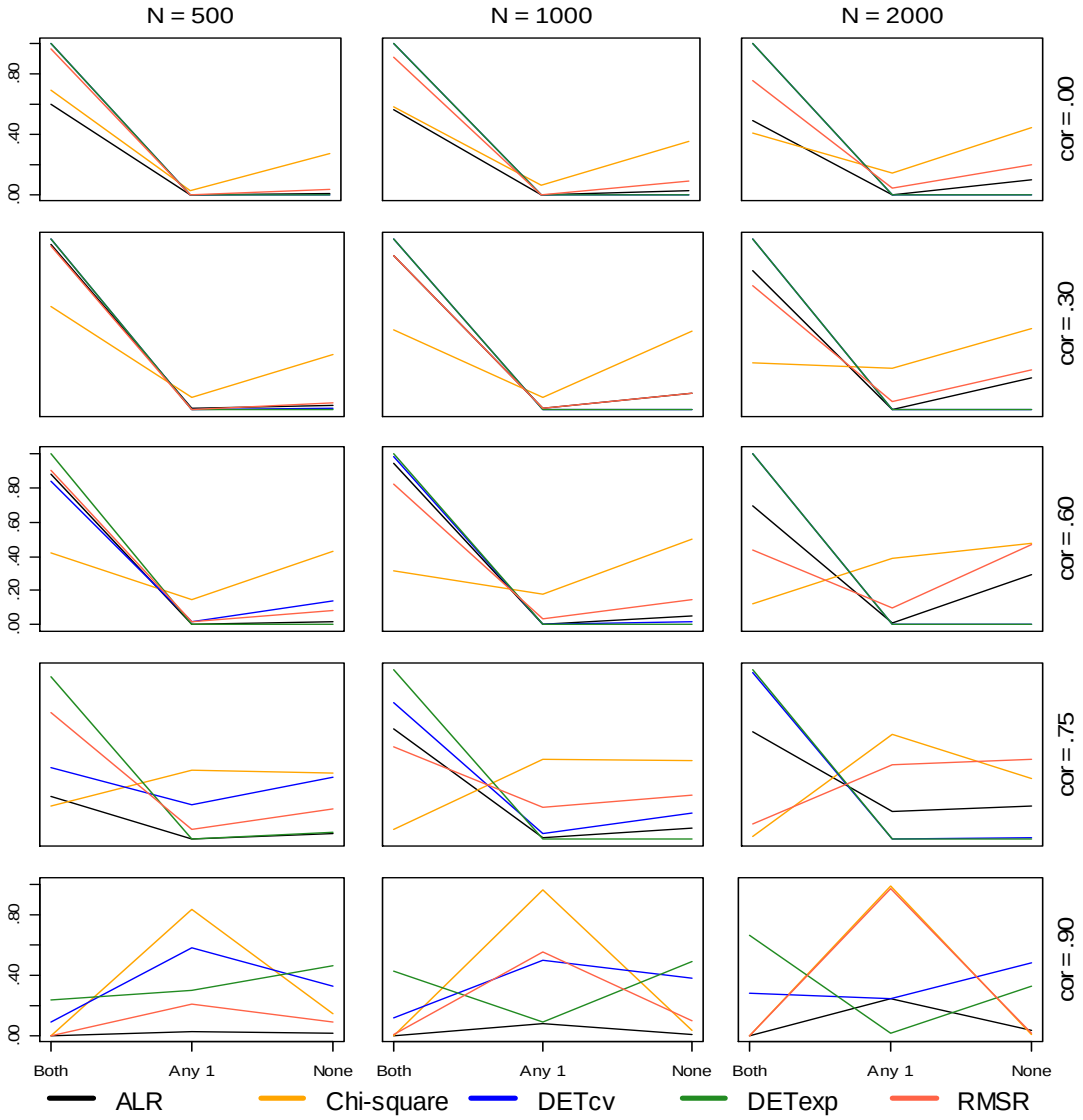

Figure B12. Marginal proportions across 500 replications that a method identified two (both), (any) one, or none of the sets of items as dimension-like (x-axis) when the data exhibit 0% complexity and follow a noncompensatory 2D MIRT model with 20 items per dimension.

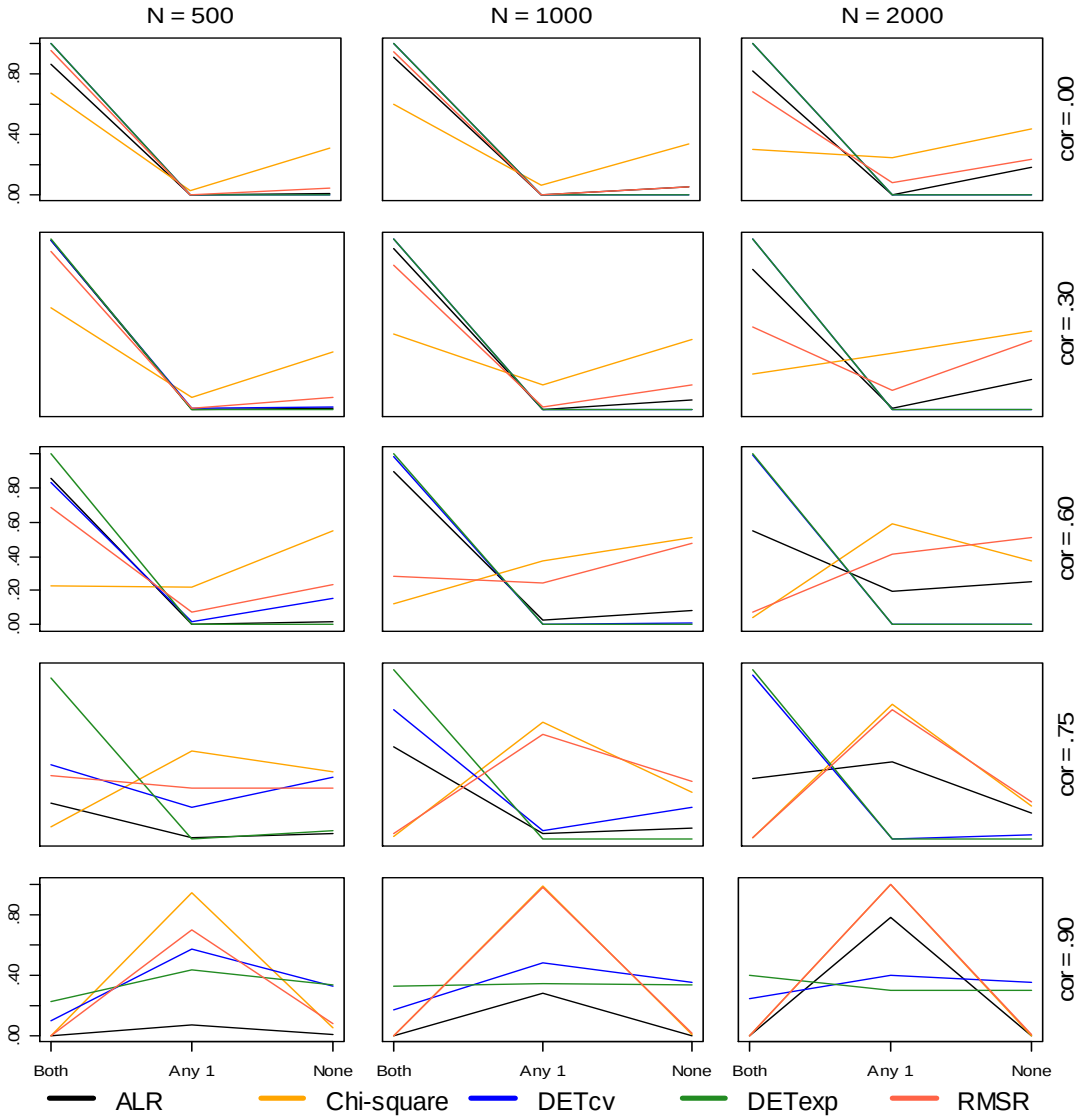

Figure B13. Marginal proportions across 500 replications that a method identified two (both), (any) one, or none of the sets of items as dimension-like (x-axis) when the data exhibit 10% complexity and follow a noncompensatory 2D MIRT model with 20 items per dimension.

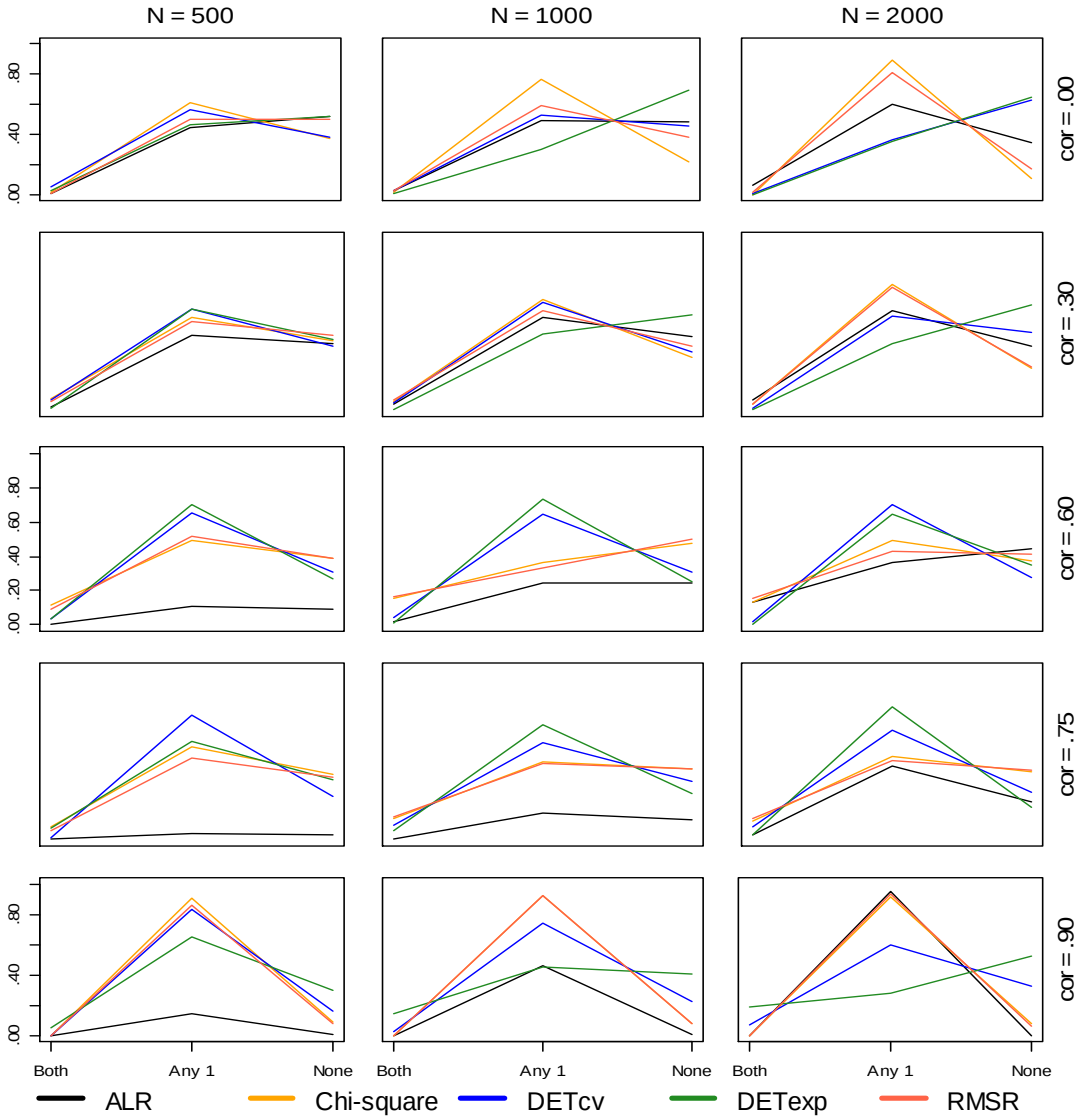

Figure B14. Marginal proportions across 500 replications that a method identified two (both), (any) one, or none of the sets of items as dimension-like (x-axis) when the data exhibit 50% complexity and follow a noncompensatory 2D MIRT model with 20 items per dimension.

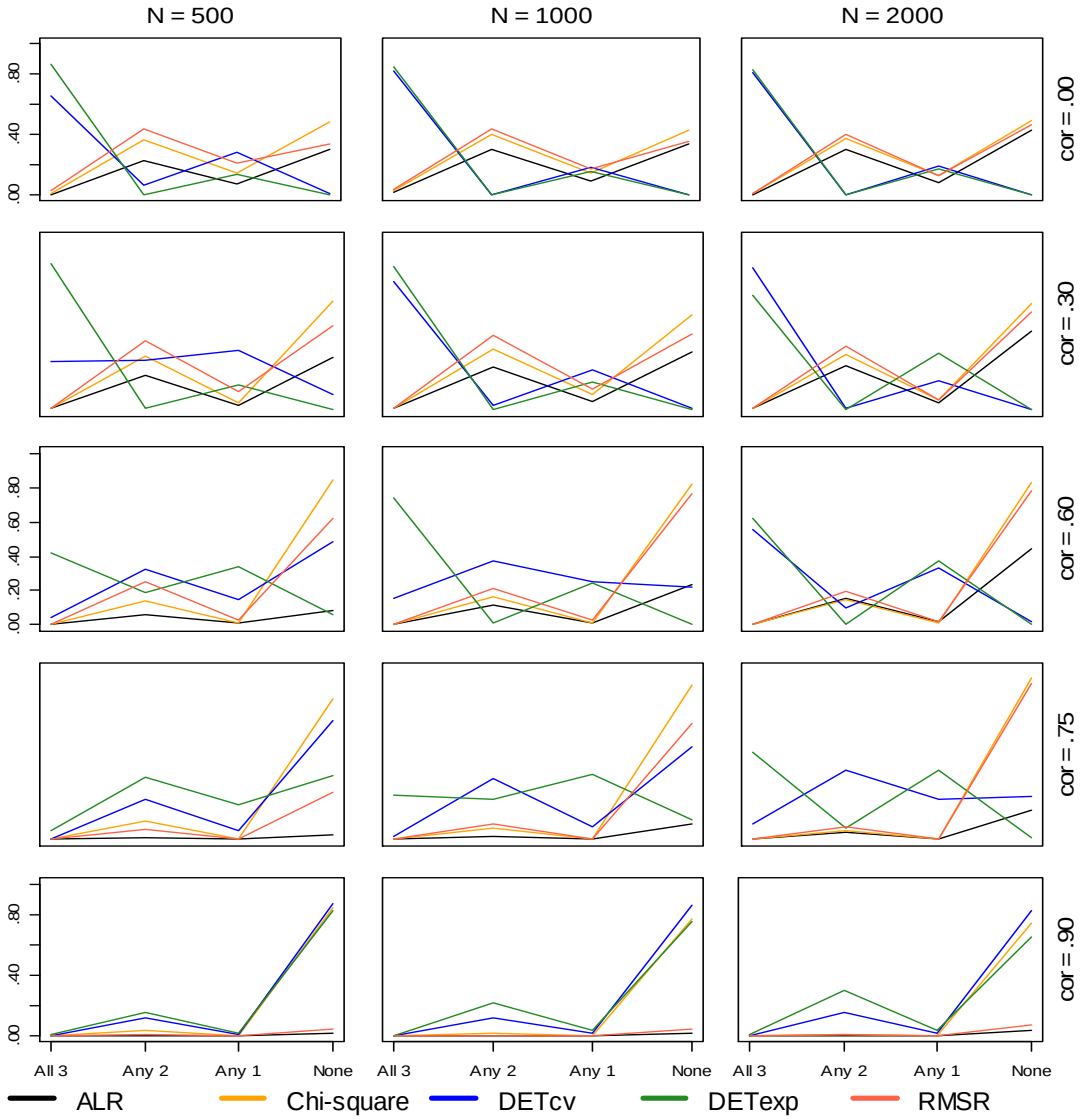

Figure B15. Marginal proportions across 500 replications that a method identified three, any two (both), (any) one, or none of the sets of items as dimension-like (x-axis) when the data exhibit 0% complexity and follow a noncompensatory 3D MIRT model with 20 items per dimension.

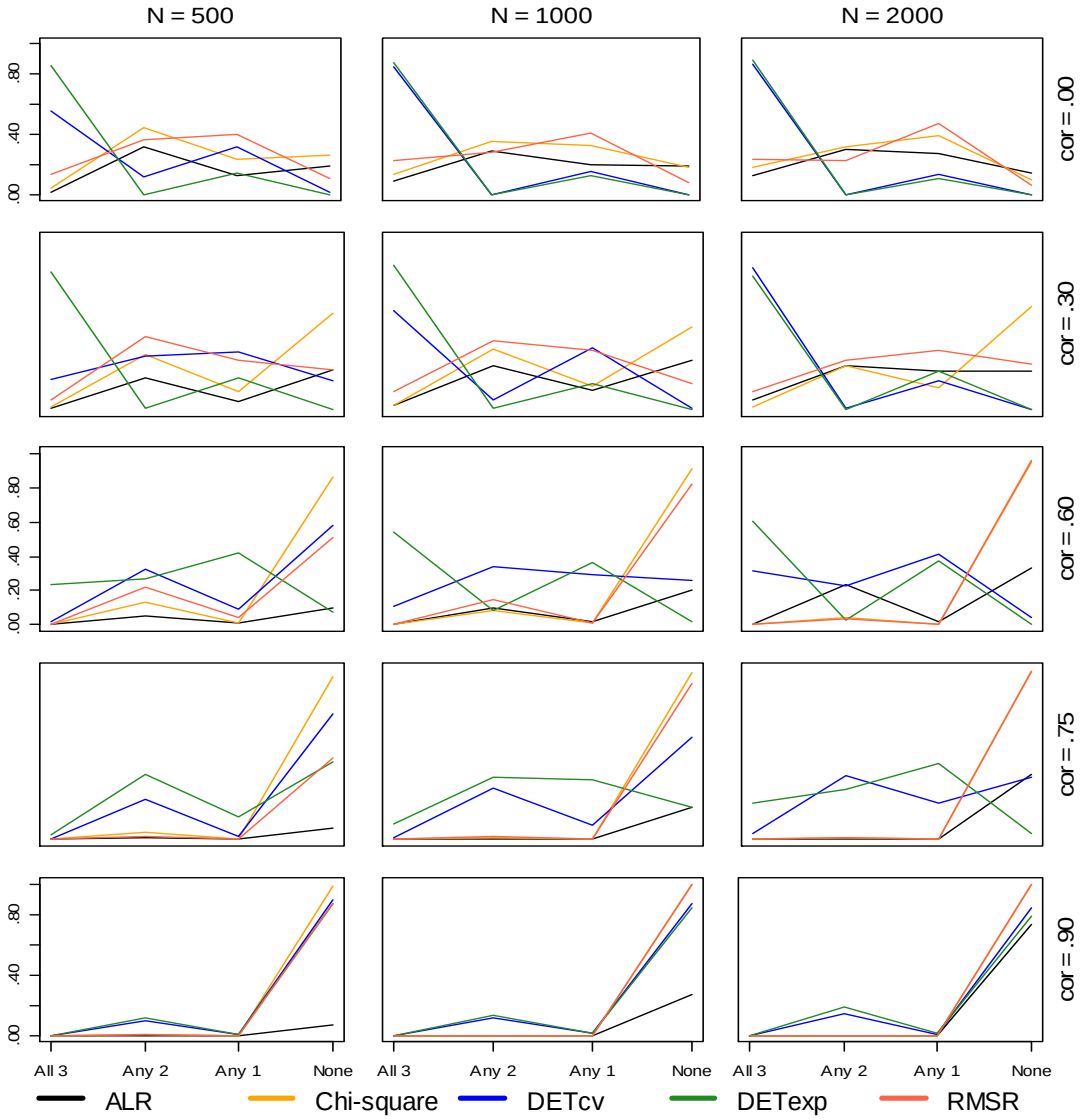

Figure B16. Marginal proportions across 500 replications that a method identified three, any two (both), (any) one, or none of the sets of items as dimension-like (x-axis) when the data exhibit 10% complexity and follow a noncompensatory 3D MIRT model with 20 items per dimension.

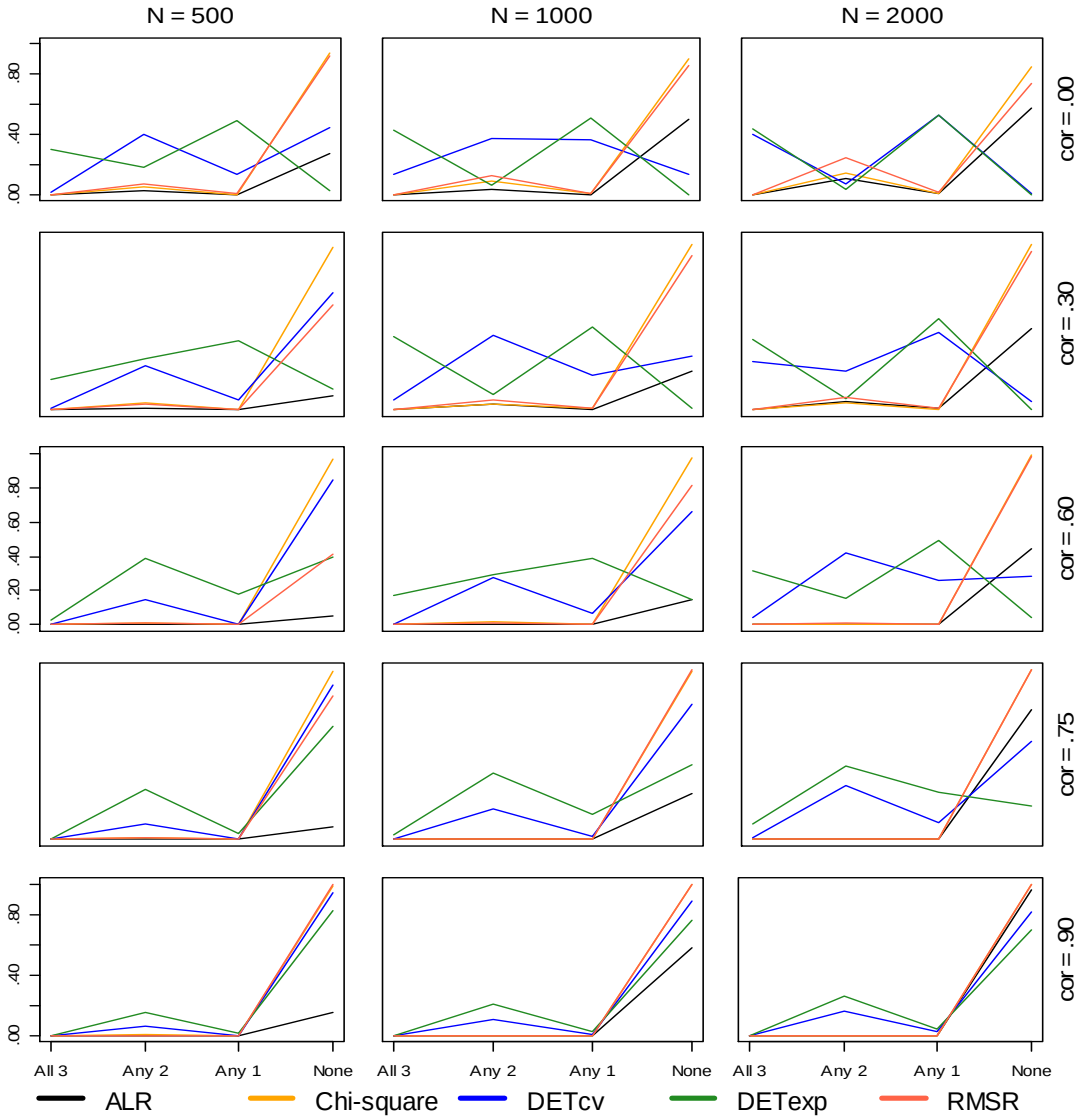

Figure B17. Marginal proportions across 500 replications that a method identified three, any two (both), (any) one, or none of the sets of items as dimension-like (x-axis) when the data exhibit 30% complexity and follow a noncompensatory 3D MIRT model with 20 items per dimension.