Sample Size and Test Length Minima for DIMTEST with Conditional Covariance -Based Subtest Selection

by

Derek Fay

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Arts

Approved April 2012 by the
Graduate Supervisory Committee:

Roy Levy, Chair
Samuel Green
Joanna Gorin

ARIZONA STATE UNIVERSITY

May 2012

ABSTRACT

The existing minima for sample size and test length recommendations for DIMTEST (750 examinees and 25 items) are tied to features of the procedure that are no longer in use. The current version of DIMTEST uses a bootstrapping procedure to remove bias from the test statistic and is packaged with a conditional covariance-based procedure called ATFIND for partitioning test items. Key factors such as sample size, test length, test structure, the correlation between dimensions, and strength of dependence were manipulated in a Monte Carlo study to assess the effectiveness of the current version of DIMTEST with fewer examinees and items. In addition, the DETECT program was also used to partition test items; a second feature of this study also compared the structure of test partitions obtained with ATFIND and DETECT in a number of ways. With some exceptions, the performance of DIMTEST was quite conservative in unidimensional conditions. The performance of DIMTEST in multidimensional conditions depended on each of the manipulated factors, and did suggest that the minima of sample size and test length can be made lower for some conditions. In terms of partitioning test items in unidimensional conditions, DETECT tended to produce longer assessment subtests than ATFIND in turn yielding different test partitions. In multidimensional conditions, test partitions became more similar and were more accurate with increased sample size, for factorially simple data, greater strength of dependence, and a decreased correlation between dimensions. Recommendations for sample size and test length minima are provided along with suggestions for future research.

DEDICATION

This work is dedicated to my wife Jaye whose love and patience has been the cornerstone to the completion of this work. This work is also dedicated to my daughter Ariadne; hopefully you will one day understand the inspiration you offer with your smile.

ACKNOWLEDGMENTS

I would like to extend enormous gratitude to several individuals who have made this work possible. Above all, I would like to thank my thesis advisor, Dr. Roy Levy, for his insightfulness, patience, and guidance. My appreciation goes to my committee members, Dr. Joanna Gorin and Dr. Samuel Green for helping to make this experience a rewarding aspect of my graduate training. Also, I wish to thank my family, both immediate and in-laws, who have provided enormous support over the years, and on countless occasions, lent an open ear and an even more open heart. An additional thank you goes to my brother-in-law, Jordan Williams, for his remarkable knowledge and assistance with batch scripting. Of course, I wish to extend a comprehensive thank you to each of those individuals listed above and those I have failed to mention; while the support of any one individual has made this work possible, it has been the combined support from those in my life that have made this work a reality.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Chapter 1

INTRODUCTION

Item response theory (IRT) methods were originally developed and have traditionally been applied to high-stakes, large-scale testing environments. Such environments can be characterized as having potentially large respondent and item pools. Low-stakes small-scale testing environments tend to have access to fewer respondents items. Despite these tendencies, IRT methods have garnered the interest of those working in these settings. Examples of small-scale testing environments include applied research, screening measures, or even classroom-based tests. Although the respondent and item pools may be more limited in these contexts, those using the tests are often interested in characterizing the factors driving respondents' item responses.

For most applications of IRT, it is assumed that test performance is primarily a function of a single dimension. However, it has been argued that test performance often depends on factors that potentially have a non-trivial effect beyond that of a unidimensional characterization. Non-trivial dimensions, if left unaccounted for, lead to the violation of key independence assumptions and may hinder many facets of testing such as linking and score reporting. Thus, many promising procedures aimed at identifying the presence of unaccounted for dimensions have been developed (see Tate, 2003). Given the high-stakes and large-scale foundations of IRT, most of these dimensionality assessment procedures have been developed and researched assuming respondent and item pools that often exceed those used in small-scale contexts. Thus, small-scale test

developers are left with little guidance with respect to choosing suitable methods for assessing whether a single dimension is sufficient for modeling observed response patterns.

Although many promising dimensionality assessment procedures are available, this work focuses on the performance of the DIMTEST procedure (Froelich & Habing, 2008; Froelich & Stout, 2003; Nandakumar & Stout, 1993; Stout, 1987) for assessing the assumption of unidimensionality in small-scale testing environments. Many extensive Monte Carlo simulation studies have shown DIMTEST to perform favorably for many realistic testing conditions. Though recommendations pertaining to minima for sample size and test length exist for DIMTEST (Gessaroli & De Champlain, 1996; Pyo, 2000), the procedure has undergone many changes that potentially render those recommendations obsolete. One goal of this research is to update the test length and sample size recommendations for the current version of DIMTEST.

A second goal is to evaluate the performance of an alternative method for partitioning test items into two clusters that are as dimensionally distinct as possible. In particular, a specially designed genetic algorithm is used in Zhang and Stout's (1999a) DETECT procedure to identify and characterize maximally distinct dimensions. This method is akin to but more exhaustive than the current method used in exploratory DIMTEST analyses. Though the genetic algorithm has been suggested for partitioning test items into dimensionally distinct clusters (Stout, Froelich, & Gao, 2001), the utility of this method for doing so has not been investigated empirically.

The remainder of this chapter will be aimed at characterizing fundamental concepts pertaining to the current work. Relevant unidimensional and multidimensional IRT models and their corresponding assumptions will be discussed first. Next, Zhang and Stout's (1999a, 1999b) theory of conditional covariances will be briefly characterized prior to presenting DIMTEST and other relevant procedures based on their theory.

Dichotomous IRT Models

Item response models are concerned with modeling a dichotomous outcome as function of a possibly vectored latent, or unobserved, characteristic of any respondent I ($\mathbf{i}_I$) and the set of characteristics of any item ($\boldsymbol{\xi}_j$) j. In the context of educational assessments, a common outcome is $X_{ij} = 1$ and $X_{ij} = 0$ for correct and incorrect responses, respectively. Although many possible dichotomous outcomes exist, IRT models are discussed within the framework of educational assessments for the purposes of this work.

Unidimensional IRT Models. As mentioned earlier, the majority of IRT applications assume examinees' responses solely depend on a single dimension θ. Using representations from Rupp, Templin, and Henson (2010), this situation is depicted graphically in Figure 1. Following normal conventions, the circle represents a latent variable, and the six squares represent observed variables. The lines cutting through the observed variables represent the location of the item (i. e., item difficulty). Arrows emanating from θ to the six observed variables indicate the direction of dependence. For any examinee, the value of observed variables $X_{1,\ldots,}X_6$ depend on their location along θ relative to location of the items

along the same continuum. Barring dependencies among respondents (e. g., through cheating) or influential, but unaccounted for dimensions, an examinee must have more of the characteristic θ than is required by the item in order to have a high probability of responding correctly. From a model-based perspective, a unidimensional IRT model represents a way to accumulate evidence about and summarize student proficiency in terms of a single summary of performance on the tasks.

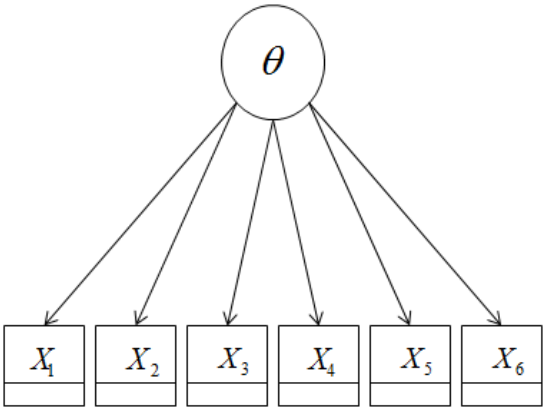


Figure 1. A unidimensional IRT model.

The most general form of the item response function (IRF) commonly found in practice is the 3-parameter logistic (PL) as given by:

$$P\left(X_{ij}=1\,|\,\theta_i,a_j,b_j,c_j\right)=c_j+(1-c_j)\frac{e^{[a_j(\theta_i-b_j)]}}{1+e^{[a(\theta_j-b_j)]}} \qquad (1)$$

where a_j, b_j, and c_j denote the discrimination, difficulty, and lower-asymptote (sometimes referred to as pseudo-guessing) parameters for item j, respectively. Notably, applying constraints on particular item parameters yield more restricted models. Fixing $c_j = 0$ while continuing to allow a_j to vary for each of J items

yields the 2-PL. Further restricting *Equation 1* such that all *a*-parameters are equal yields the 1-PL model.

Multidimensional IRT Models. Multidimensional IRT (MIRT) models extend the unidimensional models above; Figure 2 depicts the general structure of MIRT models. There are two new features in the multidimensional representation that were not necessary in the unidimensional representation. First, the curved arrow indicates the possibility of correlated dimensions. Second, solid arrows signify that performance on an item depends on one dimension exclusively while dashed arrows indicate that performance on an item depends on more than one dimension.

When tests are best characterized as reflecting multiple dimensions, it may be desirable for items to be exclusive indicators of one dimension. Tests consisting of such items are said to be factorially simple. The notion of factorially simple tests may be relaxed to factorially complex tests (Zhang & Stout, 1999a). Complex structure allows for items to be dependent on multiple dimensions. Using the language from the factor analytic framework, items load on multiple dimensions. When simple structure does not hold, it may be that approximate simple structure holds, where items have strong loadings on one dimension and trivial, but non-zero, loadings on other dimensions. Having said that, since the criteria for what counts as a trivial loading can be fairly subjective, the more general term “complex structure” will be used throughout this work to simply refer to data structures that do not follow a simple structure model.

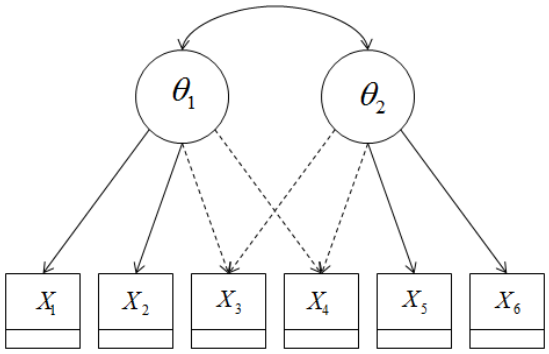


Figure 2. General representation of MIRT models.

Though various forms of MIRT exist, compensatory models are the most widely used. The 3-PL compensatory MIRT model is formally given by:

$$P(X_{ij}=1 \mid a_j, d_j, c_j, \boldsymbol{\theta}_i) = c_j + (1-c_j)\frac{e^{(\sum_{m=1}^{M} a_{jM}\theta_{iM}+d_j)}}{1+e^{(\sum_{m=1}^{M} a_{jM}\theta_{iM}+d_j)}} \qquad (2)$$

where $\boldsymbol{a_j} = (a_{j1},\ldots, a_{jM})$ denotes the vector of discrimination parameters; $\boldsymbol{\theta}_i = (\theta_{i1},\ldots, \theta_{iM})$ denotes the vector of M examinee characteristics; c_j denotes the lower-asymptote (pseudo-guessing) parameter; and d_j denotes a scalar related to the item difficulty (Reckase, 1985, 1997). Like their unidimensional counterparts, multidimensional models are hierarchically related. Fixing $c_j = 0$ yields the 2-PL MIRT model, and fixing all elements in $\boldsymbol{a_j}$ as equal yields a1-PL MIRT model.

Conditional Independence in IRT Models

Local, or conditional, independence (LI) is a central assumption of item response models. To satisfy the LI assumption, responses to test items must be statistically independent given examinees possibly vectored set of attributes $\boldsymbol{\theta}$ and the characteristics of the items, $\boldsymbol{\xi}_j$. The LI assumption is formally given by:

$$P(X_{i1},...,X_{iJ} \mid \boldsymbol{\theta}_i, \boldsymbol{\xi}_j) = \prod_{j=1}^{J} P(X_{ij} \mid \boldsymbol{\theta}_i, \boldsymbol{\xi}_j). \quad (3)$$

Local dependence occurs (LD) when LI is violated. As a short list, possible sources of LD include other examinee characteristics, speededness, fatigue, and passage dependence. More generally, finding statistical evidence of LD implies that some dimension beyond that conditioned on effects how some, or all, examinees respond to some, or all, items (Yen, 1993). Additional dimensionality includes both those that are cognitively meaningful and those that constitute as "nuisance" dimensions. Nuisance dimensions might involve, for example, unintended consequences of test design (e. g., speededness) and/or peculiarities of the item(s) (Ip, 2000; Stout, 1987).

Investigating LD is essential prior to accepting an IRT model as a tool for characterizing student proficiency. The presence of unaccounted for LD may result in imprecise estimates of person and item parameters (Ackerman, 1987; Yen, 1993) which can in turn threaten estimates of information and standard errors of estimates, test equating and linking, the reporting of scores and precision of scores, and ultimately, the interpretation and use of scores (Birnbaum, 1968; van der Linden, 1996; Yen, 1993). Given the pervasive influence of LD, identifying substantial LD is of the utmost importance prior to further measurement activities.

Weaker Forms of LI. *Equation 3* is often referred to as strong local independence (SLI). Unlike weaker forms of LI, satisfying SLI requires that all bivariate and higher-order dependencies are accounted for by $\boldsymbol{\theta}_i$ and $\boldsymbol{\xi}_j$; weaker

forms of LI are concerned with bivariate associations. Often cited forms that are weaker than SLI include weak local independence (WLI; McDonald, 1994) and essential independence (EI; Stout, 1987).

Weak local independence, also called pair-wise independence, is often investigated in practice in place of SLI. A set of items is considered pairwise independent if:

$$\text{cov}(X_{ij}, X_{ij'} \mid \boldsymbol{\theta}_i, \xi_j) = 0, \text{ for all } \boldsymbol{\theta} \text{ and } 1 \leq \text{j} < j' \leq \text{J} \qquad (4)$$

where *cov* denotes covariance. Since WLI does not account for higher-order dependencies, it is not mathematically sufficient for SLI (Zhang & Stout, 1999b). However, WLI is thought to be *empirically* sufficient for SLI when working with real datasets; although higher-order dependencies are possible, they are unlikely when WLI holds (McDonald, 1994). Central to the procedure of interest for this research—DIMTEST—is the weaker notion of EI (Stout, 1987).

Two key differences separate EI from the stronger SLI and WLI forms. First, the average conditional covariances should be small and become smaller as *J* approaches infinity. Test items are viewed as EI if:

$$\frac{\sum_{1 \leq j < j' \leq J} |\operatorname{cov}(X_{ij}, X_{ij'} \mid \boldsymbol{\theta} = \boldsymbol{\theta}')|}{\binom{J}{2}} \rightarrow 0 \qquad (5)$$

for all $\boldsymbol{\theta}'$ as $J \rightarrow \infty$. Additionally, while SLI and WLI involve conditioning on dominant and minor (i. e., all) dimensions, only *dominant* dimensions (denoted $\boldsymbol{\theta}'$ above) are of interest for EI (Stout, 1987, 1990). This viewpoint is akin to the factor analytic tradition of dimensionality assessment; dimensions characterized

by a single or relatively few items tend to be psychometrically uninteresting while dimensions characterized by many items are of greater substantive interest (Stout, 1987). The minimum number of dominant dimensions necessary to produce an EI model is termed the essential dimensionality; essential unidimensionality is achieved when a single dominant dimension (θ') is sufficient (Nandakumar & Yu, 1996).

Relationship between LI and Dimensionality. The presentation of LI has thus far aimed to establish the link between LI and the number of underlying dimensions. However, the relationship between LI and the underlying dimensionality is more general in that finding evidence of LD is akin to finding evidence that the dimensionality is underspecified (Ip, 2001). Clearly, a model is underspecified if there are too few dimensions in the model to account for the covariance among observables. The dimensionality of a model can also be underspecified even if the "correct" number of dimensions is included. For example, Levy and Svetina (2011) showed that fitting a factorially simple model to data that were factorially complex resulted in violations of LI holding constant the number of dimensions. Although the correct number of dimensions was specified, unmodeled dependencies between items and the underlying dimensions resulted in violations of LI.

Conditional Covariance-Based Dimensionality Assessment

Zhang and Stout's conditional covariance theory (CCT; 1999a, 1999b) is the common foundation for the three procedures relevant to the current work: DIMTEST, DETECT, and CCPROX/HCA. The non-parametric CCT framework

was developed as an alternative to parametric approaches to dimensionality assessment. Parametric approaches to dimensionality assessment make additional assumptions beyond those required for assessing dimensionality via CCT. Dimensionality assessment via CCT only requires monotonicity ($P(X_{ij} = 1 \mid \boldsymbol{\theta}_i \rightarrow 1$ as $\boldsymbol{\theta}_i \rightarrow \infty$); assumptions about the distribution of $\boldsymbol{\theta}_i$ or the particulars of the IRFs (e. g., guessing) are not made. The form of the IRF for CCT is Zhang's (1996) generalized *m*-dimensional compensatory model as given by:

$$P(X_{ij} = 1 \mid \boldsymbol{\theta}_i) = H_j\left(\sum_{m=1}^{M} a_{jm}\theta_{\mathrm{m}} - b_j\right) \tag{6}$$

where H_j is any non-decreasing function (i. e., monotonicity); the a_j and b_j terms retain their usual meaning. This form of the IRF represents a more general form of the common logistic and other (e. g., normal ogive) compensatory MIRT models. In particular, whereas the common versions of compensatory models assume a specific form of the IRF, H_j in the generalized compensatory model is arbitrary.

The central features of CCT-based dimensionality assessment include the item, unidimensional test composites for each dimension in the test space, and total test composite as measured by the total test score (Stout et al., 1996; Zhang & Stout, 1999a). Items cluster together to yield the weighted unidimensional composites corresponding to each dimension in the test space, and the total test composite represents the unit-weighted aggregate of all unidimensional composites in the test space. Each of these features of a test can be represented geometrically by vectors as shown in Figure 3, which shows a two dimensional

test. The total test composite is denoted by θ_{TT}; unidimensional composites corresponding to the two dimensions are denoted by θ_{C1} and θ_{C2}; and the items associated with each dimension are shown as vectors clustered around θ_{C1} and θ_{C2}.

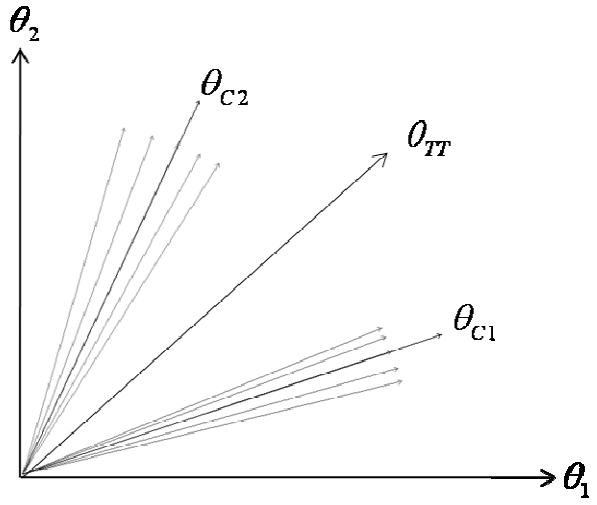


Figure 3. Geometric representation of a two-dimensional test.

Two key features of the vectors reveal the dimensional structure of the test: the angle relative to the θ_1 axis and length. The angle relative to the θ_1 axis for all vectors represents the direction of maximum discrimination. Ideally, the direction of all θ_{C1} items would be aligned with the θ_1 axis, and all θ_{C2} items would be aligned with the θ_2 axis; this situation would result in a model with simple structure and orthogonal dimensions. Figure 3 depicts a test structure that is non-ideal but more likely to occur with real test data. As shown by the $\theta_{C1}$and θ_{C2} vectors deviations from the θ_1 and θ_2 axes, the dimensions are correlated; that is, the dimensions are oblique to each other rather than orthogonal. Importantly, even with oblique dimensions, a simple structure model would obtain if all items were identically aligned with the respective unidimensional composite vectors

(Zhang & Stout, 1999a). Since item vectors are spread around the respective unidimensional composites, however, the test shown in Figure 3 follows a model with complex structure. That is, all items measure one dimension best and have smaller loadings on the other dimension.

From the CCT perspective, the directionality of item vectors has implications for quantifying the extent of multidimensionality. Taken relative to the direction of the θ_{TT} vector, any two items with directions on the same side of θ_{TT} will exhibit a positive conditional covariance; any two items with directions on opposite sides θ_{TT} will exhibit a negative conditional covariance; and if either item from a pair is aligned with θ_{TT}, the conditional covariance will be zero (Zhang & Stout, 1999a). Referring back to Figure 3, all items on the same side of θ_{TT} will be positively related. Therefore, the covariance among all pairs of θ_{C1} items will be positive conditional on θ_{TT} as will the covariance among all pairs of θ_{C2} items. Any pair of items taken from opposite clusters, however, will result have a negative covariance conditional on θ_{TT}.

The CCT perspective provides a strong theoretical framework for DIMTEST, DETECT, and CCPROX/HCA. Although DIMTEST is the primary interest of this work, the other CCT-based procedures are used for conducting exploratory runs of DIMTEST. In particular, each of the methods for partitioning items into maximally distinct clusters employs a combination of CCPROX/HCA and DETECT. With that, the remainder of this chapter briefly characterizes each of these procedures.

DIMTEST

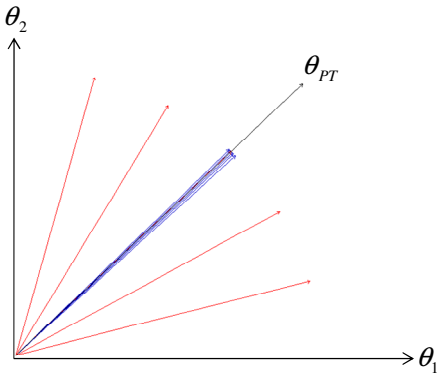
As mentioned, it is hypothesized for most applications of IRT that a single dimension is sufficient to satisfy the LI assumption. The DIMTEST procedure subjects the weaker hypothesis of essential independence to a formal test by splitting items into two clusters and evaluating their statistical distinctiveness (Froelich & Habing, 2008; Stout et al, 2001; Nandakumar & Stout, 1993; Stout, 1987). One cluster, called the assessment subtest (AT), ideally consists of items that are dimensionally homogenous to each other but dimensionally distinct from the remaining test items. The second cluster, called the partitioning subtest (PT), consists of the remaining items and is used to form ability subgroups as measured by the total score across the items that comprise the PT. The splitting of test items into the AT and PT can be done using either confirmatory or exploratory methods. This work focuses on exploratory methods; exploratory strategies for obtaining the AT/PT partition are reviewed after presenting the DIMTEST procedure.

In essence, the DIMTEST statistic is the bias-corrected standardized difference between the total variability and variability due to a single dimension (herein referred to as unidimensional variance). Equation 1.10 in Froelich and Stout (2003) indicates that the sum of the difference between the two variance estimates across all ability groups results in the total estimated covariance among test items. If the total variance is equal to the unidimensional variance, then a single dominant dimension accounts for all of the variability in the observed data. The conclusion would be that the AT and PT items measure the same single dimension. Small deviations between the two variance estimates suggest that

minor dimensions may be present, but a single dominant dimension is sufficient to satisfy the LI assumption; that is, the data are essentially unidimensional. Large deviations provide evidence that the AT and PT subtests measure two distinct dimensions. This situation would result in the rejection of the null hypothesis that the assumption of EI is satisfied given a single dimension.

While the description above captures the essence of the DIMTEST procedure, the following steps briefly show the formalized translation. The first step, which is also the most crucial, requires that the test items be split into the AT and PT subtests. Figures 4a and 4b depict poor and good partitions, respectively, for testing the dimensional distinctiveness of ten items with the DIMTEST procedure. Blue and red item vectors correspond to items in the θ_{PT} and θ_{AT}, respectively. The length of each item vector represents the composite magnitude of the item discrimination, and the angle of the item vector from theθ_1 axis represents the composite item direction.

(a) Poor Partition



(b) Good Partition

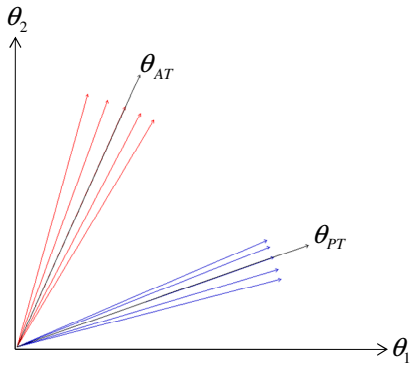


Figure 4. Possible partitions of items into the AT and PT subtests.

The DIMTEST procedure is more likely to fail to reject the null hypothesis of essential unidimensionality for the partition in Figure 4a. The AT is clearly not comprised of a set of dimensionally homogenous items; this is seen by the spread of the AT items throughout the test space. That is, the AT items are heterogeneous in that some items measure θ_1 best while others measure θ_2 best. In addition, the direction of θ_{AT} is not differentiable from that of θ_{PT}. Put differently, the AT items do not measure a dimension that is distinct from the PT items. Figure 4b shows a partitioning of AT and PT that would likely result in DIMTEST rejecting the null hypothesis of essential unidimensionality. The AT items are localized within the test space, indicating that the items homogenously measure one dimension. Furthermore, that the angle between the θ_{AT} and θ_{PT} unidimensional composites is wide indicates the measurement of two distinct dimensions.

Calculating the DIMTEST statistic. After obtaining an AT/PT partition and forming *K* ability subgroups based on total PT scores, the DIMTEST statistic can be calculated. The total score on the AT subset is given by:

$$Y_{ik} = \sum_{j \in AT}^{J} X_{ijk} \tag{7}$$

where X_{ijk} denotes the response from examinee *i* from subgroup *k* for item *j*. The average total score for examinees in subgroup *k* is calculated as:

$$\overline{Y}_k = \frac{1}{I_k} \sum_{i=1}^{I_k} Y_{ik} \,, \tag{8}$$

with I_k denoting the total number of examinees in subgroup k. The usual variance estimate of examinee total scores for the kth subgroup is:

$$\hat{\sigma}_k^2 = \frac{1}{I_k}\sum_{i=1}^{I_k}(Y_{ik} - \overline{Y}_{ik})^2 . \quad (9)$$

The estimate of the unidimensional variance for the kth subgroup is given by:

$$\hat{\sigma}_{U,k}^2 = \sum_{j=1}^{J}\hat{p}_{jk}\left(1-\hat{p}_{jk}\right), \quad (10)$$

where $\hat{p}_{jk}$ is the estimated difficulty of item j for subgroup k; this value is calculated as:

$$p = \frac{1}{I_k}\sum_{i=1}^{I_k} X_{ijk} .$$

The difference between the usual and unidimensional variance estimates for each subgroup yields the estimate of the conditional covariance among all item pairs for that subgroup. To assess the null hypothesis of essential unidimensionality, the total estimated conditional covariance across all subgroups is aggregated and standardized. The variance of the estimated conditional covariance for the kth subgroup is given by:

$$S_k^2 = \frac{(\hat{\mu}_{4,k} - \hat{\sigma}_k^4) - \delta_{4,k}}{J_k}, \quad (11)$$

such that

$$\mu_{4,k} = \frac{\sum_{i=1}^{I_k}(Y_{ik} - \overline{Y}_k)^4}{I_k}$$

and

$$\delta_{4,k} = \sum_{j=1}^{J} \hat{p}_{jk}(1-\hat{p}_{jk})(1-2\hat{p}_{jk})^2$$

Aggregating the estimated conditional covariance and its corresponding variance across all K subgroups yields the DIMTEST statistic,

$$T_L = \frac{\sum_{k=1}^{K} T_{L,k}}{\sqrt{\sum_{k=1}^{K} S_k^2}}. \quad (12)$$

The T_L statistic tends to be positively biased for short tests for both unidimensional and multidimensional tests (Stout, 1987; Stout et al., 2001). Keeping in mind that DIMTEST is designed to assess the hypothesis of essential unidimensionality, the positive bias can inflate the Type I error rate. That is, the DIMTEST statistic may suggest the presence of additional dominant dimensions when the test is actually essentially unidimensional. The original formulation of DIMTEST corrected this bias with a second assessment subtest (AT2). For each item on AT1, an item of approximately equal difficulty from the remaining pool of test items was chosen to be on AT2. Using the AT2 items, a second DIMTEST statistic (T_B) was calculated using the above formulas. The bias-corrected DIMTEST statistic is calculated as:

$$T = \frac{T_L - T_B}{\sqrt{2}}. \quad (13)$$

The T statistic was proven by Stout (1987) to follow a standard normal distribution under the null hypothesis of unidimensionality as the number of examinees and test items tend to infinity.

As stated in Stout et al. (1996), the null hypothesis of essential unidimensionality ($d_E = 1$) and alternative hypothesis ($d_E > 1$) tested by DIMTEST are:

H_0: $AT \cup PT$ satisfies essential unidimesionality ($d_E = 1$) and H_A: $AT \cup PT$ fails to satisfy $d_E = 1$, respectively. That is, the null hypothesis is that the AT and PT both measure the same underlying dominant dimension as measured by total score; the alternative is that the items on the AT measure a dimension other than that measured by the PT. Formally, the decision reached via the DIMTEST statistic T regarding the dimensionality of a test is:

$$= \begin{cases} \text{Fail to Reject, if } T < Z_{100(1-\alpha)} \\ \text{Reject, if } T \geq Z_{100(1-\alpha)} \end{cases}.$$

That is, the T statistic rendered from the DIMTEST procedure is used to decide whether to reject or fail to reject the null hypothesis of essentially unidimensionality ($d_E = 1$).

The current DIMTEST procedure no longer uses AT2 to remove the positive bias from the T_L statistic. The AT2 tended to not remove enough of the bias when the AT1 items were homogenous in terms of difficulty and/or had large discriminations (Nandakumar & Stout, 1993; Stout et al., 2001). In addition, since the power of DIMTEST increases as the test length increases, selecting items purely for the purpose of bias removal seems wasteful. The current DIMTEST procedure removes bias via a bootstrapped DIMTEST statistic. The bootstrapping procedure begin with estimating unidimensional IRFs based on the observed data; based on the estimated IRFs, unidimensional data sets are generated; and a

DIMTEST statistic is calculated for each generated data set. Then, the average DIMTEST statistic across the simulated data sets, $\overline{T_G}$, is used to remove bias from the T_L statistic; the new bias-corrected DIMTEST statistic, as shown in Stout et al. (2001), is given by:

$$T = \frac{T_L - \overline{T}_G}{\sqrt{1 + 1/N}} \tag{14}$$

This version of DIMTEST, which is used in this study, has been shown to have greater power while maintaining control of the Type I error rate than earlier versions using the FAC and AT2 (Finch & Habing, 2007; Froelich & Habing, 2008; Stout et al., 2001).

Strategies for Partitioning Test Items

The DIMTEST procedure can either be conducted in a confirmatory or exploratory manner. Users with substantively-based hypotheses can conduct DIMTEST as a confirmatory analysis by manually selecting items for the PT and AT. Users can also conduct DIMTEST as an exploratory analysis by obtaining the test partition using statistical methods. The use of conditional covariance-based exploratory methods is one focus for this work. Although DIMTEST can be a powerful confirmatory dimensionality analysis tool, the typical user will often use DIMTEST in an exploratory fashion. Notably, users with substantive hypotheses can also benefit from exploratory analyses in two key ways. First, an exploratory approach may yield a test partition that is in agreement with substantive hypotheses; this finding would provide empirical evidence for the user's intuition about the underlying dimensionality of the test. Second, an exploratory approach

can also suggest alternative test partitions that may yield additional insights into underlying test dimensionality. While any approach thought to be suitable for partitioning into two unidimensional item sets (Stout, 1987), the exploratory approaches used in this work are based on the same conditional-covariance based framework from which DIMTEST was developed. In particular, a conditional covariance-based hierarchical clustering procedure and Zhang and Stout's (1999b) DETECT index form the basis of the both of the exploratory methods investigated. Accordingly, key features of these two procedures are discussed prior to presenting the two conditional-covariance based exploratory test partitioning methods.

CCPROX/HCA. For the exploratory methods considered here, a hierarchical clustering procedure as described by Roussos, Stout, and Marden (1998) serves as the basis for finding candidate test partitions. The clustering procedures used in this work consist of two key steps. First, an item-level proximity matrix is obtained based on a conditional covariance-based measure of proximity (CCPROX) as given by:

$$p_{\text{ccov}} = \frac{1}{N_k}\sum_{k=0}^{J-2} N_k \operatorname{cov}(X_j, X_{j'} \mid S_{j,j'}) \text{ + constant,} \qquad (15)$$

where $S_{j,j'}$ denotes examinees total score excluding items X_j and $X_{j'}$, and N_k denotes the number of examinees with $S_{j,j'}$ (Stout et al., 1996). The second step is to conduct the agglomerative hierarchical cluster analysis (HCA). At the beginning of the HCA, all items are one *object* in *J clusters*; at the end of the HCA, all items comprise a single *J* object cluster. For all iterations in between the

beginning and end points, items are combined into multiple object clusters based on (a) the item-level proximity matrix and (b) the linking method used to calculate the proximity between clusters.

Roussos et al. (1998) described and investigated four linking methods, three of which are pertinent to the subtest selection methods used in this work. The three relevant linking methods for calculating the proximity between clusters include single link (Florek, Lukaszewicz, Perkal, Steinhaus, & Zubrzycki, 1951), complete link (McQuitty, 1960), and unweighted pair-group method of averages (UPGMA; Sokal & Michener, 1958). The single link algorithm calculates the minimum value of proximity for all pairwise item proximities among all objects between clusters; in contrast, the complete link algorithm calculates the maximum value of proximity for all pairwise item proximities among all objects between clusters; and the UPGMA calculates the average pairwise proximity among all objects between clusters. Irrespective of the chosen linking method, clusters yielding the smallest value of the proximity measure are combined together with the other clusters left unaffected.

The CCPROX/HCA procedure is built into the two algorithms for partitioning test items (which are described below). Within these algorithms, the CCPROX/HCA procedure serves as a basis for finding candidate partitions that best approximate simple structure, the situation that yields the greatest chance for DIMTEST to correctly find true departures from essential unidimensionality. However, CCPROX/HCA does not provide the two cluster AT/PT partition needed for DIMTEST. More generally, when and if the true underlying structure

of the test is found via CCPROX/HCA, the program provides no indication that this has occurred. For both subtest selection methods, candidate partitions are supplied to the DETECT program (discussed next) to search for the test partition that indeed maximizes the degree of multidimensionality underlying a test.

DETECT. For a given partition of items into clusters, denoted *P*, Zhang and Stout's (1999b) theoretical DETECT index quantifies the extent to which observed data exhibit multidimensionality. The theoretical DETECT index $D(P, \theta_{TT})$ is formally given by:

$$D(P,\theta) = \frac{2}{J(J-1)} \sum_{1 \le j < j' \le J} \delta_{j,j'} E\left[\mathrm{cov}(X_j, X_{j'} \mid \theta_{TT})\right], \qquad (16)$$

where $\delta_{j,j'}$ assumes a value of 1 if items *j* and j' are part of the same cluster and -1 otherwise. The fundamental goal of DETECT is to find the partitioning of a test that maximizes the scatter of items away from the unidimensional composite θ_{TT}. This partition, which is also deemed the true structure, will uniquely yield the highest value of the DETECT index for the observed data. The value of the DETECT index associated with this partition is called DETECT_{max}.

In practice, the value of DETECT_{max} and the test structure that produces it is often unknown. When conducting as an exploratory analysis, the partition yielding DETECT_{max} is searched for via a genetic algorithm (Zhang & Stout, 1999b). As a starting point for the algorithm, the results from three CCPROX/HCA analyses are supplied as parents; the three parents are found via the single link, complete link, and UPGMA linking methods. For each parent, *J*/5 of the items are "mutated" by cycling the *J*/5 items through all possible cluster

memberships that differ from their original clusters. Each crossing of the mutated items into different clusters represents a new generation, and the DETECT index is calculated for each new generation. If the DETECT index for the new generation is larger than the previous generation, then the partition associated with the new generation is declared as $DETECT_{max}$. The key idea underlying the genetic algorithm is to calculate the DETECT index for many possible cluster solutions towards the end of searching for the partition that yields $DETECT_{max}$. Further details about the algorithm can be found in Zhang and Stout's (1999b) original work.

The exploratory DETECT program can suggests up to a maximum of 10 dominant dimensions, but the algorithm will terminate if fewer dimensions are found to yield $DETECT_{max}$. With that said, users also have the option to set the maximum number of dimensions to fewer than 10. The current work pursues the DETECT genetic algorithm as an alternative method for obtaining an AT/PT partition that maximizes the possibility of finding evidence of additional dimensionality when it is indeed present. This was achieved by setting the maximum of number of dimensions to two and therefore searching for the two cluster test partition that yields $DETECT_{max}$.

Conditional Covariance-Based Subtest Selection Methods. The two conditional covariance-bases subtest selection methods relevant to the current work include the DETECT genetic algorithm (described above) and Froelich and Habing's (2008) ATFIND program. The ATFIND program and the exploratory version of DETECT are similar in that both programs involve a two-stage

partitioning method that includes (a) finding candidate partitions via CCPROX/HCA and then (b) finding the partition that yields the highest value of DETECT. However, there are key differences between ATFIND and the exploratory version of DETECT. First, restrictions placed on the minimum test length for ATFIND are not in place for DETECT. The ATFIND program requires a minimum of 15 PT items and four AT items and therefore cannot be used with fewer than 19 items. Second, the two-stage partitioning method is operationalized with less rigor using the ATFIND program than with DETECT. The ATFIND program first conducts a CCPROX/HCA analysis using only the UPGMA linking method (as described above) to find candidate partitions that are supplied to DETECT. Then, the DETECT index is calculated for each potential AT/PT partition. The AT for each potential test partitions includes between four and half of the items on the test; the remaining items serve as the PT. Unlike the DETECT genetic algorithm, the items are not "mutated" to search for $DETECT_{max}$.

Hypotheses

The existing research on DIMTEST (reviewed further in the following chapter) and features of the conditional covariance-based framework suggest key variables that are hypothesized to affect performance. They hypotheses relating the variables of interest to the performance of DIMTEST are discussed in this section.

Sample Size and Test Length. It is hypothesized that neither sample size nor test length will affect performance for true unidimensional conditions. In true multidimensional conditions, however, it is expected that power will become

higher with increases in sample size and/or test length. As reviewed in greater detail in the following chapter, research on earlier versions of DIMTEST suggest that the procedure is a reliable test of the null hypothesis of essential unidimensionality with a minimum of 750 examinees and 25 items. It is hypothesized that the current version of DIMTEST will yield sufficiently high power without compromising Type I error rates with fewer examinees and/or test items than required for previous versions.

Strength of Dependence. The item discrimination parameter is a measure of the strength of relationship of an item on the latent variable(s); this will herein be referred to as the strength of dependence. It is hypothesized that the strength of dependence will not affect the performance of DIMTEST in true unidimensional conditions. However, it is hypothesized that power will be positively related to increases in the strength of dependence.

Correlation between Dimensions. It was hypothesized that power will be decrease to the extent that the correlation between dimensions increases holding all else constant.

Test Structure. Holding all else constant, deviations from simple structure are hypothesized to decrease power.

AT Selection Method. It is of interest to compare the performance of DETECT and ATFIND in terms of selecting AT items. The two methods were hypothesized to similarly yield similar rejection rates across all conditions in both unidimensional and multidimensional conditions. For tests consisting of fewer than 19 items, only DETECT was be used to select AT items. The hypotheses

stated earlier pertaining to sample size, test length, the strength of dependence, the correlation between dimensions, and test structure also apply to the DETECT-only conditions.

Summary

The performance of the DIMTEST procedure (Stout, 1987; Froelich & Habing, 2008) has been well-documented for large-scale testing conditions. In small-scale contexts, however, recommendations for the use of DIMTEST (and dimensionality assessment in general) are relatively sparse. Existing recommendations, which will be reviewed in the second chapter, generally pertain to older versions of the DIMTEST procedure. The current DIMTEST procedure employs a combination of theoretically consistent methods, namely CCPROX/HCA and DETECT, to obtain the AT/PT partition and removes bias from the statistic with a bootstrapping technique. As discussed in the next chapter, the new version of DIMTEST has empirically demonstrated better performance than previous versions in that the Type I error rate remains closer to or lower than the nominal value of α and marked increases in power have been achieved for even less ideal testing conditions (Finch & Habing, 2007; Froelich & Habing, 2008). This goal of this study is extend research on the current DIMTEST procedure by providing recommendations for its use in small-scale testing conditions.

Beyond evaluating the performance of DIMTEST in small-scale testing conditions, a second purpose of this study is to evaluate the performance of the DETECT genetic algorithm for obtaining an AT/PT partition. This choice is in

part motivated by the need for obtaining an AT/PT partition for conditions with fewer than 19 items. In addition, given that the DETECT genetic algorithm and the ATFIND program share the same logic, it is of interest to compare the two methods in terms of selecting AT items for conducting DIMTEST analyses.

Chapter 2

REVIEW OF LITERATURE

This chapter reviews relevant existing research on the DIMTEST procedure and presents the hypotheses for the current work. Although the goals and logic of the DIMTEST procedure have remained consistent, aspects of the procedure and the resulting test statistic have evolved significantly since Stout's (1987) original work. In particular, the removal of AT2 and replacing FAC with ATFIND yield the current version of DIMTEST. Each instantiation of DIMTEST can be thought of as being qualitatively different in that recommendations for the use of DIMTEST are tied to the version in consideration. From this perspective, existing research will be presented in three sections pertaining to each instantiation. After presenting the relevant existing research on DIMTEST, the hypotheses for the current will be detailed.

DIMTEST with AT2 and FAC

Building off of Stout's (1987) original work, Nandakumar and Stout (1993) proposed a number of changes aimed at establishing a better match between the item difficulties for the AT1 and AT2 items and adjusting the standard error of the test statistic. With the proposed changes, DIMTEST was found to maintain the Type I error rate and achieve greater power than Stout's original formulation. One goal of the proposed changes was to reduce the Type I error rate for tests primarily consisting of highly discriminating items. Towards that end, the proposed changes resulted in a Type I error rate that was drastically reduced compared to DIMTEST without the changes. With respect to sample size

and test length requirements, Nandakumar and Stout's work suggested that DIMTEST was a reliable test of essential unidimensionality with 750 examinees and 25 items. Notably, at these levels of sample size and test length, power was significantly reduced when the correlation between θ_1 and θ_2 increased from ρ = .5 to ρ = .7.

Gessaroli and De Champlain (1996) generated data from a 2-PL model to compare the performance of their $\chi^2_{G/D}$ statistic to DIMTEST with the FAC program and AT2. Although the Type I error rate never exceeded 8 rejections out of 100 independent trials, they found that inflated error rates tended to occur with tests comprised of weakly or moderately discriminating items. In multidimensional conditions, power was primarily affected by the dominance of the second dimension and to a lesser extent by test length. Power tended to be lowest with a test length of 15 items[1]with 80% of the items measuring θ_1 and 20% measuring θ_2; increasing the length of the test and/or having strongly discriminating items increased power to acceptable levels under normal standards (rejecting the null hypothesis 80% of the time when it is false). Given their results, and in keeping with previous recommendations (Nandakumar, 1987;

[1]While the current version of DIMTEST with ATFIND requires a minimum of 19 items, it is not clear from previous research whether test length restrictions were in place with the FAC program. The fact that Gessaroli and De Champlain (1996) was able to investigate DIMTEST with as few as 15 items may be indicative that the either (a) the FAC program did not impose a constraint on the required test length or (b) the authors had access to a version of the program that is not available to the typical user.

Nandakumar & Stout, 1993; Stout, 1987), Gessaroli and De Champlain suggested that DIMTEST with AT2 and FAC should only be used with sample sizes of 750 examinees and 25 items.

Although necessary for the removal of bias, the use of AT2 reduced the flexibility of the DIMTEST procedure. In particular, in order to optimize the removal of bias from the T_L statistic, every AT2 item would ideally be matched to an item equal in difficulty to each AT1 item. The goal of doing so was to enhance the sensitivity of the test statistic T to sources of dependencies that were not a byproduct of the FAC method used for selecting AT1 items. A well-known issue of applying linear factor analysis via the FAC program was that items similar in difficulty tended to be extracted as a unique factor (McDonald & Ahlawat, 1974, Stout, 1987). Bias would not be removed from the DIMTEST statistic to the extent that the AT1 items were homogenous in terms of difficulty and distinct from the AT2 items. In these cases, DIMTEST was more prone to an inflated Type I error rate particularly with small samples and short tests (Nandakumar & Stout, 1993; Stout, 1987). With the goal of reducing the Type I error rate, and ultimately making DIMTEST more flexible, Froelich and Stout (2003) developed the bootstrapping procedure described earlier.

DIMTEST with Bootstrap Bias Correction and FAC

Froelich and Stout (2003) conducted a Monte Carlo simulation study to evaluate DIMTEST without AT2 while still including the FAC program. When

the size of the assessment subtest (AT) was automatically determined[2], the Type I error rate was consistently below the specified level of significance ($\alpha = .05$) in unidimensional conditions. For multidimensional conditions[3], Froelich and Stout investigated the influence of test structure on the bootrapped version of DIMTEST. In the simple structure model, the lowest power was observed with 750 examinees, 25 items, and highly correlated dimensions ($\rho = .7$); out of the 100 replications, 94 of which correctly resulted in a rejection of the null hypothesis. When data followed a factorially complex model, the correlation between dimensions had a large effect. Using normal conventions, power was adequately high when the correlation between dimensions was low ($\rho = .3$). Holding all else constant, increasing the correlation to $\rho = .7$ resulted in marked decreases in power; adequate levels of power were never achieved in conditions with test lengths of 25 items. Froelich and Stout concluded that the significant drop in power was attributable to a failure of the FAC program to select a sufficient number of dimensionally similar items for AT. In light of this view, they suggested an alternative method employing HCA/CCPROX and DETECT, which eventually became the version of DIMTEST currently in use.

---

[2] Three levels of AT size were varied in the Type I error study.The three levels included J/4, J/2, or the size of the AT was allowed to vary between replications.

[3] For the power study, three methods were used for the selection of AT items. The method closest to that of the current exploratory DIMTEST procedure used the FAC program and allowed the size of AT to vary between replications. For the sake of comparison to research on the current version, only the results pertaining to this method were considered.

It is noteworthy that the performance of Froelich and Stout's (2003) DIMTEST without AT2 was comparable to Nandakumar and Stout's (1993) DIMTEST with AT2. However, the use of the bootstrapping procedure for removing bias is more flexible than selecting AT2 items. Whereas the AT2 items were chosen based on their similarity to AT1 items in terms of difficulty, the bootstrapping procedure is more flexible in that other features of the ICCs, such as the estimated discrimination and pseudo-guessing parameters, can be taken into account. Overall, the bootstrapping procedure makes the DIMTEST procedure more flexible than AT2 while not increasing the Type I error rate or compromising power.

DIMTEST with Bootstrap Bias Correction and ATFIND

As mentioned, the current version of DIMTEST removes the need for AT2 via Froelich and Stout's (2003) bootrapping method and employs ATFIND, which uses a combination of HCA/CCPROX and DETECT, to select AT items. Froelich and Habing (2008) conducted a Monte Carlo simulation study to compare the performance of the ATFIND and FAC methods for selecting AT items; in all conditions, bias in the test statistic was removed with the bootrapping procedure. Compared to the FAC program, they found that DIMTEST with ATFIND maintained similar control of the Type I error rate while improving power, particularly for tests deviating from simple structure with dimensions correlated at $\rho \geq .7$.

Finch and Habing (2007) compared the performance of the current version of DIMTEST to other statistics for testing the unidimensionality assumption.

They found that DIMTEST generally maintained the Type I error rate and was highly powered even with highly correlated dimensions (ρ = .8) with tests as short as 15 items[4]. Notably, their smallest sample size condition was 1000 examinees; their results suggest that DIMTEST may be a reliable test of the unidimensionality hypothesis with fewer items and/or smaller samples.

[4] The version of ATFIND is a constrained version of the program that is not available to typical users. Since the typical user cannot release the constraint that at least 19 items are required, it is unclear how Finch and Habing (2007) managed to use the program for 15 item tests. This suggests that the authors were using a version of ATFIND that is no longer in use or was a modified version available to the authors.

Chapter 3

METHODOLOGY

The goal of this Monte Carlo simulation study was to re-evaluate the sample size and test length minima for the current DIMTEST procedure. Drawing from previous research, other factors known to affect the performance of DIMTEST were also manipulated. In addition to sample size and test length, other manipulations included the strength of dependence, and for the power study only, the correlation between dimensions and test structure; these manipulations are detailed below. Then, the process for conducting DIMTEST is described. Finally, the data analytic procedures are described.

Data Generation

Using R 2.13.0 (R Core Team Development, 2011), *Equation 2* was used to generate dichotomous item responses such that $X_{ij} = 1$ for a correct response and $X_{ij} = 0$ for an incorrect response. For unidimensional conditions, the unidimensional IRT (*Equation 1*) model was obtained by setting the correlation between dimensions and all *a*-parameters on the second dimension to zero. The *c*-parameter was set to zero for all items; that is, all datasets were generated from the 2-PL model. Although it is well-recognized that $c > 0$ is common in practice, recent work has suggested that the *c*-parameter has little impact on the performance of the current version of DIMTEST (Finch & Habing, 2007) for the sample sizes considered in this work.

Number of Replications. From an empirical perspective, the goal was to determine the number of replications that would be necessary to estimate Type I

error rates and power with sufficient precision (i.e., low standard error). Each replicate dataset within a condition is an independent trial with a binary outcome (0 = do not reject H_O, 1 = reject H_O); to estimate the rejection rate for a given condition, the outcomes are summed and divided by the number of trials to form a proportion. Each condition is therefore a representative sample proportion (p) of a population proportion that can be characterized by a binomial distribution. The standard error (SE_p) for the binomial distribution is given by:

$$SE_p = \sqrt{\frac{p(1-p)}{R}}, \tag{17}$$

such that p is the value of the sample proportion and R represents the number of replications (i.e., independent trials). Given the expected value of p (denoted $\hat{p}$) and the desired standard error of $\hat{p}$ ($SE_{\hat{p}}$), R was determined by:

$$R = \frac{\hat{p}(1-\hat{p})}{SE_{\hat{p}}^2}. \tag{18}$$

A confidence interval, which is based on the normal approximation to the binomial distribution, can also be formed around $\hat{p}$ such that:

$$\hat{p} \pm Z_{1-\alpha/2} \times SE_{\hat{p}} \tag{19}$$

where $Z_{1-\alpha/2}$ represents the two-tailed critical value associated with the level of α chosen for the binomial distribution.

The choices of $\hat{p}$ for Type I error rates and power were chosen based on past research on DIMTEST and standards that are common in social science research. Accordingly, the Type I error rate was set to α = .05 ($\hat{p}$ = .05) and

DIMTEST was deemed sufficiently powerful if 80% ($\hat{p} = .8$) of the replicate datasets resulted in rejection of a false null hypothesis. Given these values of $\hat{p}$, R was calculated for different values of $SE_{\hat{p}}$. Table 1 shows the resulting values of R across the values of $SE_{\hat{p}}$ considered along with 95% and 99% confidence intervals around $\hat{p}$. Based on these results, it was decided that $SE_{\hat{p}} = .0145$ struck an appropriate balance between precision and available computational resources. For the Type I error study and power studies, $R = 226$ and $R = 761$, respectively, to achieve $SE_{\hat{p}} = .0145$. To be conservative, the number of replications was set to 800 for both the Type I error studies and power studies. In effect, the precision was greater in the Type I error study ($SE_{\hat{p}} = .008$) than in the power ($SE_{\hat{p}} = .014$) study.

Table 1

Number of Replications (R) Given $SE_{\hat{p}}$ *with 95% and 99% Confidence Intervals (CI)*

	R		*q* = .05 CI		*q* = .8 CI	
$SE_{\hat{p}}$	*q* = .05	*q* = .8	95%	99%	95%	99%
.0200	119	400	[.011 .089]	[-.002 .102]	[.761 .839]	[.748 .852]
.0195	125	421	[.012 .088]	[.000 .100]	[.762. 838]	[.750 .850]
.0190	132	443	[.013 .087]	[.001 .099]	[.763. 837]	[.751 .849]
.0185	139	467	[.014 .086]	[.002 .098]	[.764. 836]	[.752 .848]
.0180	147	494	[.015 .085]	[.004 .096]	[.765. 835]	[.754 .846]
.0175	155	522	[.016 .084]	[.005 .095]	[.766. 834]	[.755 .845]
.0170	164	554	[.017 .083]	[.006 .094]	[.767. 833]	[.756 .844]
.0165	174	588	[.018 .082]	[.007 .093]	[.768. 832]	[.757 .843]
.0160	186	625	[.019 .081]	[.009 .091]	[.769. 831]	[.759 .841]
.0155	198	666	[.020 .080]	[.010 .090]	[.770. 830]	[.760 .840]
.0150	211	711	[.021 .079]	[.011 .089]	[.771. 829]	[.761 .839]
.0145	226	761	[.022 .078]	[.013 .087]	[.772. 828]	[.763 .837]
.0140	242	816	[.023 .077]	[.014 .086]	[.773. 827]	[.764 .836]
.0135	261	878	[.024 .076]	[.015 .085]	[.774. 826]	[.765 .835]
.0130	281	947	[.025 .075]	[.017 .083]	[.775. 825]	[.767 .833]
.0125	304	1024	[.026 .074]	[.018 .082]	[.776. 824]	[.768 .832]
.0120	330	1111	[.026 .074]	[.019 .081]	[.776. 824]	[.769 .831]

Generation of Person Parameters. Three levels of sample size (250, 500, 750) were chosen for the current work. The largest sample size considered in this study has commonly been used in other research on DIMTEST, thereby allowing for comparison across studies. In contrast, sample sizes of 250 and 500 have not been investigated with the current version of DIMTEST and therefore are reflective of the goals for this study. It is recommended that different subsets of examinees be used for selecting AT items and calculating the DIMTEST statistic. Accordingly, one third of the examinees were used for selecting AT items (83, 167, 250) with the remaining examinees (167, 333, 500) used for calculating the DIMTEST statistic.

Person parameters were generated for each replication such that $\boldsymbol{\theta} = (\theta_1, \theta_2) \sim N(0, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is the variance-covariance matrix; the variance for each dimension was set to one, and the off-diagonal element of $\boldsymbol{\Sigma}$ was set to the desired correlation. The correlation between dimensions was manipulated in the power study to be either uncorrelated ($\rho = 0$), moderate ($\rho = .35$), or strong ($\rho = .7$).

Generation of Item Parameters. Five levels of test length (9, 15, 21, 27, 33) were chosen based on previous research and with the goals of this work in mind. One aspect of the items that was hypothesized to affect the performance of DIMTEST was the strength of dependence of item responses on the underlying dimensions (weak, moderate, strong). Item discrimination parameters were generated from a random uniform distribution for each replication within a condition. Table 2 shows the minima and maxima of the random uniform distribution used for each level of strength of dependence in the IRT

parameterization and the standardized factor loading metric. In terms of item difficulty parameters, Table 3 shows the fixed values of *d*-parameters for each level of test length; the values of *d*-parameters shown for any test length were simply repeated three times in order to arrive at the full test length. It is recognized that there is a confound between test length and the variability in the fixed item difficulties. Research on earlier versions of DIMTEST indicated that performance in multidimensional conditions is affected by the relative location of item difficulty parameters to person parameters and the variability of the generating item difficulty distribution (Seraphine, 2000). More recent research by Froelich and Habing (2008) suggests that these features of item difficulty parameters do not affect the performance of the current version of DIMTEST.

Table 2

Minima and Maxima of Random Uniform Distributions for Generating Discrimination Parameters

Strength of Dependence	IRT Discrimination		Standardized Factor Loading	
	Minimum	Maximum	Minimum	Maximum
Weak	.50	1.00	.28	.51
Moderate	.75	1.25	.40	.59
Strong	1.25	1.75	.59	.72

Note. The IRT discrimination parameters were transformed to standardized factor loadings via $\lambda_j = a_j / D \Big/ \sqrt{1 + (a_j / D)^2}$ where *D* is a scaling constant equal to 1.7 (Wirth & Edwards, 2007).

Table 3

Generating Item Difficulty Parameters

Test Length				
9	15	21	27	33
-.75	-1.5	-1.5	-1.5	-1.5
0	-.75	-1	-1.13	-1.2
.75	0	-.5	-.75	-.9
	.75	0	-.38	-.6
	1.5	.5	0	-.3
		1	.38	0
		1.5	.75	.3
			1.13	.6
			1.5	.9
				1.2
				1.5

Note. The sequence for each test length was repeated three times to produce the full test length.

Test Structure. Froelich and Habing's (2008) method was used to manipulate test structure. Their method consisted of (1) generating unidimensional *a*-parameters (denoted a_j), (2) assigning each item an angle (denoted β_j), and (3) calculating the discrimination parameters for the first (a_{j1}) and second (a_{j2}) dimensions with the following equations:

$$a_{j1} = a_j \cos(\beta_j) \text{ and } a_{j2} = a_j \sin(\beta_j). \quad (20)$$

For unidimensional conditions, $\beta_j = 0$ for all items; that is, all items exclusively measured θ_1 and did not measure θ_2. In simple structure conditions $\beta_j = 0$ (measure θ_1 exclusively) for two thirds of the items and $\beta_j = 90$ (measuring θ_2 exclusively) for the remaining items. In complex structure conditions angles were uniformly distributed for each replication such that $0 \leq \beta_j \leq 20$ (measure θ_1 best) for two thirds of the items and $70 \leq \beta_j \leq 90$ (measure θ_2 best) for the remaining items.

Conducting DIMTEST

A number of decisions must be made by the user when conducting a DIMTEST analysis. The first decision is also the most important; users must partition the items into the PT and AT. As mentioned earlier, one goal of this study was to compare the performance and behavior of the ATFIND and exploratory DETECT programs for selecting AT items. The details for using these programs will be discussed first. Then, the specification of parameters for running DIMTEST will be elicited.

Parameters of Subtest Selection Methods. Although similar, the ATFIND and DETECT programs require different decisions on the user's part. When conducting ATFIND, the only input necessary is an estimate of the *c*-parameter. Since data were generated without guessing, $c = 0$ for all conditions in the study. The DETECT program requires more input than ATFIND. Specifically, one must select the type of analysis (confirmatory or exploratory), whether a

cross-validation analysis will be conducted, the maximum number of dimensions, and a seed number. For this work, exploratory DETECT with no cross-validation was used to maximize the DETECT index for a two dimensional solution; the default seed number (99991) was used.

DIMTEST Parameters. The parameters that must be specified for DIMTEST exclusively pertain to the bootstrapping procedure for removing bias from the test statistic. Three parameters are used for estimating the ICCs for each item – an estimate of the *c*-parameter, the number of evaluation points for smoothing ICCs, and the seed number. Since data were generated without guessing, $c = 0$ for all conditions. To smooth ICCs, 50 evaluation points (which is the default) were used. The default seed number (99991) was used. The final parameter that must be specified is the number of bootstrap replications; the default of 100 replications was used.

Data Analysis

Type I error and power estimates were calculated for each condition by counting up the number rejections and dividing by the number of replications which was set to 800.

Summary

This chapter has detailed the process for generating data, conducting the DIMTEST analyses for each replicated dataset, and analyzing the results. The factors manipulated included sample size, test length, and the strength of *a*-parameters for the Type I error and power studies. The inter-dimensional correlation and test structure were also manipulated in the power study. The

minimum number of replications required to achieve pre-specified levels of precision was determined empirically and tempered by practical considerations; the number of replications for each condition was set to 800. In terms of data analyses, Type I error rates and power were estimated by simply calculating the proportion of replications that resulted in a rejection of the null hypothesis.

Chapter 4

RESULTS

This presentation of results is split into two sections. The first section presents the rejection rates for DIMTEST; the performance of ATFIND and DETECT with respect to selecting subtests is presented in the second section. Within each of these sections, the results for unidimensional and multidimensional conditions are shown separately.

Rejection Rates

Unidimensional Conditions: Estimation of Type I Error Rates. Figure 5 presents the estimated Type I error rates based on 800 independent trials in each of the 45 unidimensional conditions. Each of the 15 panels corresponds to one of the combinations of sample size (the three rows of panels) and test length (the five panels within each row). The Type I error rate appears on the vertical axis, and the three levels of strength of dependence (weak, moderate, and strong) are shown on the horizontal axis. The black dashed line cutting across each panel marks the nominal rate of $\alpha = .05$. The rejection rates for DIMTEST are shown separately for test partitions obtained with ATFIND (denoted by circle markers) and DETECT (denoted by triangle markers). In conditions generated with 21, 27, and 33 items, the ATFIND and DETECT programs were applied to the same simulated datasets and are therefore directly comparable. Estimated Type I error rates for nine and 15-item tests are only shown for DETECT since it was not possible to use ATFIND for these conditions.

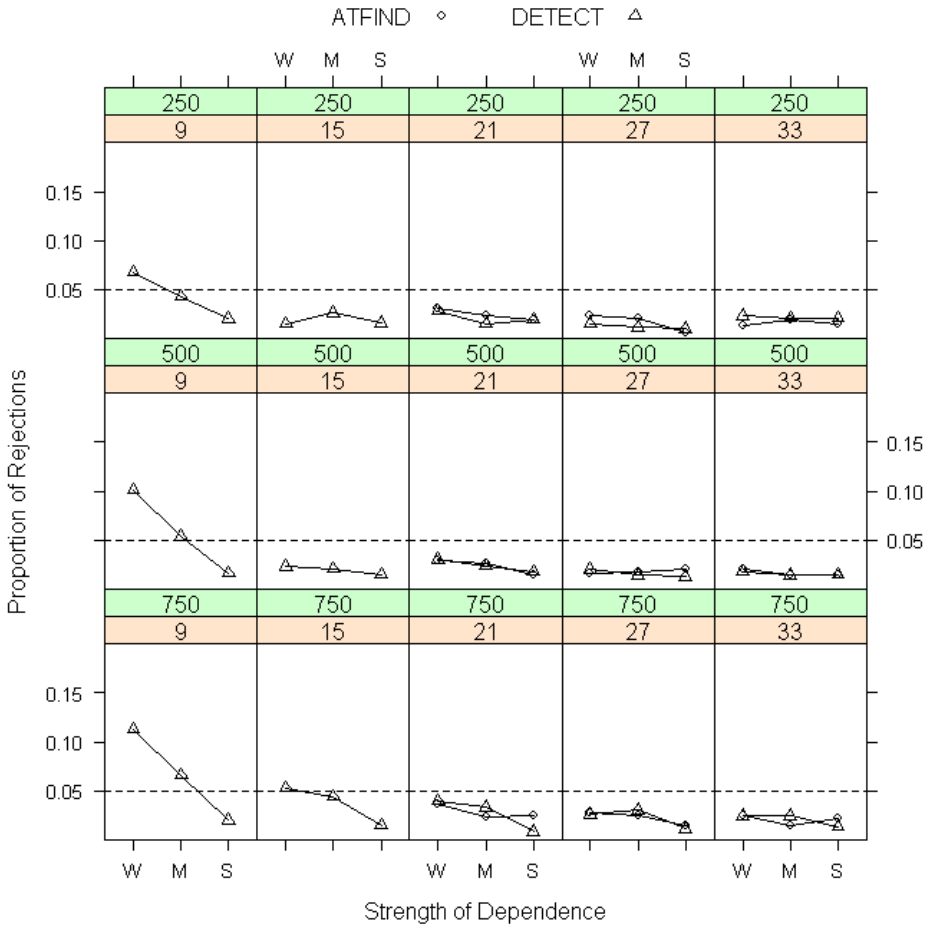


Figure 5. Proportion of rejections in 800 independent trials for unidimensional conditions. Values in the green and yellow bars correspond to the levels of sample size and test length, respectively. The dashed line marks the nominal rate of rejection ($\alpha = .05$). W = weak strength of dependence, M = moderate strength of dependence, S = strong strength of dependence.

In general, DIMTEST was quite conservative across most conditions; that is, Type I error rates were below the nominal value of $\alpha = .05$ for most conditions. Type I error rates were slightly higher than the nominal value for some conditions but the degree of inflation may have been the result of sampling error. However,

Type I error rates were approximately twice the nominal rate for nine item tests consisting of weakly discriminating items and at least 500 examinees. With the exception of short tests consisting of weakly discriminating items, the manipulated variables had trivial effects on the observed Type I rates. Notably, DIMTEST performed similarly when ATFIND and DETECT were applied to the same datasets for partitioning test items.

Multidimensional Conditions: Estimation of Power. Figures 6 through 8 graphically present the proportion of rejections for multidimensional conditions, or power, for each of the three levels of strength of dependence. The power of DIMTEST for all other combinations of the manipulated factors (sample size, test length, strength of dependence, test structure, test partitioning method) is shown in each figure. Each figure consists of 15 panels that represent the combinations of test length and sample size. Within any one panel, the three levels of dimensional correlation appear on the horizontal axis, and power is shown on the vertical axis. The lines within each panel represent one of the four combinations of test structure and test partitioning method. The markers differentiate the test partitioning method such that DETECT is denoted with triangle markers and ATFIND is denoted by circle markers. The lines differentiate test structure such that solid lines correspond to simple structure and dashed lines correspond to complex structure.

Figure 6 presents the power results for strongly discriminating items. Several main effects and interactive relationships were found among the manipulated variables and power. Looking within each panel, the main effect of

increasing the correlation between dimensions is seen by looking across the horizontal axis and the main effect of test structure is seen by comparing the solid and dotted lines. The main effect of test length is seen looking across panels within a row holding constant the correlation between dimensions and subtest selection method. Finally, the main effect of sample is seen looking across panels within a column holding constant the correlation between dimensions and subtest selection method.

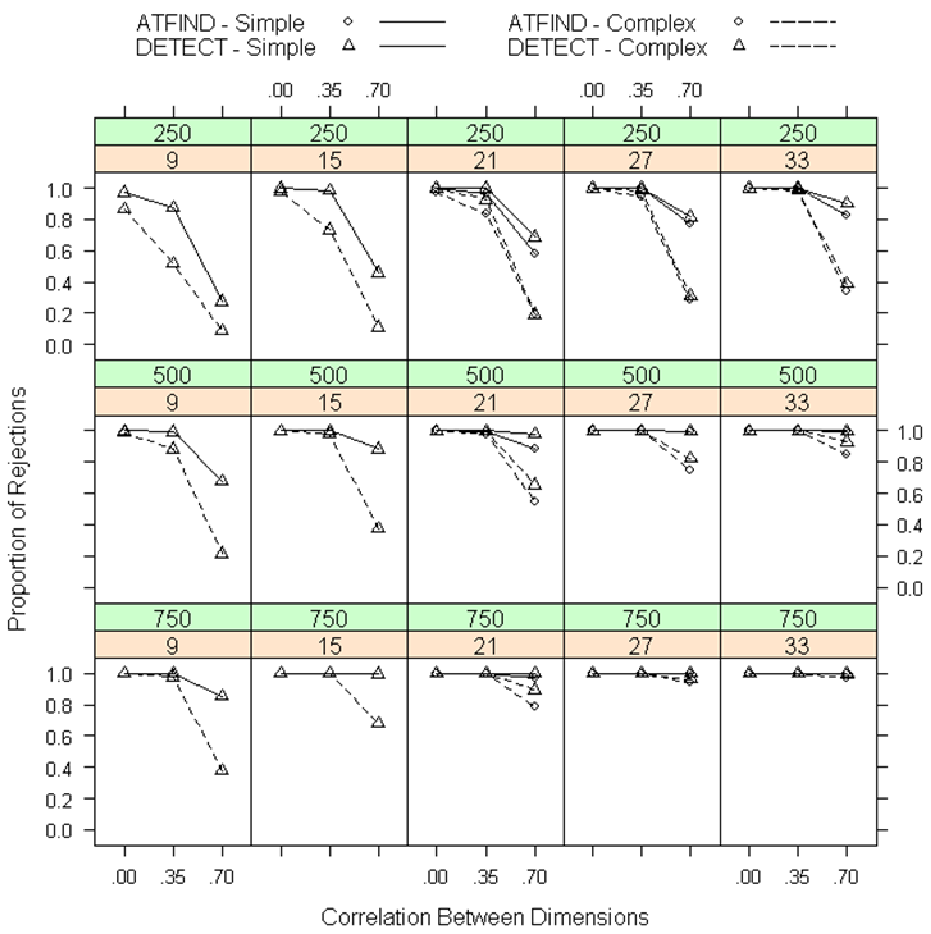


Figure 6. Proportion of rejections in 800 independent trials for multidimensional conditions with strongly discriminating items. Values in the green and yellow bars correspond to the levels of sample size and test length, respectively.

Owing to the presence of ceiling effects for many conditions with dimensions correlated at $\rho = 0$ and $\rho = .35$, the main effects of test length, sample size, test structure, and subtest selection method were most clearly seen with dimensions correlated at $\rho = .7$. That is, there were conditional main effects across the levels of test length, sample size, test structure, and subtest selection method conditional on dimensions being correlated at $\rho = .7$. With dimensions correlated

at $\rho = .7$, key main effects included positive relationships test length, sample size, simple structure relative to complex structure, and DETECT compared to ATFIND.

Key interactive relationships were also observed with strongly discriminating items. The gains in power with increases in test length were amplified with simultaneous increases in sample size. The positive effects of the two-way interaction between sample size and test length, however, were dependent on test structure and the correlation between dimensions. For example, the positive effect of test length on power for test following a factorially complex model with dimensions correlated at $\rho = .7$ was smaller with 250 examinees than with 500 examinees. Holding all else constant, the story was differed somewhat for data following a factorially simple model. Power was generally high across the levels of test length with 500 examinees and therefore left relatively little room for improvement; the same was not true with 250 examinees and allowed power to consistently improve with increases in test length.

Figure 7 shows the power results for moderately discriminating items. Compared to conditions with strongly discriminating items, a general drop in power was observed with moderately discriminating items. This finding highlights the positive main effect of having more strongly discriminating items. The main effects of other manipulated variables as discussed above were also seen with moderately discriminating items. Looking within panels, the effect of correlated dimensions is clearly seen to have a negative relationship with power particularly for data following exhibiting complex structure. For many conditions

generated with complex structure and dimensions correlated at ρ = .7, floor effects were commonly observed with 250 and 500 examinees and to a lesser extent with 750 examinees.

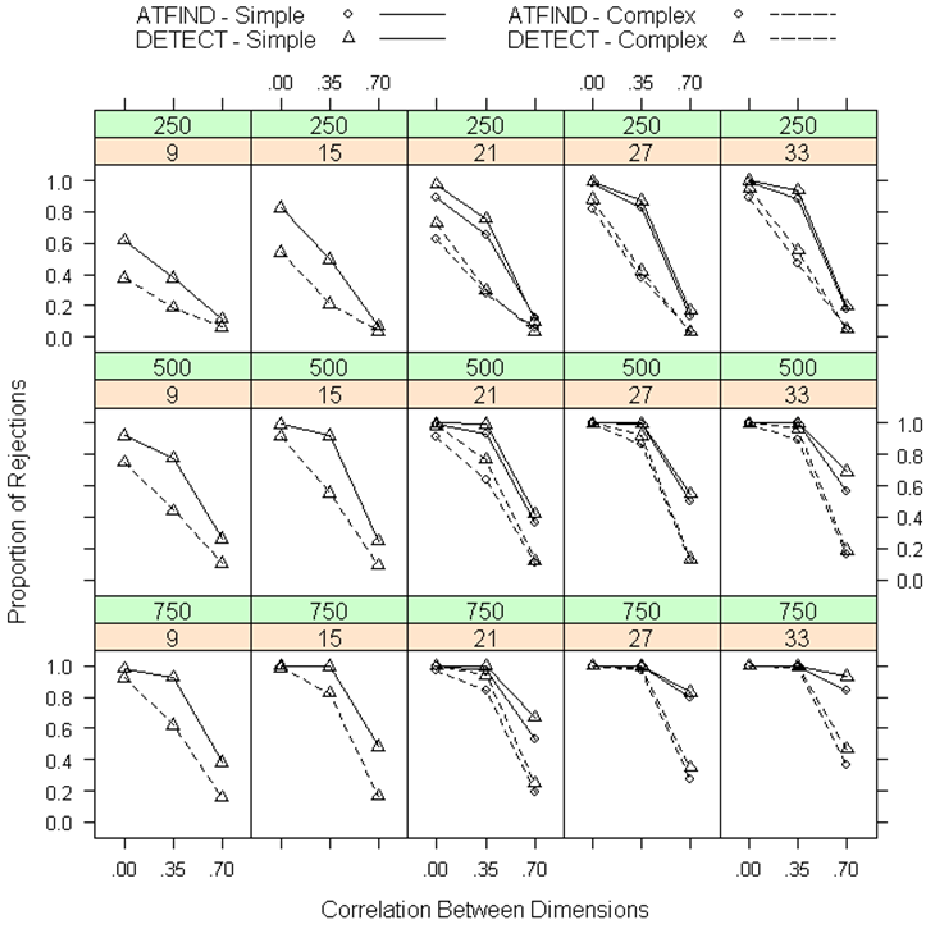


Figure 7. Proportion of rejections in 800 independent trials for multidimensional conditions with moderately discriminating items. Values in the green and yellow bars correspond to the levels of sample size and test length, respectively.

Interactive relationships with power were also observed with moderately discriminating items. As was the case with strongly discriminating items, the positive relationship of test length was amplified by increased sample size; this

relationship again depended on the levels of the correlation between dimensions and test structure. For example, for data following a factorially simple model with dimensions correlated at $\rho = .7$, the positive moderating relationship between sample size and test length is seen with 500 and 750 examinees, but on account of floor effects, not with 250 examinees. A similar, but less pronounced trend was also observed for tests following a complex loading structure with dimensions correlated at $\rho = .35$. Importantly, the nature of these interactive relationships was not seen with strongly discriminating; this highlights the moderating influence of strength of dependence.

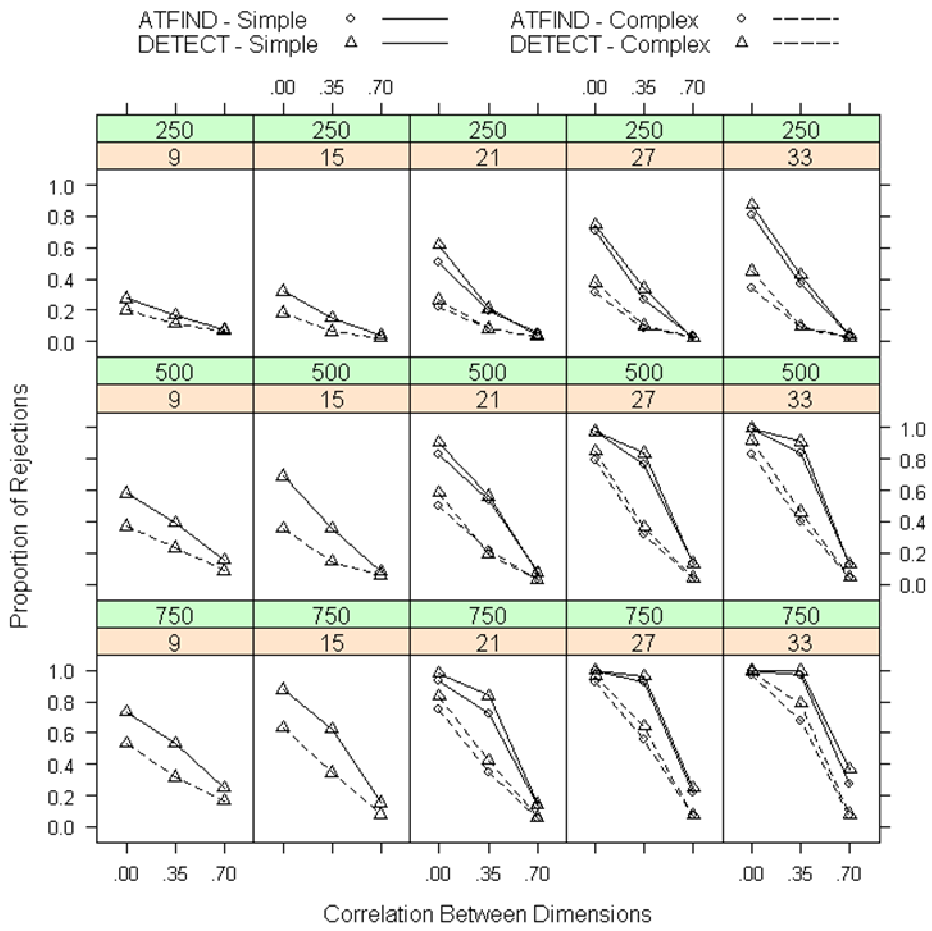


Figure 8. Proportion of rejections in 800 independent trials for multidimensional conditions with weakly discriminating items. Values in the green and yellow bars correspond to the levels of sample size and test length, respectively.

Figure 8 shows the power obtained with weakly discriminating items was lower than was observed in conditions with more strongly discriminating items; this highlights the main effect of strength of dependence. The positive moderating relationship of sample size and test length was not observed with dimensions correlated at $\rho = .7$ irrespective of test structure on account of very low power across all combinations of test length and sample size. The positive moderating

role of sample size and test length on power, however, was more clearly seen with dimensions correlated at either $\rho = 0$ or $\rho = .35$. This was particularly true with 250 and 500 examinees for data that followed a factorially simple model and with 750 examinees for data that followed a factorially complex model.

Summary of Trends in Power. Key trends between the manipulated variables and power include positive relationships with sample size, test length, and strength of dependence; negative relationships were found between the correlation between dimensions and for test following a factorially complex rather than factorially simple model. However, the magnitude of main effects often depended on the levels of other variables. The positive effect of test length was amplified by increased sample size; this moderating relationship in turn was moderated by characteristics of the underlying model such as the test structure, the correlation between dimensions, and the strength of dependence.

Evaluation of Test Partitions

The secondary focus of this study was to compare the test partitions obtained with ATFIND and DETECT. The previous section showed *how* the rejection rates obtained with DIMTEST were affected by the manipulated variables. As acknowledged earlier, the performance of DIMTEST is inherently tied to the partitioning of test items. By focusing on the test partitions obtained with ATFIND and DETECT, this chapter provides insight as to *why* the DIMTEST program performed as it did with respect to rejection rates.

Unidimensional Conditions. The test partitions obtained in unidimensional conditions were evaluated in two ways. First, the average AT

length was calculated separately for ATFIND and DETECT across the 800 replications in each condition; the results are shown in Figure 9 for each of the 45 unidimensional conditions. The overall structure is similar to Figures 5 – 8 with the exception of the vertical axis. The positive relationship between total test length and AT length is inherent. In order to evaluate the role of test length meaningfully, it was necessary to place the average AT length on a common scale. Towards that end, the vertical axis shows the ratio of the average observed AT length to the total test length; values are interpreted as the average percentage of the total test length that were partitioned to the assessment subtests across 800 independent trials. While the manipulated variables generally had little bearing on the average AT length, clear differences were found between ATFIND and DETECT for tests consisting of 21 or more items. In particular, DETECT tended to produce longer assessment subtests on average than ATFIND. The difference in average AT length between ATFIND and DETECT, however, tended to decrease with increases in test length.

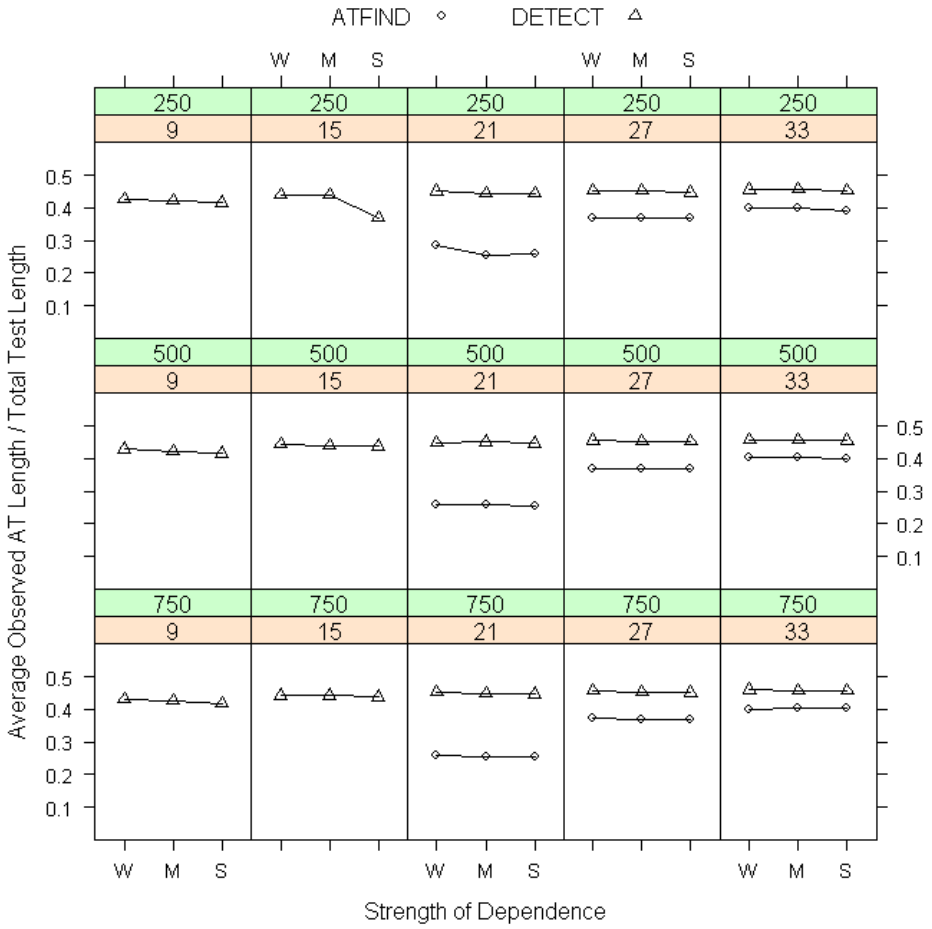


Figure 9. Average length of assessment subtests across 800 independent trials. Values in the green and yellow bars correspond to the levels of sample size and test length, respectively. W = weak strength of dependence, M = moderate strength of dependence, S = strong strength of dependence.

The ATFIND and DETECT-generated test partitions were also compared directly. For the purposes of the current work, three categories were used for comparing the test partitions. The first of which was strict agreement. For any one test partition, strict agreement was satisfied if the AT found with ATFIND was identical to the AT found with DETECT. The second category, which consists of

two possibilities, required that the AT obtained with one method be a proper subset of the AT obtained with the other. This situation occurred if (a) the assessment subtests obtained with the ATFIND and DETECT programs differed in length, and (b) the shorter AT was a proper subset of the longer AT. If one AT was a subset of the other, then the primary difference was length. The third category was disagreement. Assessment subtests were deemed to disagree if (a) they differed in terms of length and the shorter AT was not a proper subset of the longer AT, or (b) they were the same length but differed by at least one item.

The results for comparing test partitions are shown in Figure 10. The panels in the figure correspond to one of the nine combinations of test length (21, 27, 33) and sample size (250, 500, 750) in which ATFIND and DETECT were both used to partition test items. The four levels of comparison among the test partitions are denoted by different markers. Agreement is denoted by circles; disagreement is denoted by triangles; the addition operator denotes ATFIND being a subset of DETECT; and the multiplication operator denotes DETECT being a subset of ATFIND. The proportion of rejections out of 800 independent trials is shown on the vertical axis, and the three levels of strength of dependence are shown on the horizontal axis.

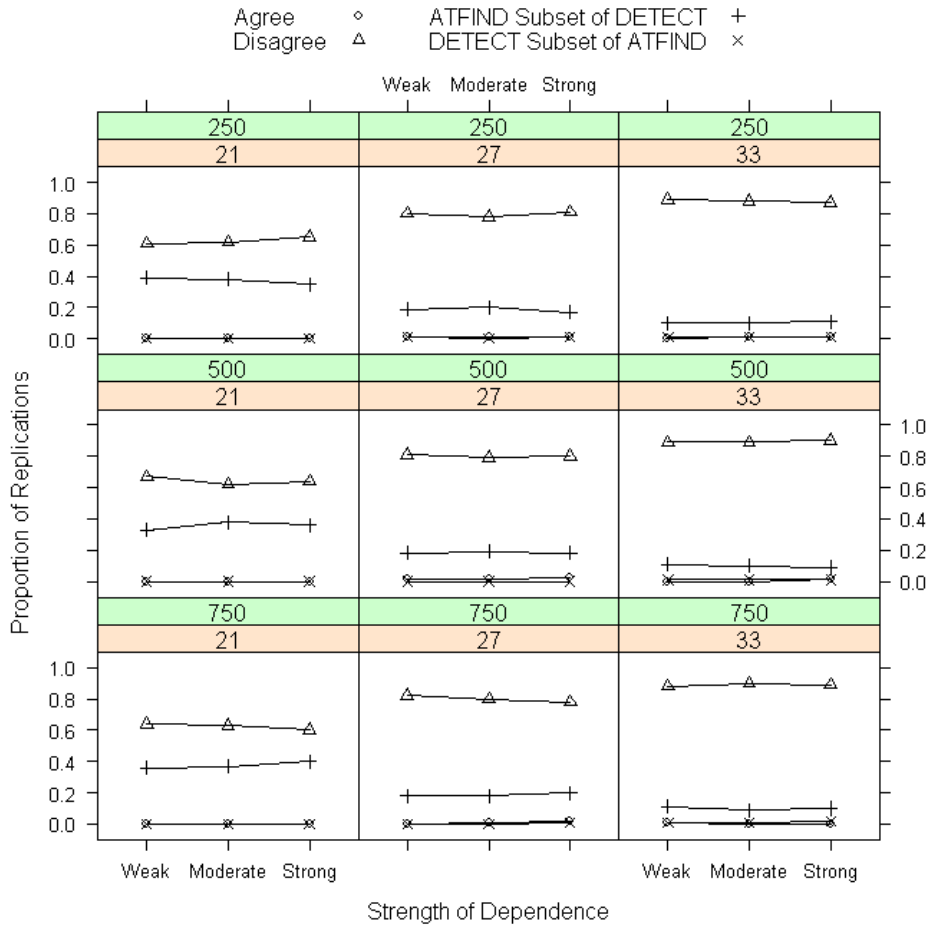


Figure 10. Comparison of similarity of ATFIND and DETECT-generated AT subtests across 800 independent trials for unidimensional conditions. Values in the green and yellow bars correspond to the levels of sample size and test length, respectively.

For tests following a unidimensional structure, ATFIND- and DETECT-generated test partitions were most likely to disagree. When test partitions did not disagree, ATFIND was most likely to produce an AT that was a subset of the AT found with DETECT. As evidenced above, ATFIND produced shorter assessment subtests, on average, than DETECT. Taken together these results indicate that

when directly compared the AT found with ATFIND would be shorter and either (a) be a proper subset of the DETECT AT, or more likely, (b) include at least one item that was not observed on the DETECT AT. Notably, disagreement became more likely to the extent that test length increased.

Multidimensional Conditions. Like the unidimensional conditions, test partitions in multidimensional conditions were evaluated in terms of length and similarity between the assessment subtests found with ATFIND and DETECT. Unlike the unidimensional conditions, the test partitions that ought to result when applying a subtest selection method were known for multidimensional conditions. Therefore, this afforded the opportunity to also evaluate the accuracy of the test partitions obtained with ATFIND and DETECT. The following presentation of results is ordered such that accuracy is shown first, followed by the similarity of test partitions, and finally the average AT length across conditions is shown last.

Accuracy of Test Partitions. Before presenting the results, it's necessary to first define accuracy. A test partition was deemed accurate if the observed AT (and therefore the PT) was identical to the true test partition. Multidimensional data were simulated such that the first two thirds of the items best measured the first dimension, and the remaining one third of the items best measured the second dimension. Accordingly, a test partition was labeled as accurate if the PT consisted only of the first two thirds of the items and the AT consisted only of the final one third of the items. Although this requirement for accuracy is strict, this definition has been employed by Zhang and Stout (1999b) and Roussos and Ozbek (2006) in their investigations of DETECT. The results for accuracy are

presented graphically. Figures 11 through 13 shares an identical structure to those used to present rejection rates in multidimensional conditions. The primary difference is the vertical axis; instead of the proportion of rejections, the vertical axis now shows the proportion of 800 replications that resulted in accurate test partitions.

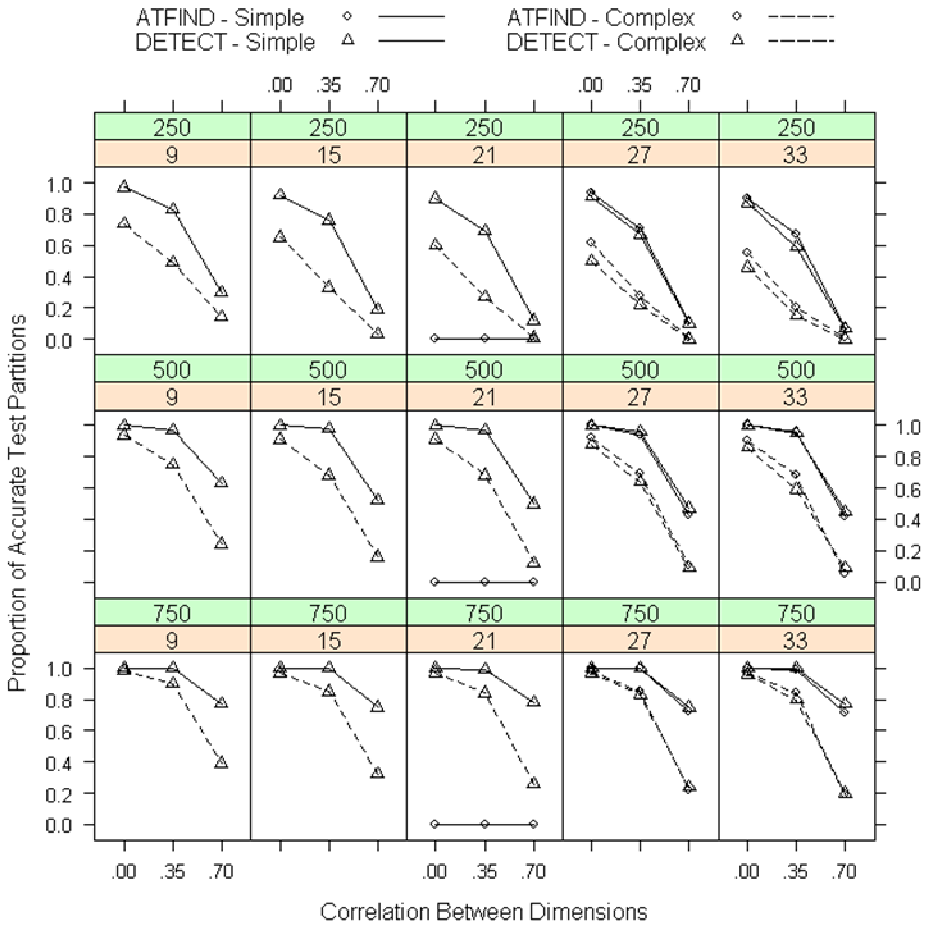


Figure 11. Proportion of accurate test partitions across 800 independent trials for multidimensional conditions with strongly discriminating items. Values in the green and yellow bars correspond to the levels of sample size and test length, respectively.

Figure 11 shows the proportion of accurate test partitions across all conditions generated with strongly discriminating items. Key main effects on accuracy include a (a) positive relationship with sample size, (b) negative relationship with the correlation between dimensions, (c) negative relationship with test length, and (d) a decline in accuracy for data following a factorially complex rather than factorially simple model. With exception to tests generated with 21 items, the performances of ATFIND and DETECT were similar. An interactive relationship involving sample size, test structure, and the correlation between dimensions was also found. With 250 examinees, the effect of test structure was most clearly seen with dimensions correlated at $\rho = 0$ and $\rho = .35$; accuracy was consistently low with dimensions correlated at $\rho = .7$ irrespective of test structure. In contrast, accuracy was consistently high with uncorrelated dimensions with 500 and 750 examinees. For these conditions, increasing the correlation between dimensions more strongly separated out the effects of test structure.

Similar trends were also seen in conditions with moderately discriminating conditions as shown in Figure 12. The main effects enumerated above with highly discriminating items also apply with moderately discriminating items. The nature of the interactive relationship among sample size, test structure, and the correlation between dimensions was moderated by decrease in the strength of dependence. Across all levels of sample size, the effect of test structure was strongest when dimensions were either uncorrelated or correlated at $\rho = .35$.

However, the magnitude of the effect of test structure was generally larger with sample sizes of 500 and 750 examinees than with 250 examinees.

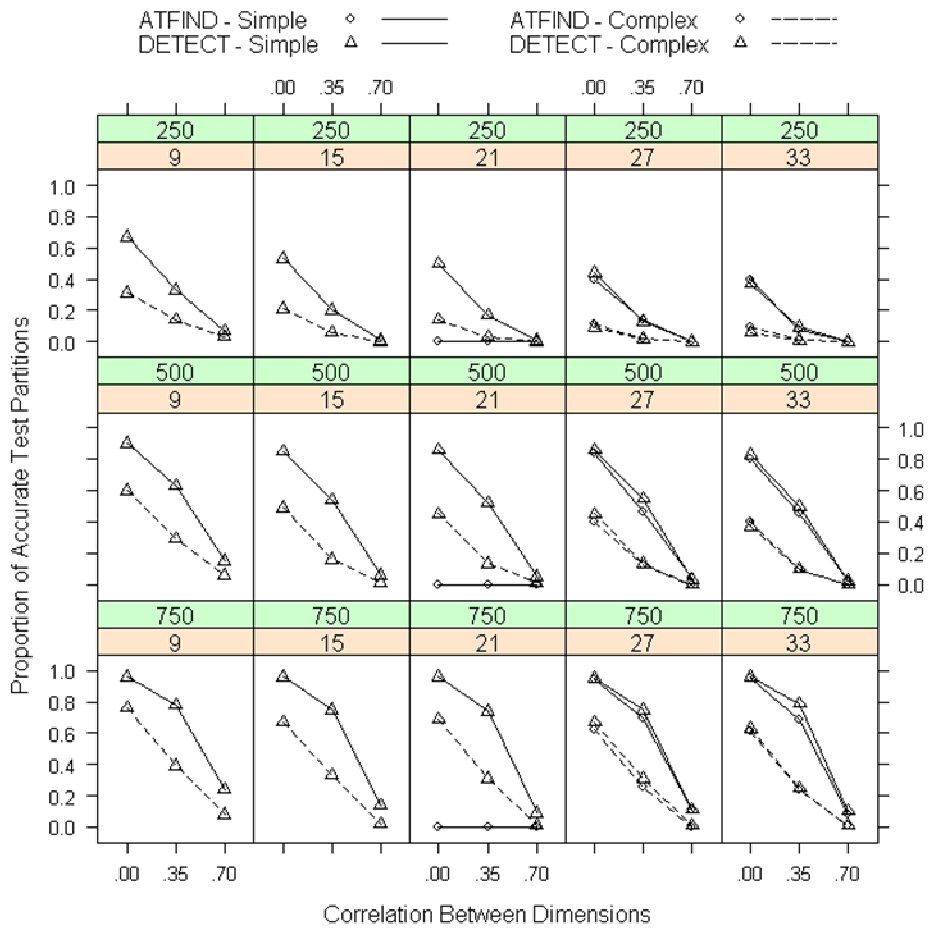


Figure 12. Proportion of accurate test partitions across 800 independent trials for multidimensional conditions with moderately discriminating items. Values in the green and yellow bars correspond to the levels of sample size and test length, respectively.

Figure 13 shows the proportion of accurate rejections for weakly discriminating items. The main effects enumerated with more strongly discriminating conditions also applied here, and the general structure of the

interactive effects was similar to those found with moderately discriminating conditions. With respect to the interactive relationship, however, differences between test structure became more pronounced with either uncorrelated or dimensions correlated at $\rho = .35$ to the extent that sample size increased. Across all levels of sample size, accuracy was consistently close to zero with for tests that were factorially complex with dimensions correlated at $\rho = .35$; holding the correlation between dimensions constant at $\rho = .35$, accuracy tended to improve with factorially simple tests to the extent that sample size was increased. A similar trend was also found with uncorrelated dimensions; differences found between factorially complex and factorially simple tests were smaller with 250 examinees than with 500 and 750 examinees.

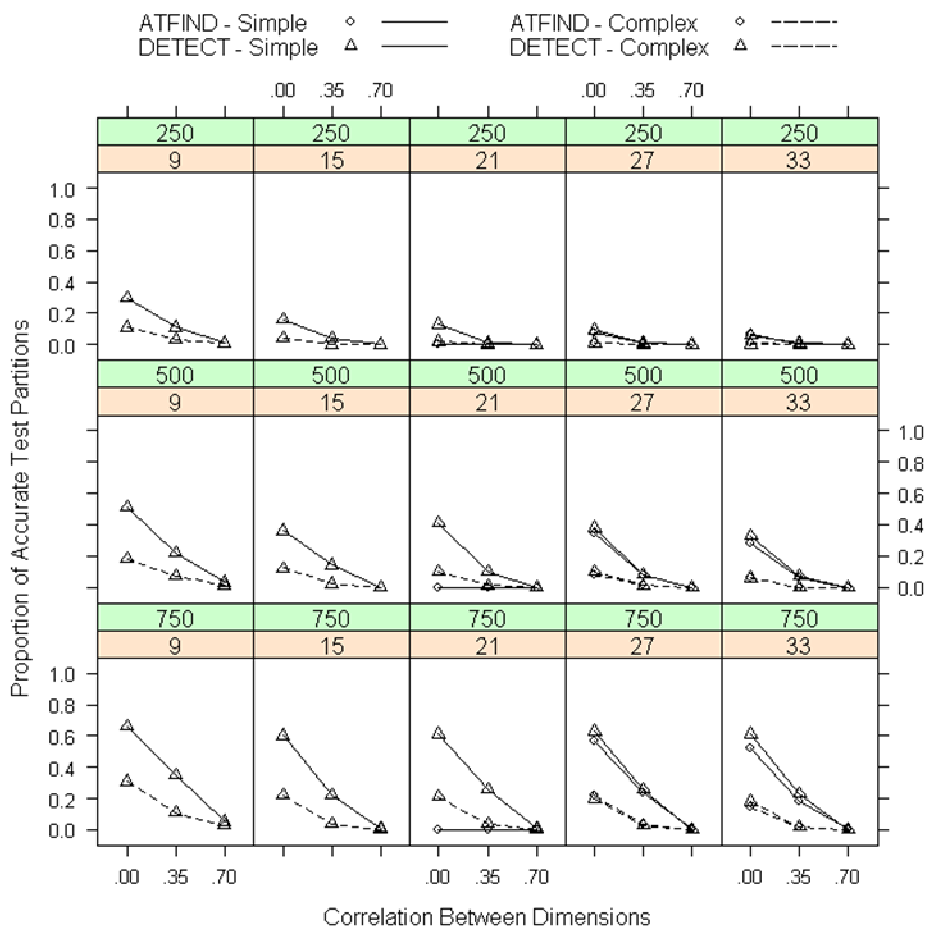


Figure 13. Proportion of accurate test partitions across 800 independent trials for multidimensional conditions with weakly discriminating items. Values in the green and yellow bars correspond to the levels of sample size and test length, respectively.

Similarity of Test Partitions. Figures 14 through 19 show the results for gauging the similarity between ATFIND- and DETECT-generated test partitions in multidimensional conditions. The axes and markers for each figure are identical to those used for Figure 10. Each figure presents the results for one of the six combinations of the strength of dependence and test structure.

Ideally, the circle marker for each level of strength of dependence within a panel would be to the far right. This result would indicate that ATFIND and DETECT suggested identical test partitions for all replications within a condition; this was not the case for most multidimensional data structures. Generally speaking, the test partitions found with ATFIND and DETECT were most likely to agree for tests consisting of 27 or more strongly discriminating items and uncorrelated dimensions. When test partitions were not identical, it was generally the case that (a) the AT found with ATFIND was a subset of DETECT or (b) the assessment subtests disagreed. In cases in which non-identical test partitions were found, disagreement became less likely to the extent that conditions became more ideal (increased sample size, test length, strength of dependence; lower correlation between dimensions; simple rather than complex test structure).

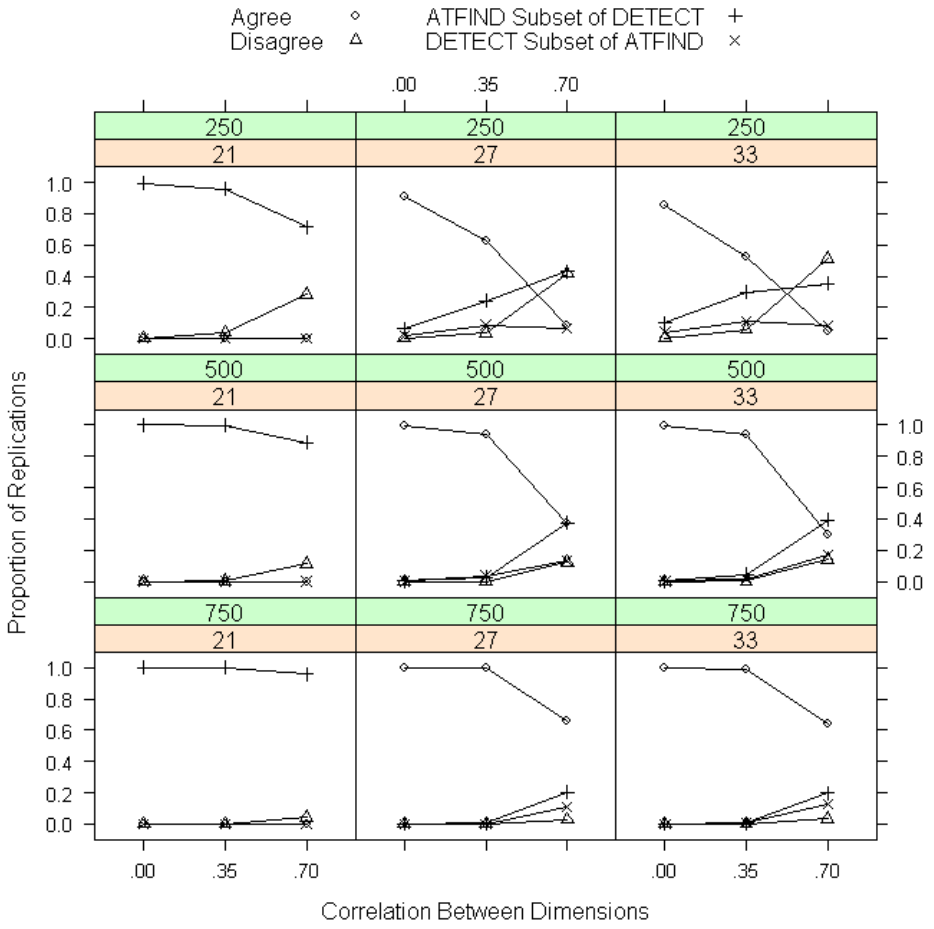


Figure 14. Comparison of ATFIND and DETECT test partitions across 800 replications for data exhibiting simple structure with strongly discriminating items in multidimensional conditions. Values in the green and yellow bars correspond to the levels of sample size and test length, respectively.

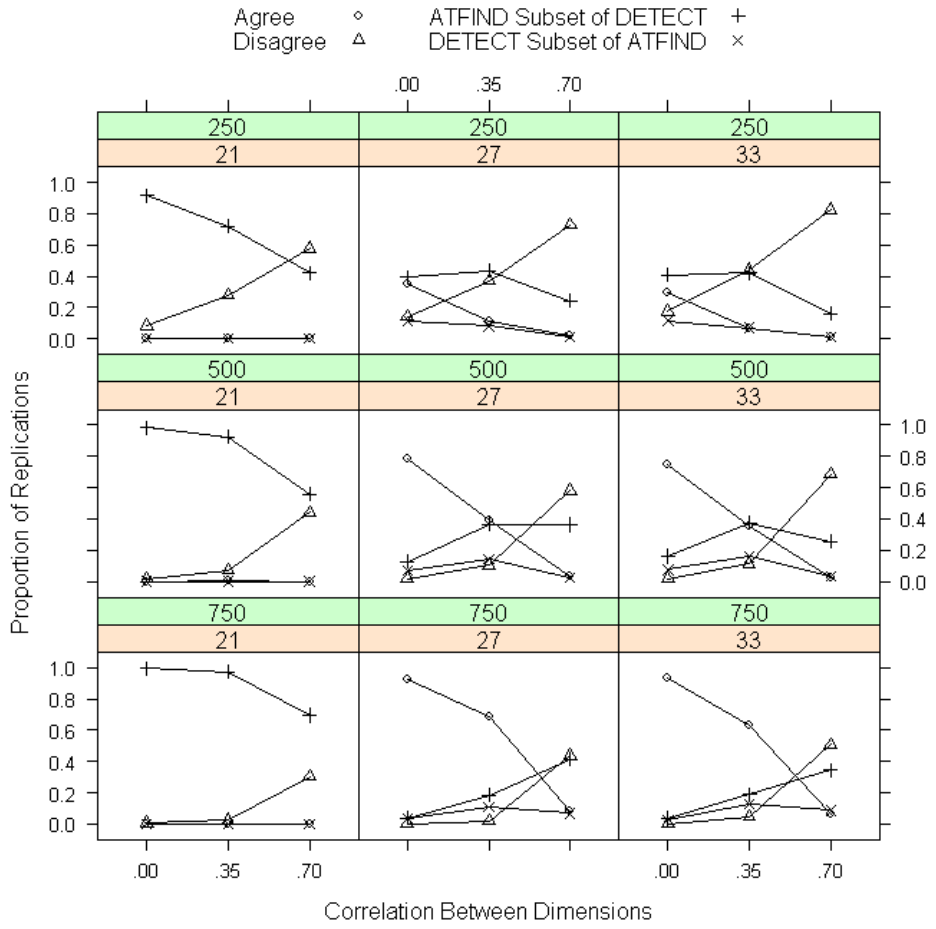


Figure 15. Comparison of ATFIND and DETECT test partitions across 800 replications for data exhibiting simple structure with moderately discriminating items in multidimensional conditions. Values in the green and yellow bars correspond to the levels of sample size and test length, respectively.

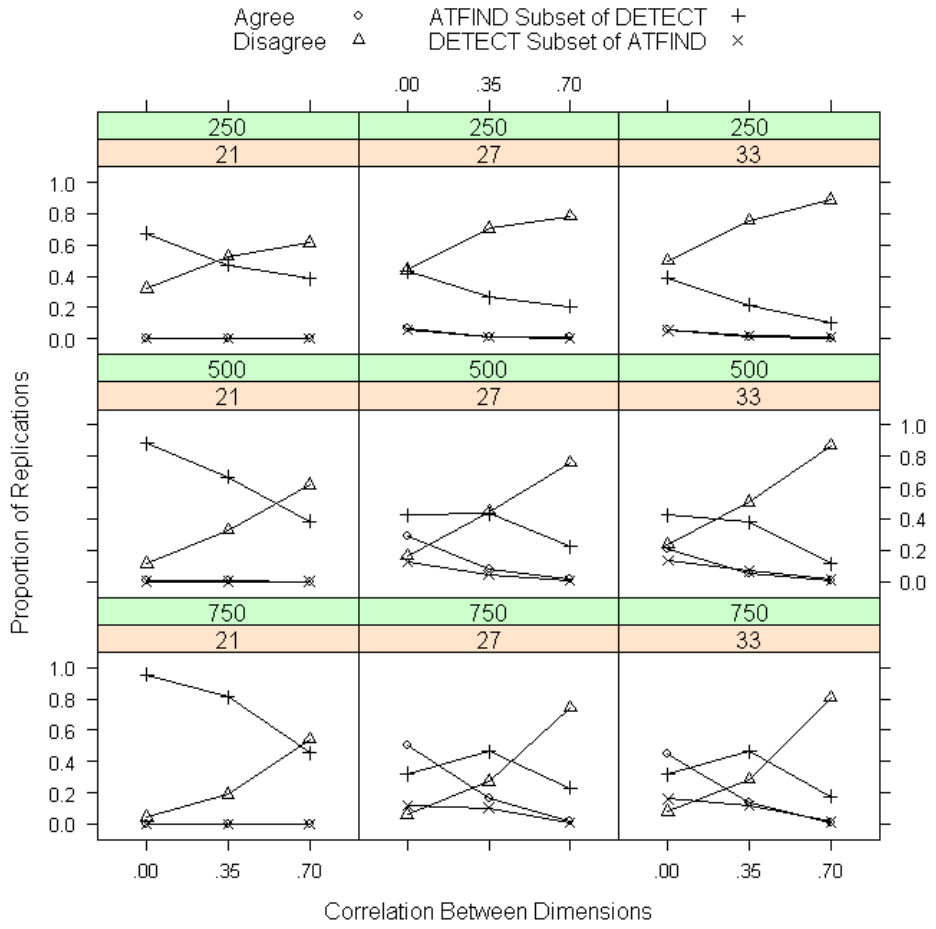


Figure 16. Comparison of ATFIND and DETECT test partitions across 800 replications for data exhibiting simple structure with weakly discriminating items in multidimensional conditions. Values in the green and yellow bars correspond to the levels of sample size and test length, respectively.

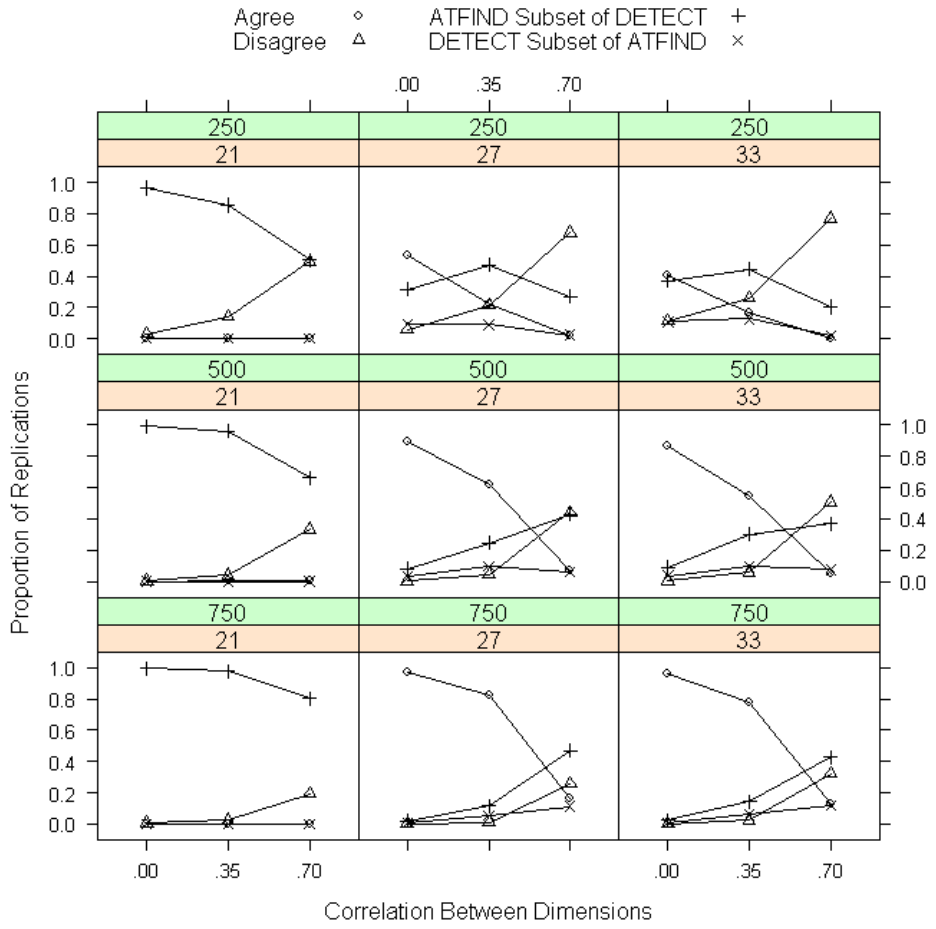


Figure 17. Comparison of ATFIND and DETECT test partitions across 800 replications for data exhibiting complex structure with strongly discriminating items in multidimensional conditions. Values in the green and yellow bars correspond to the levels of sample size and test length, respectively.

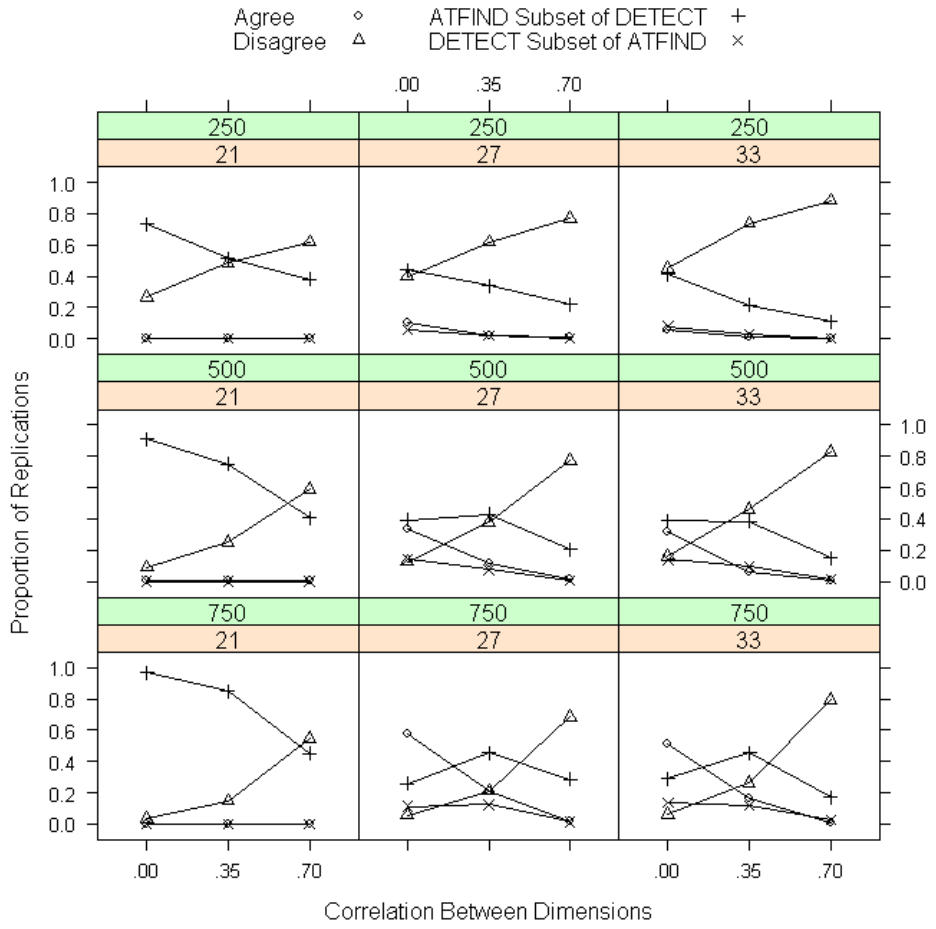


Figure 18. Comparison of ATFIND and DETECT test partitions across 800 replications for data exhibiting complex structure with moderately discriminating items in multidimensional conditions. Values in the green and yellow bars correspond to the levels of sample size and test length, respectively.

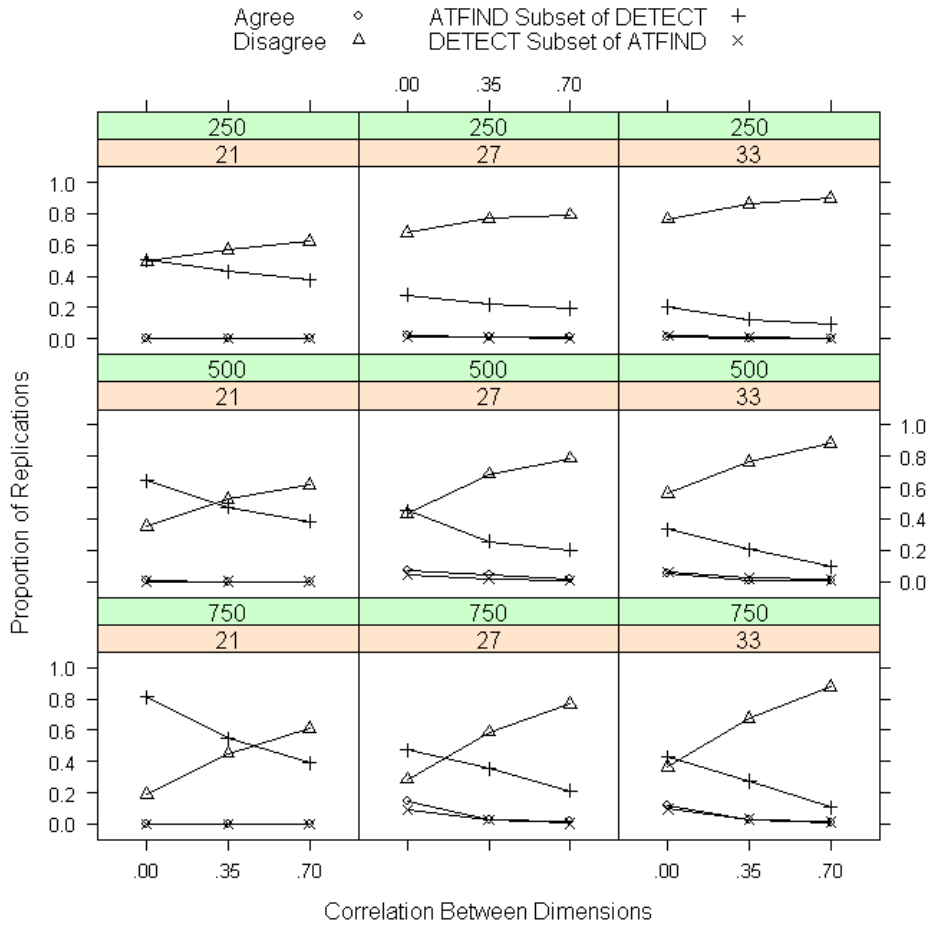


Figure 19. Comparison of ATFIND and DETECT test partitions across 800 replications for data exhibiting complex structure with weakly discriminating items in multidimensional conditions. Values in the green and yellow bars correspond to the levels of sample size and test length, respectively.

Average AT Length. The results for the average AT length in multidimensional conditions are shown in Figures 18 through 20. Each figure corresponds to one of the three levels of strength of dependence such that Figures 18, 19, and 20 present the results for test consisting of strong, moderate, and weak item discriminations, respectively. With the exception of the vertical axis, the structure of each figure is identical to those shown for presenting rejection rates and the accuracy of test partitions. Since the data were simulated, the length of the true AT (3, 5, 7, 9, 11) for each level of total test length (9, 15, 21, 27, 33) was known in multidimensional conditions. The vertical axis shows the ratio of the average observed AT length across 800 independent trials to the true AT length. By forming the ratio, the average AT length across the levels of total test length are made comparable. A ratio equal to one indicates that the typical observed AT length was equal to the true AT length; a ratio equal to 1. 5 indicates that the typical test length was longer than the true AT length by half, which is the maximum length the AT can be; and a ratio equal to . 5 indicates that the typical AT length was half of the true AT length.

A few key trends were found between the manipulated variables and the typical length of assessment subtests. First, the average observed AT length became closer to the true AT lengths dimensions became less correlated, for tests exhibiting simple rather than complex structure, and/or as the strength of dependence increased. Second, neither test length nor sample size affected the average length of assessment subtests. Third, as hinted at earlier, the length of assessment subtests found with the ATFIND program for 21 item tests were

consistently shorter than the true AT length. The typical observed AT length found with ATFIND was between five and six items while the true AT length was seven items for 21 item tests. Finally, the average AT length deviated from the true AT length for many of the multidimensional conditions such that assessment subtests were longer than the true AT length; this was particularly true assessment subtests found with DETECT.

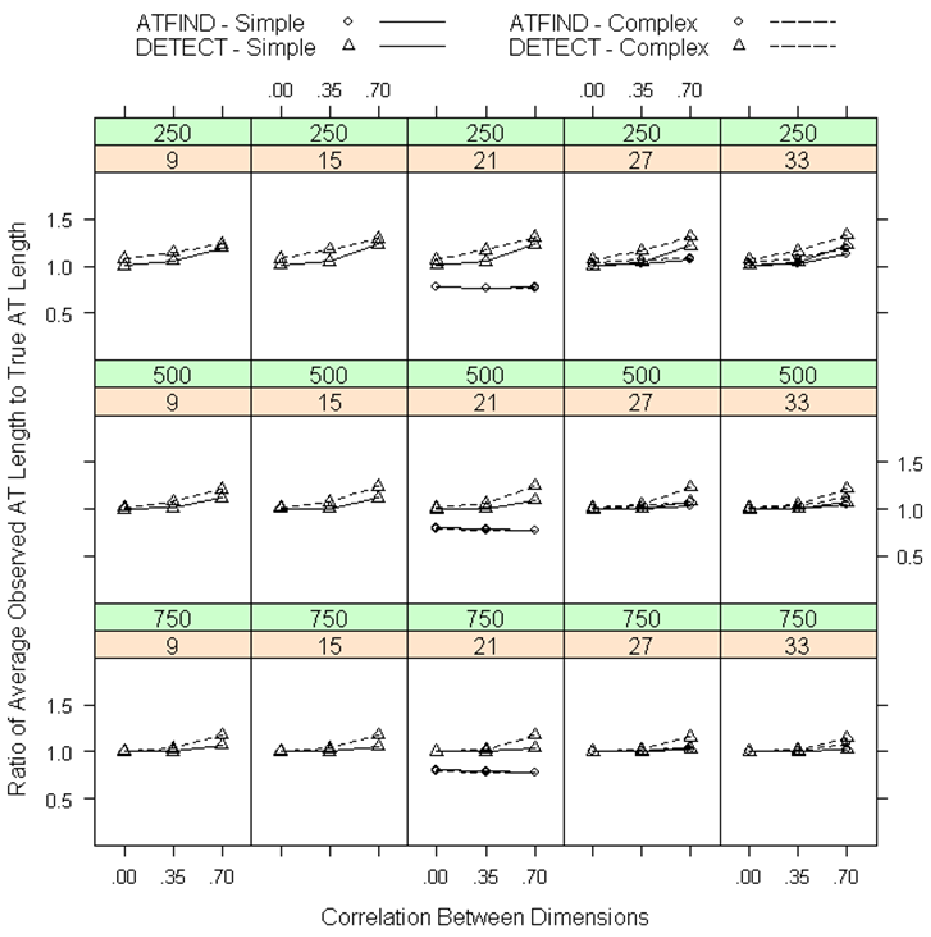


Figure 20. Ratio of the average AT length across 800 replications to the true AT length for multidimensional conditions with strongly discriminating items. Values in the green and yellow bars correspond to the levels of sample size and test length, respectively.

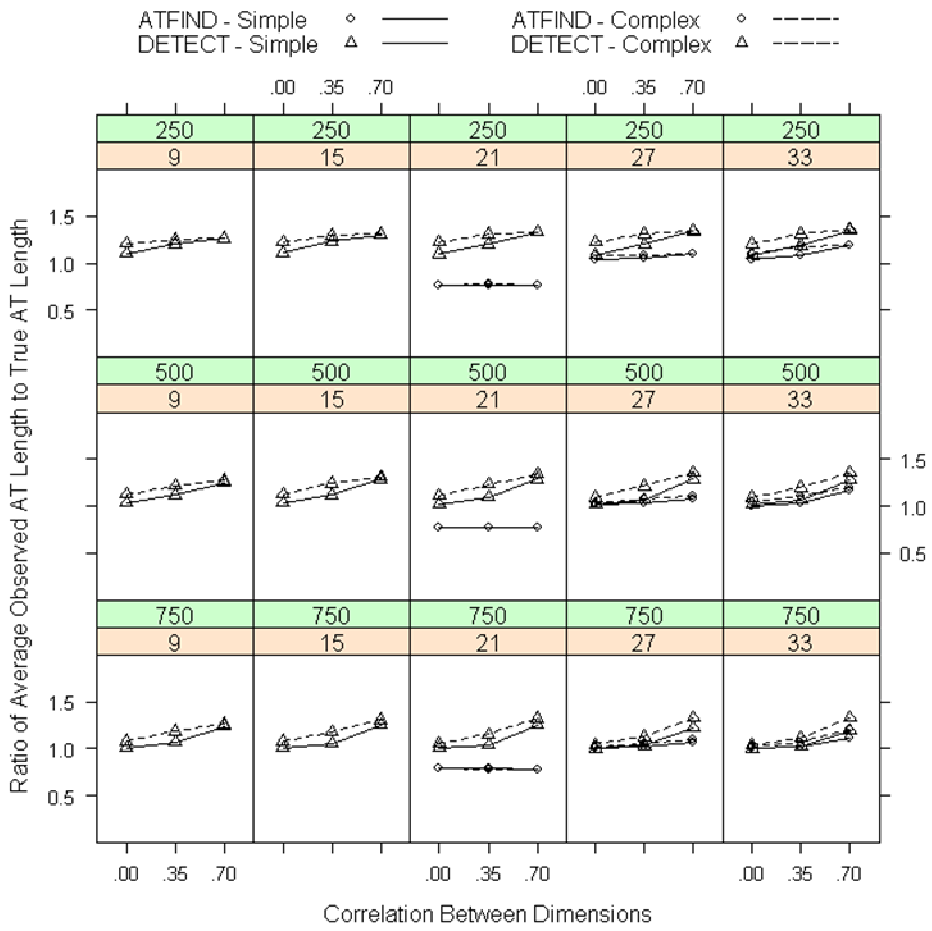


Figure 21. Ratio of the average AT length across 800 replications to the true AT length for multidimensional conditions with moderately discriminating items. Values in the green and yellow bars correspond to the levels of sample size and test length, respectively.

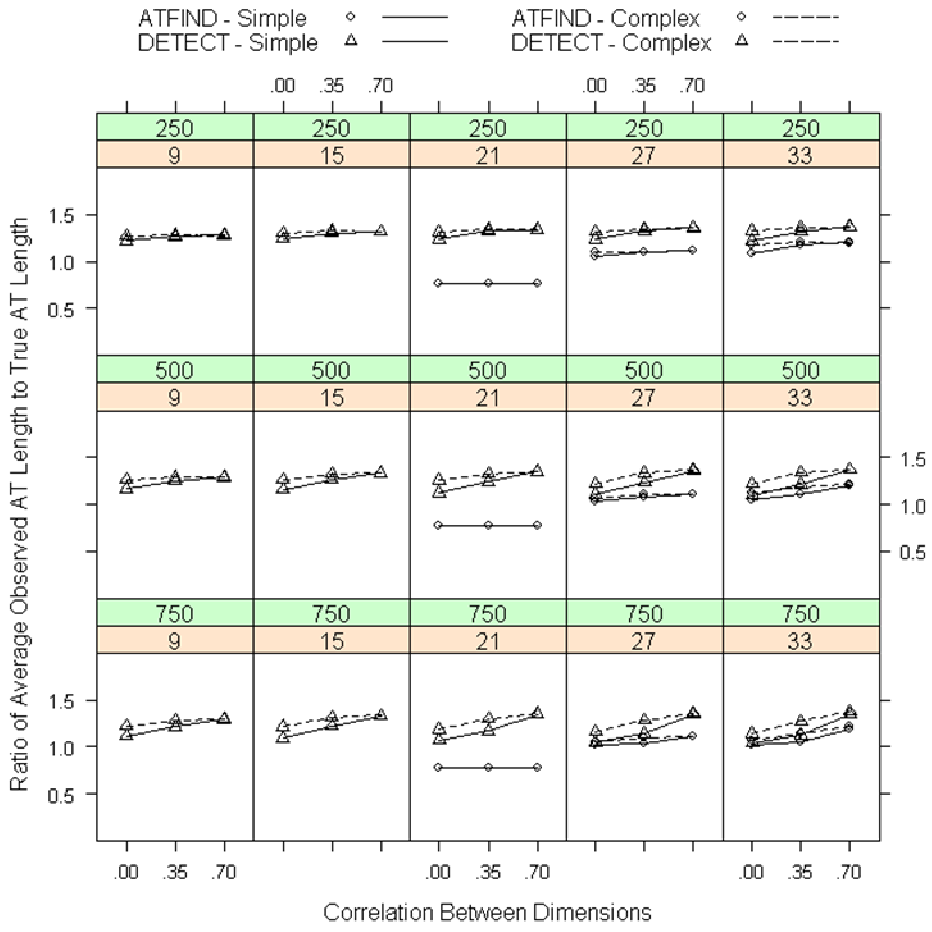


Figure 22. Ratio of the average AT length across 800 replications to the true AT length for multidimensional conditions with weakly discriminating items. Values in the green and yellow bars correspond to the levels of sample size and test length, respectively.

Summary

This chapter has presented the results for unidimensional and multidimensional conditions. The estimation of Type I error rates and power in unidimensional and multidimensional conditions, respectively, served as the

primary focus of this work. In general, the Type I error rate with DIMTEST tended to be conservative for most unidimensional conditions. Type I error rates approximately doubled, however, with either 500 or 750 examinees responding to nine weakly discriminating items. For multidimensional conditions, the manipulated variables exerted main effects and also combined interactively to influence the performance of DIMTEST. Put simply, the power of DIMTEST tended to improve as conditions became more ideal (larger sample size, longer tests, simple structure, less correlated dimensions, more discriminating items). To briefly summarize the interactive effects, *combining* a larger sample size, longer tests, less correlated dimensions, and more discriminating in generally resulted in high power with the exception of ceiling and floor that were observed for some conditions.

The secondary focus of this work was to compare the test partitions obtained the two conditional covariance-based subtest selection methods, ATFIND and DETECT. Test partitions were evaluated in terms of accuracy and test length, and were directly compared to assess their similarity between methods. In unidimensional conditions, assessment subtests found with DETECT tended to be longer, on average, than those found with ATFIND. Accordingly, the assessment subtests tended to disagree or the AT found with ATFIND was a proper subset of the AT found with DETECT; disagreement was observed more often for longer tests. In general, ATFIND and DETECT produced accurate test partitions, were in agreement more often, and more closely approximated the true

AT length in multidimensional conditions to the extent that conditions became more ideal.

Chapter 5

DISCUSSION AND CONCLUSIONS

Each of manipulated variables had meaningful impact on the rejection rates and the structure of test partitions. Key findings pertaining to these outcomes are separately discussed for unidimensional and multidimensional conditions. In light of the results, sample size and test length recommendations are given for the current DIMTEST procedure. Finally, concluding statements, limitations of the study, and suggestions for future work are provided.

Unidimensional Conditions

Type I Error Rates. It was hypothesized for unidimensional conditions that none of the manipulated variables would systematically impact rejection rates; this hypothesis was generally supported (see Figure 5 and Appendix B). Irrespective of subtest selection method, Type I error rates were consistently below the nominal rate of $\alpha = .05$ for tests consisting of 21 or more items. However, Type I error rates were systematically affected by the manipulated variables for shorter tests. Specifically, Type I error rates were inflated for tests comprised of nine weakly discriminating items, particularly to the extent that sample size was increased. Holding all else constant, Type I error rates were reduced for nine item tests by increasing the strength of dependence. Finch and Habing (2007) found that Type I error rates became inflated with DIMTEST by increasing the sample size from 1000 to 2000 examinees for short tests. The current study extends their work by showing that the role of sample size on Type I error rates for short test may be moderated by the strength of dependence.

In general, rejections rates were lower than the nominal Type I error rate of $\alpha = .05$ suggesting that DIMTEST is a conservative tests of the null hypothesis of essential unidimensionality for the sample sizes and test lengths considered. On one hand, this indicates that an analyst is quite unlikely to reject the null hypothesis with DIMTEST for tests that are indeed best characterized by one dimension. On the other hand, the empirical Type I error rate did not approach specified rate with increases in test length and/or sample size, which is an asymptotic relationship expected from theory (Stout, 1987).

Structure of Test Partitions. With no existing work to draw from, hypotheses were not made about the relationship between the manipulated variables and the structure of test partitions; however, the characteristics of test partitions were explored. Test partitions in unidimensional conditions were explored by (a) calculating the average length of assessment subtests across the 800 replications for each condition, and (b) directly comparing the assessment subtests obtained with ATFIND and DETECT. It was found that the manipulated variables had almost no effect on the average AT length (see Figure 9 and Appendix C). However, it was found that the average AT lengths obtained with ATFIND were consistently shorter than those found with DETECT. In turn, it was found that test partitions tended to disagree or ATFIND produced an AT that was a proper subset of the AT found with DETECT (see Figure 10). Disagreement became more likely to the extent that test length was increased. This finding may be expected in that increasing the test length yields a greater chance that at least

one item will differbetween the assessment subtests generated with any two subtest selection methods.

Multidimensional Conditions

Power. In support of the hypotheses for this study, finding evidence for multidimensionality was easier with larger samples, longer tests, stronger strength of dependence, less correlated dimensions, and for test exhibiting simple rather than complex structure (see Figures 6 – 8). These findings were also consistent with previous research on DIMTEST (Nandakumar & Stout, 1993; Finch & Habing, 2007; Froelich & Habing, 2008; Froelich & Stout, 2003) and Zhang and Stout's (1999a) conditional covariance theory. The main effects, however, were not consistent across the levels of the other manipulated variables.

To highlight the general nature of the interactive effects, consider increasing the test length for conditions with dimensions correlated at $\rho = .7$. In the weak strength of dependence conditions, power was consistently low across all levels of test length irrespective sample size and test structure. The same was true for tests with complex structure and moderately discriminating items. In contrast, for tests following a simple structure model with at least 500 examinees, the low levels of power were offset by increasing test length. In general, increasing the test length for tests following a simple structure model had less of an effect with dimensions correlated at $\rho = .7$ on account of already high power.

Structure of Test Partitions. Three aspects of the test partitions were explored in multidimensional conditions: accuracy, the similarity of assessment subtests obtained with ATFIND and DETECT, and the average AT length. With

respect to the accuracy of test partitions, the trends in accuracy (see Figures 11 – 13) were similar to the trends seen with rejection rates in multidimensional conditions. Specifically, test partitions were more likely to be accurate with (a) increased sample size, (b) less correlated dimensions, (c) a stronger strength of dependence, and (d) for tests following a simple rather complex structure model. The relationship between test length and accuracy, however, differed in two ways from the trends seen with rejection rates. First, test partitions were slightly less likely to be accurate with increases in test length. One possible explanation is that the *potential* for assigning at least one item to the wrong cluster becomes greater for longer tests. Second, test partitions obtained with ATFIND were never accurate for 21 item tests. This finding highlights a nuance in the ATFIND program that is not implemented in the DETECT program; the AT must consist of at least four items, and the PT must consist of at least 15 items. Since the true AT in 21 item conditions consisted of seven items, the true PT of 14 items was shorter than allowed by the ATFIND program. It was therefore impossible for the ATFIND program to produce accurate test partitions based on the definition of accuracy used here.

Distinct patterns were observed when directly comparing the assessment subtests found with ATFIND and DETECT directly (see Figures 14-19). First, the results obtained for 21 items differed markedly from 27 and 33 item tests. For 21 item tests, the assessment subtests obtained with DETECT were never observed to agree with those found with ATFIND or be a subset of ATFIND assessment subtests. Rather, ATFIND assessment subtests were more likely to be a subset of

DETECT assessment subtests with (a) increased sample size, (b) less correlated dimensions, (c) a stronger strength of dependence, and (d) for tests following a simple rather than complex structure model. The results for 27 and 33 item tests were quite similar to each other. Strict agreement between ATFIND and DETECT assessment subtests became more likely with (a) increased sample size, (b) less correlated dimensions, (c) a stronger strength of dependence, and (d) for tests following a simple rather than complex structure model. When the assessment subtests between methods did not agree, the patterns were somewhat more mixed for 27 and 33 item tests. However, as was the case for 21 item tests, the most likely scenarios were that the assessment subtests obtained with ATFIND either disagreed with or were subset of the DETECT assessment subtests.

The relationships pertaining to the accuracy of test partitions and the agreement between ATFIND and DETECT assessment subtests were strongly tied to the typical length of assessment subtests (see Figures 20 – 22). Across all levels of strength of dependence, they eye is drawn to the lines showing the results for ATFIND in 21 item conditions, which are the only lines that show an average AT length lower than the true AT length. This finding reflects the requirements placed on the length of the two subtests by the ATFIND program. The remaining results indicate that the average observed AT length became closer to the true AT length with (a) increased sample size, (b) a stronger strength of dependence, (c) less correlated dimensions, and (d) for tests following a simple rather complex structure model. Test length had little effect on the relative length of observed assessment subtests to the true assessment subtest length. These

trends shed light on the accuracy of test partitions in that accuracy tended to diminish on account of ATFIND and DETECT including more items than necessary on the assessment subtests. It was also observed that DETECT tended to produce slightly longer assessment subtests than ATFIND, on average, particularly with an increased correlation between the dimensions and/or decreased strength of dependence. In these conditions, ATFIND tended to produce assessment subtests that were a subset of those found with DETECT or the assessment subtests disagreed. To the extent that the average length of ATFIND and DETECT-generated subtests became more similar, the assessment subtests were more likely to agree.

Relating the Structure of Test Partitions to DIMTEST Performance. For true multidimensional tests, the performance of DIMTEST depends on the utility of the test partitioning method to maximize dimensional homogeneity within subtests and dimensional heterogeneity between subtests. The maximization of dimensional homogeneity and heterogeneity was captured by the operationalization of test partition accuracy. In general, the observed trends in power were also observed with the proportion of perfect test partitions, or accuracy; as the proportion of accurate test partitions increased, so did power. The relationship between accuracy and power, however, was not as strong as intuition may suggest. In particular, gains in accuracy were did not perfectly mirror gains in power; the latter were generally greater than the former.

The discrepancy between power and accuracy is an issue of the robustness of DIMTEST in multidimensional conditions to inaccurate test partitions. In the

context of DIMTEST, robustness refers to the degree by which a test partition can be inaccurate and still yield a correct result. As robustness was not the focus of this work, it was not investigated directly. However, key relationships in the average AT length, accuracy, and power provide some insight into the issue of robustness.

One key finding was that DIMTEST was similarly powered with ATFIND and DETECT for 21 item test even though ATFIND always produced assessment subtests that were shorter than the true AT length, and were therefore never accurate. Restrictions on the length of partitioning subtests with ATFIND ($J_{PT} \geq 15$) forced assessment subtests to be shorter than the true AT length. Despite this restriction, the differences in the power of DIMTEST obtained with ATFIND and DETECT were substantially smaller than the differences seen with respect to the accuracy of test partitions. At the extreme, DIMTEST rejected the null hypothesis nearly 100% of the time with ATFIND for some conditions even though the test partitions submitted for testing were never perfectly accurate.

Instances of high power despite lacking accuracy of test partitions were seen in other conditions as well. Although somewhat restrictive, conditions with a strong strength of dependence, complex test structure, and dimensions correlated at $\rho = .7$ are discussed for illustration. For these conditions, the average length of assessment subtests were consistently longer than the true subtests, and in turn, resulted in diminished proportions of accurate test partitions. Moreover, a slight decline in accuracy was observed with increases in test length. Despite the low levels of accuracy, the power of DIMTEST improved with increases in test length

and sample size. Taken together, these results indicate that DIMTEST became more robust to inaccurate test partitions with increases in sample size and test length. Similar trends were also seen with uncorrelated and moderately correlated dimensions with weakly and moderately discriminating items.

Sample Size and Test Length Requirements

The primary goal of this work was to update the sample size and test length requirements for the current DIMTEST procedure. Previous research has suggested that DIMTEST can be expected to maintain the Type I error rate at or below the nominal rate without sacrificing power with a minimum of 750 examinees and 25 items (Gessoroli and De Champlain, 1996; Stout, 1987). Drawing on the results from this work, it is argued that the selection of sample size and test length minima for DIMTEST depends on the approach used for making recommendations.

The approach for making recommendations in previous research on DIMTEST was to select the lowest combination of sample size and test length that yielded sufficiently low Type I error rates without sacrificing power. Given that DIMTEST is intended for use prior to selecting a unidimensional IRT model, the analyst is unlikely to know much about the data at hand beyond the sample size and test length. The results in this study indicates that unknown features of the data such as the strength of dependence, the correlation between dimensions, and test structure strongly impacted the performance of DIMTEST. Since these features of the data are most likely unknown to the analyst, one reasonable solution is to recommend a combination of sample size and test length that yields

the lowest opportunity for making an incorrect inference about the dimensionality of a test based on the DIMTEST analysis. From this perspective, the recommended combination of sample size and test length based on the current work would be 750 examinees and 33 items. For this sample size and test length combination, Type I error rates were low and power was reasonably high with moderately correlated dimensions and weakly discriminating items.

The above approach assumes a passive analyst and may unnecessarily detract those with shorter tests and/or smaller samples from using DIMTEST. However, DIMTEST maintained control of the Type I error rate and was reasonably powered for some conditions with as few as 250 examinees and nine items. A more flexible approach to offering recommendations assumes an active analyst who has gained some understanding about the data at hand.

While features of a test such as the strength of dependence of items, test structure, and the correlation between dimensions cannot be known, analysts can gain some insight about these features of the data at a hand via classical test theory indices and subject matter expertise (when available). For example, the strength of dependence between items and the latent variable(s) can be approximated via item-total score correlation coefficients. Two features of the item-total score correlation coefficients can be considered in light of the current work: (a) the strength of the coefficients and (b) the pattern of the coefficients. Higher coefficients indicate a stronger strength of dependence and potentially a more powerful test of the unidimensionality assumption. The pattern of coefficients can suggest (a) unidimensionality if all items are highly related to the

total score or (b) multidimensionality if some items are not strongly related to the total score. Subject matter expertise can also be called on to gain insight about the expectations with respect to test structure in terms of (a) the number of latent variables and (b) which items are associated with the latent variable(s). If multiple dimensions are suspected, the correlation between dimensions can be estimated via the correlation between the total scores for the expected dimensions. The key point is that analysts can evaluate how well the features of real data align with the features of the simulated data in this work (and other work) and decide whether DIMTEST would yield a trustworthy result given their sample size and test length.

Conclusions

This research extends the existing research on the current DIMTEST procedure and provides avenues for additional research. Central to this work was the belief that recent chances to the DIMTEST statistic would require smaller samples and fewer items than research on earlier versions has suggested. The results indicate that sample size and test length requirements are tied to the features of the data such as the strength of discrimination parameters, and in multidimensional conditions, the correlation between dimensions and the test structure. Given that the range of sample size and test length requirements varied for different data structures, it is recommended that analysts become familiar with the data at a descriptive level via classical test theory statistics and subject matter expertise (when possible) prior to conducting DIMTEST.

To date, this work is the first to investigate the performance of DIMTEST using Zhang and Stout's (1999b) DETECT program with the genetic algorithm (the ATFIND program uses DETECT, but does not employ the genetic algorithm). Although the performance of DIMTEST with ATFIND and DETECT was similar for most conditions, DETECT yielded non-trivially higher power than ATFIND for some multidimensional conditions. Moreover, it was possible to use DETECT to obtain subtests with fewer items than ATFIND. These features call for additional research on the utility of DETECT for obtaining the required test partition for DIMTEST

The relationship between the performance of and the structure of test partitions obtained with ATFIND, DETECT, and possibly other methods deserves further attention, particularly for multidimensional conditions. In general, rejection rates tended to be larger than the proportion of accurate test partitions; this indicates that DIMTEST is robust to imperfect partitions. Additional research is necessary for (a) understanding factors that influence the structure of test partitions and (b) gauging the degree of departure of the AT/PT test partitioning from the true test structure than can be withstood before DIMTEST no longer performs effectively.

REFERENCES

Birnbaum, A. (1968). Some latent trait models and heir use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 392-479). Reading, Mass.: Addison-Wesley.

Florek, K., Lukaszewicz, J., Perkal, J. Steinhaus, H., & Zubrzycki, S. (1951). Sur la liaison et la division des points d'un ensemble fini. *Colloquium Mathematicum, 2,*282-285.

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 responses to a set of items. *Journal of Educational Measurement*, *33*, 157-179.

Hattie, J., Krakowski, K., Rogers, J., & Swaminathan, H. (1996). An assessment of Stout's index of essential unidimensionality. *Applied Psychological Measurement*, *20*, 1- 14.

Ip, E. H. (2000). Adjust for information inflation due to local dependency in moderately large item clusters. *Psychometrika*, *65¸*79-91.

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

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

McDonald, R. P. (1994). Testing for approximate dimensionality. In D. Laveault, B. Zumbo, M. Gessaroli, & M. Boss (Eds.), *Modern theories of measurement: Problems and issues* (pp. 63-85). Ottawa, Canada: University of Ottawa Press.

McQuitty, L. L. (1960). Hierarchical linkage analysis for the isolation of types. *Educational and Psychological Measurement*, *20*, 55-67.

Nandakumar, R., & Stout, W. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational and Behavioral Statistics*, *18*, 41-68.

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

Pyo, K. H. (2000, April). *Assessing dimensionality of a set of language test data.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-9000051-07-0, URL http://www. R-project. org/.

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

Reckase, M. D. (1997). A linear logistic 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).

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. F., &Marden, J. I. (1998). Using new proximity measures with hierarchical cluster analysis to detect multidimensionality. *Journal of Educational Measurement*, *35*, 1-30.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guildford Press.

Seraphine, A. E. (2000). The performance of DIMTEST when latent trait and item difficulty distributions differ. *Applied Psychological Measurement*, *24*, 92 -94.

Sokal, R. R., & Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Bulletin*, *38*, 1409-1438.

Stout, W. F., Habing, B., Douglas, J., Kim, H. R., Roussos, L., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement*, *19*, 331-354.

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

Stout, W. F. (1990). A new item response theory modeling approach with applications in unidimensionality assessment and ability estimation. *Psychometrika*, *55*, 293-326.

Stout, W. F., Froelich, A. G., & Gao, F. (2001). Using resampling 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. 357-376). New York: Springer-Verlag.

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

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12*, 58-79.

Yen, W. M. (1993). Scaling performance assessments: strategies for managing local item dependence. *Journal of Educational Measurement, 30,* 187 -213.

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

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

APPENDIX A

DATA GENERATION CODE

```
#####################################################################
## Code for binary generating IRT data.
#####################################################################

#SPECIFY THE SAMPLE SIZE AND TEST LENGTH.
#SPECIFY UNIDIMENSIONAL OR MULTIDIMENSIONAL GENERATION
(1=unidim, 2=multidim).
#SPECIFY THE STRUCTURE OF MIRT DATA (1=simple, 2=appss).
#SPECIFY THE NUMBER OF REPLICATIONS.
#SPECIFY ALL SIMULATION SETTINGS.

I = 750
exam_prop = 2
J = 33
type=2
struct_type = "simple"
r="modcor" #nocor, modcor, or highcor
alevel="high" #low, moderate, or high
nreps=800

#DIRECTORIES.
condition = paste(I, "N", J, "J", struct_type, r, alevel, sep="_")
gen. dir = paste("C:\\Documents and Settings\\Derek\\Desktop\\", condition, "\\",
sep="")
#gen. dir = paste("F:\\Users\\Derek\\Desktop\\", condition, "\\", sep="")
#gen. dir = paste("I:\\", condition, "\\", sep="")
dir. create(gen. dir)
dir. create(paste(gen. dir, "items\\", sep=""))
dir. create(paste(gen. dir, "thetas\\", sep=""))
dir. create(paste(gen. dir, "data\\", sep=""))
dir. create(paste(gen. dir, "figures\\", sep=""))
item_fold=paste(gen. dir, "items\\", sep="")
data_fold=paste(gen. dir, "data\\", sep="")
theta_fold=paste(gen. dir, "thetas\\", sep="")
figure_fold=paste(gen. dir, "figures\\", sep="")
setwd(gen. dir)
```

```
# SPECIFY THE DISTRIBUTION OF ITEM AND PERSON PARAMETERS.
# LOAD NECESSARY PACKAGES.
#difficulty parameters.
if (J==9)dp = as. matrix(c(rep(c(-. 75,0,. 75),3)),ncol=1,nrow=J)
if (J==15) dp = as. matrix(c(rep(c(-1. 5,-. 75,0,. 75,1. 5),3)),ncol=1,nrow=J)
if (J==21) dp = as. matrix(c(rep(c(-1. 5,-1,-. 5,0,. 5,1,1. 5),3)),ncol=1,nrow=J)
if (J==27) dp = as. matrix(c(rep(c(-1. 5,-1. 13, -. 75, . 38,0,. 38,. 75,1. 13,1.
50),3)),ncol=1,nrow=J)
if (J==33) dp = as. matrix(c(rep(c(-1. 5,-1. 2,-. 9, -. 6, . 3,0,. 3,. 6,. 9,1. 2,1.
5),3)),ncol=1,nrow=J)

# correlations.
if (r=="nocor") r=0
if (r=="modcor") r=. 35
if (r=="highcor") r=. 70

library(MASS)
library(mvtnorm)

#GENERATE ABILITY ESTIMATES and DATA.
at_data = matrix(NA,at. I,J) #SET UP AT DATA MATRIX.
dim_data = matrix(NA,dim. I,J) #SET UP DIMTEST DATA MATRIX.

system. time( #start the clock for timing the generation of data.

for(which. rep in 1:nreps){

#generate unidimensional discrimination parameters.
if (alevel=="low") apars = as. matrix(runif(J, min=. 5, max=1))
if (alevel=="moderate") apars = as. matrix(runif(J, min=. 75, max=1. 25))
if (alevel=="high") apars = as. matrix(runif(J, min=1. 25, max=1. 75))

#manipulate the structure of item parameters.
 if (struct_type=="simple") structure=as. matrix(c(runif((round((2/3)*J)),min=0,
max=0), runif((round((1/3)*J)),min=(90*(pi/180)), max=(90*(pi/180)))))
 if (struct_type=="apss") structure=as.
matrix(c(runif((round((2/3)*J)),min=(0*(pi/180)), max=(20*(pi/180))),
runif((round((1/3)*J)),min=(70*(pi/180)), max=(90*(pi/180)))))
 if (struct_type=="unidim") structure=as.
matrix(c(runif((round((2/3)*J)),min=(0*(pi/180)), max=(0*(pi/180))),
runif((round((1/3)*J)),min=(0*(pi/180)), max=(0*(pi/180)))))
```

```
#pseudo-guessing parameter.
cp = as. matrix(runif(J, min=0, max=0))

#GENERATE DISCRIMINATION PARAMETERS TO BE
MULTIDIMENSIONAL IF TYPE==2.

#APPROXIMATELY 1/3 OF THE ITEMS MEASURE THE SECOND
DIMENSION, AND 2/3 MEASURE THE FIRST DIMENSION.
ap1=as. matrix(apars*cos(structure),ncol=type)
ap2=as. matrix(apars*sin(structure),ncol=type)
ap = as. matrix(cbind(ap1,ap2), ncol=type, nrow=J)
ap = as. matrix(ap, nrow=J, ncol=2)
apf = as. matrix(((apars/1. 7)/(sqrt(1+((apars/1. 7)*(apars/1. 7))))), ncol=1,
nrow=J) #unidimensional standardized factor loading - calculation based on Wirth
& Edwards, 2007.
ap1f = as. matrix(((ap1/1. 7)/(sqrt(1+((ap1/1. 7)*(ap1/1. 7))))), ncol=1,
nrow=J)#dimension 1 standardized factor loading - calculation based on Wirth &
Edwards, 2007.
ap2f = as. matrix(((ap2/1. 7)/(sqrt(1+((ap2/1. 7)*(ap2/1. 7))))), ncol=1,
nrow=J)#dimension 2 standardized factor loading - calculation based on Wirth &
Edwards, 2007.

#generate test structure plots.
plot(ap1, ap2, type="n", xlim=c(0,2), ylim=c(0,2), xlab=" ", ylab=" ")
arrows(0, 0, ap1, ap2, code=2, length=. 05)
savePlot(paste(figure_fold, "structure_", which. rep, sep=""), type="png")

#WRITE OUT GENERATED PARAMETERS.
item. id = matrix(seq(from=1, to=J, by=1), ncol=1, nrow=J)
structure. angle = matrix(structure*180)/pi
items = matrix(cbind(item. id, apars, apf, ap, ap1f, ap2f, structure, structure.
angle, dp, cp), ncol=11, nrow=J)
items = matrix(round(items,digits=2), ncol=11, nrow=J)
write. table(items, paste(item_fold, "items_", which. rep, ". dat", sep=""),
col. names=c("item", "apars", "apf", "ap1", "ap2", "ap1f", "ap2f", "beta", "angle",
"dp", "cp"),
row. names=F,
sep="\t",
quote=F)
```

```
at_theta = as. matrix(rnorm(at. I, mean=0, sd=1), ncol=1, nrow=at. I)
dim_theta = as. matrix(rnorm(dim. I, mean=0, sd=1), ncol=1, nrow=dim. I)
sigma=matrix(c(1, r, r, 1),ncol(ap))
mu=rep(0, ncol(as. matrix(ap)))
at_theta = as. matrix(mvrnorm(at. I, mu, sigma))
dim_theta = as. matrix(mvrnorm(dim. I, mu, sigma))

#write out separate theta files for the at and dimtest files.
write. table(at_theta,
paste(theta_fold, "at_thetas_", which. rep, ". dat", sep=""),
row. names=T,
col. names=F,
quote=F,sep="\t")
write. table(dim_theta,
paste(theta_fold, "dim_thetas_", which. rep, ". dat", sep=""),
row. names=T,
col. names=F,
quote=F,
sep="\t")


for(a in 1:at. I){
for(j in 1:J){
 #p_at=cp[j]+((1-
cp[j])*(exp((t(ap[j,1])%*%at_theta[a,1]+t(ap[j,2])%*%at_theta[a,2])+dp[j]))/(1+(
exp((t(ap[j,1])%*%at_theta[a,1]+t(ap[j,2])%*%at_theta[a,2])+dp[j])))
          at_num =
(exp((t(ap[j,1])%*%at_theta[a,1]+t(ap[j,2])%*%at_theta[a,2])+dp[j]))
          at_den = 1 + at_num
          p_at=(cp[j] + ((1-cp[j])*((at_num)/(at_den))))
uni_at = runif(1)
if (p_at>uni_at) at_data[a,j]=1
if (p_at<uni_at) at_data[a,j]=0
} #closes MIRT AT item loop.
} #closes MIRT AT person loop.
for(d in 1:dim. I){
for(j in 1:J){
 #p_dim=cp[j]+((1-
cp[j])*(exp((t(ap[j,1])%*%dim_theta[d,1]+t(ap[j,2])%*%dim_theta[d,2])+dp[j]))/
(1+(exp((t(ap[j,1])%*%dim_theta[d,1]+t(ap[j,2])%*%dim_theta[d,2])+dp[j])))
          dim_num =
(exp((t(ap[j,1])%*%dim_theta[d,1]+t(ap[j,2])%*%dim_theta[d,2])+dp[j]))
          dim_den = 1 + dim_num
          p_dim=(cp[j] + ((1-cp[j])*((dim_num)/(dim_den))))
uni_dim = runif(1)
```

```
if (p_dim>uni_dim) dim_data[d,j]=1
if (p_dim<uni_dim) dim_data[d,j]=0
} #closes MIRT DIMTEST item loop.
} #closes MIRT DIMTEST person loop.

#write out data files for the AT and DIMTEST.
write. table(at_data,
paste(data_fold,"at_data_", which. rep, ". dat", sep=""),
row. names=F,
col. names=F,
quote=F,
sep="")

write. table(dim_data,
paste(data_fold, "dim_data_", which. rep, ". dat", sep=""),
row. names=F,
col. names=F,
quote=F,
sep="")

} #close nreps loop.

) #close system time function.
```

APPENDIX B

TABLE OF TYPE 1 ERROR RATES

Proportion of d_E = 1 Rejections in 800 Replications for Unidimensional Conditions

		$J = 9$	$J = 15$	$J = 21$		$J = 27$		$J = 33$	
N	SoD	D	D	A	D	A	D	A	D
250	Weak	.068	.015	.030	.028	.024	.015	.013	.023
	Moderate	.043	.026	.024	.015	.020	.011	.019	.021
	Strong	.020	.016	.019	.019	.006	.010	.015	.021
500	Weak	.101	.023	.029	.030	.016	.020	.021	.018
	Moderate	.054	.021	.026	.024	.018	.014	.015	.014
	Strong	.016	.015	.014	.018	.020	.013	.014	.015
750	Weak	.113	.053	.036	.040	.028	.026	.025	.024
	Moderate	.066	.044	.023	.033	.024	.030	.014	.024
	Strong	.020	.015	.024	.008	.015	.011	.021	.013

Note. SoD = strength of dependence. A = ATFIND. D = DETECT.

APPENDIX C

TABLE OF AVERAGE AT LENGTH IN UNIDIMENSIONAL CONDITIONS

Average AT Length across 800 Replications for Unidimensional Conditions

		$J = 9$	$J = 15$	$J = 21$		$J = 27$		$J = 33$	
N	SoD	D	D	A	D	A	D	A	D
250	Weak	3.85	6.62	6.00	9.49	9.91	12.22	13.16	15.02
	Moderate	3.81	6.60	5.36	9.34	9.99	12.24	13.17	15.08
	Strong	3.75	5.53	5.39	9.35	9.95	12.08	12.92	14.98
500	Weak	3.86	6.66	5.40	9.43	9.95	12.28	13.32	15.09
	Moderate	3.81	6.60	5.43	9.47	9.96	12.26	13.25	15.10
	Strong	3.74	6.55	5.38	9.39	9.99	12.25	13.23	15.03
750	Weak	3.89	6.65	5.42	9.50	10.04	12.37	13.20	15.19
	Moderate	3.84	6.61	5.34	9.43	9.99	12.24	13.36	15.09
	Strong	3.75	6.57	5.37	9.37	9.99	12.17	13.31	15.13

Note. SoD = strength of dependence. A = ATFIND. D = DETECT.