A Comparison of DIMTEST and Generalized Dimensionality Discrepancy Approaches to Assessing Dimensionality in Item Response Theory

by

Ray E. Reichenberg

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

Approved June 2013 by the
Graduate Supervisory Committee:

Roy Levy, Co-Chair
Marilyn Thompson, Co-Chair
Samuel Green

ARIZONA STATE UNIVERSITY

August 2013

ABSTRACT

Dimensionality assessment is an important component of evaluating item response data. Existing approaches to evaluating common assumptions of unidimensionality, such as DIMTEST (Nandakumar & Stout, 1993; Stout, 1987; Stout, Froelich, & Gao, 2001), have been shown to work well under large-scale assessment conditions (e.g., large sample sizes and item pools; see e.g., Froelich & Habing, 2007). It remains to be seen how such procedures perform in the context of small-scale assessments characterized by relatively small sample sizes and/or short tests. The fact that some procedures come with minimum allowable values for characteristics of the data, such as the number of items, may even render them unusable for some small-scale assessments. Other measures designed to assess dimensionality do not come with such limitations and, as such, may perform better under conditions that do not lend themselves to evaluation via statistics that rely on asymptotic theory. The current work aimed to evaluate the performance of one such metric, the standardized generalized dimensionality discrepancy measure (SGDDM; Levy & Svetina, 2011; Levy, Xu, Yel, & Svetina, 2012), under both large- and small-scale testing conditions. A Monte Carlo study was conducted to compare the performance of DIMTEST and the SGDDM statistic in terms of evaluating assumptions of unidimensionality in item response data under a variety of conditions, with an emphasis on the examination of these procedures in small-scale assessments. Similar to previous research, increases in either test length or sample size resulted in increased power. The DIMTEST procedure appeared to be a conservative test of the null hypothesis of

unidimensionality. The SGDDM statistic exhibited rejection rates near the nominal rate of .05 under unidimensional conditions, though the reliability of these results may have been less than optimal due to high sampling variability resulting from a relatively limited number of replications. Power values were at or near 1.0 for many of the multidimensional conditions. It was only when the sample size was reduced to $N = 100$ that the two approaches diverged in performance. Results suggested that both procedures may be appropriate for sample sizes as low as $N = 250$ and tests as short as $J = 12$ (SGDDM) or $J = 19$ (DIMTEST). When used as a diagnostic tool, SGDDM may be appropriate with as few as $N = 100$ cases combined with $J = 12$ items. The study was somewhat limited in that it did not include any complex factorial designs, nor were the strength of item discrimination parameters or correlation between factors manipulated. It is recommended that further research be conducted with the inclusion of these factors, as well as an increase in the number of replications when using the SGDDM procedure.

DEDICATION

This work is dedicated to my wife, Erynn, without whom I would never have been able to undertake such a project. Her support and efforts to motivate me are as much a part of my success as any of my own doing.

ACKNOWLEDGEMENTS

There are several people whose input made the completion of this work possible. I would like to thank Drs. Samuel Green and Marilyn Thompson for their feedback and suggestions. In particular, I would like to acknowledge Dr. Roy Levy's contribution throughout the process that led to the eventual completion of this work. Finally, I would like to thank Derek Fay, Nedim Yel, and Yuning Xu for their assistance with some of the technical aspects of the project.

TABLE OF CONTENTS

Page

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Item response theory (IRT) models have garnered significant attention amongst both researchers and practicing psychometricians since their introduction in the mid-20th Century. These models have been applied in a wide variety of fields, perhaps most notably in the areas of psychological and educational assessment. Typical applications of IRT models often include both large samples of examinees and large item pools. Such scenarios might be characterized as "large-scale" testing environments. As access to these approaches increases, whether due to a heightened awareness of their advantages over traditional methods or advances in the computational resources required to estimate such models, practitioners are seeking to apply them in situations that may lack these large-scale characteristics. This may include small pilot studies, classroom or individual school-level assessments, or applied studies with limited participant access. Researchers working under such conditions may lack access to large participant or item pools while still harboring the same goals as those often found in large-scale testing scenarios.

An assumption underlying the use of IRT models for many applications is that of unidimensionality, or that a single dimension, denoted θ, drives examinee responses. Violation of this assumption may result in inaccurate estimates of the modeled parameters and incorrect interpretations of the resulting test scores (Yen, 1993). Existing methods of assessing this unidimensionality assumption (e.g., DIMTEST; Nandakumar & Stout, 1993; Stout, 1987; Stout, Froelich, & Gao,

2001) have been shown to work well under certain conditions, such as when sample sizes and item pools are large and items tend to be highly discriminating. These methods may not, however, be robust to use under less desirable conditions. Other approaches, such as those that do not require partitioning items into subtests or that take advantage Markov Chain Monte Carlo sampling methods may prove more useful under small-scale testing conditions. The primary goal of the current work was to investigate the performance of one such method, the *standardized generalized dimensionality discrepancy measure* (SGDDM; Levy, Xu, Yel, & Svetina, 2012), relative to the DIMTEST approach. Though these approaches can be applied to item responses stemming from tests designed for use in any number of fields, the discussion in the following chapters will be focused on applications in educational assessment.

The remainder of this chapter is dedicated to providing an overview of the concepts central to the primary goal of the current study. The theory and assumptions underlying item response theory (IRT), as well as relevant IRT models will be presented first. Following that, Zhang and Stout's (1999a; 1999b) conditional covariance theory will be summarized and its use in assessing dimensionality will be discussed. Next, the logic and process of the DIMTEST and model-based covariance approaches to dimensionality assessment (e.g., SGDDM) will be outlined. Finally, as the SGDDM approach is applied using a posterior predictive model-checking (PPMC) framework, a brief overview of that framework, as well as Bayesian approaches to inference and estimation in general, will be given.

IRT Models

A number of item response models with varying levels of complexity exist in practice. The complexity of the model may be a function of the number of item parameters (difficulty, discrimination, etc.) that it specifies, the number of underlying dimensions that it assumes, the number of item response categories that it is capable of accommodating, or of some other source. A brief overview of dimensionality and the handling of item and person parameters in IRT are presented in the following section(s). With respect to the nature of the item responses, the current work focuses, in particular, on models designed to deal with dichotomous item responses in which a binary outcome is hypothesized to be a function of a latent, or unobserved characteristic of any respondent *i* and a set of characteristics for any item *j*. These responses are typically denoted as $X_{ij} = 1$ and $X_{ij} = 0$, indicative of a correct or incorrect response, respectively.

Unidimensional IRT models. As was mentioned earlier, most IRT applications assume that participant responses depend on a single underlying dimension, θ. Figure 1 depicts this scenario graphically using conventions similar to those typically used structural equation modeling (see Kline, 2010 for examples). The circle represents a latent variable while the squares represent observed variables. In this case, these are representative of the latent person abilities and observed examinee responses, respectively. Lines through the observed variables indicate the thresholds that delineate the amount of the latent characteristic that is required to endorse (i.e., correctly answer) a particular item. The number of thresholds estimated is a function of the number of available

response categories. For dichotomous type data, only one threshold is present. Arrows emanating from the latent variable to the observed variables indicate the direction of dependency. The realized values of the observable variable(s), then, are a function of the examinee's latent ability relative to the item's location (difficulty), and possibly other parameters, as described below. Both the item and person parameters are relative to an identical scale under the IRT framework. This implies that, unmodeled local dependencies (e.g., unaccounted for dimensionality, cheating, group problem-solving, etc.) or other threats to data-model misfit notwithstanding, the probability of an examinee endorsing that item is a function of the difference between the value of the examinee's ability on the latent scale and the value of the item's location (difficulty) on the latent scales. Furthermore, for the vast majority of IRT models used in education, including the models employed here, as latent ability increases, the probability of endorsing an item should also increase, assuming the item parameters are held constant.

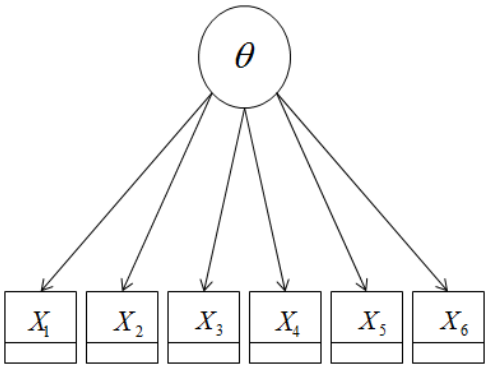

Figure 1. Six-item test with a single latent dimension and dichotomously scored items.

The majority of dichotomous, unidimensional IRT applications utilize one of three hierarchically related item response functions (IRFs; see de Ayala, 2009;

Embretson & Reise, 2000 for examples). The most general of these is the three-parameter logistic model (3-PL), which defines the probability of examinee *i* endorsing an item given their ability, θ_i, as:

$$P(X_{ij}=1|\theta_i,\alpha_j,\delta_j,c_j)=c_j+(1-c_j)\frac{exp[D\alpha_j(\theta_i-\delta_j)]}{1+exp[D\alpha_j(\theta_i-\delta_j)]}, \quad (1)$$

where α_j, δ_j, and c_j denote the discrimination, difficulty, and lower-asymptote (pseudo-guessing) parameters for item *j* and *D* is usually taken to be 1 but may take on other values as discussed below. Applying constraints to particular item parameters yields one of the more restricted models nested within the 3-PL. Fixing $c_j = 0$ while allowing α_j to vary for each of *J* items yields the two-parameter logistic model (2-PL), while constraining α_j to be equal across items, effectively forcing there to be a single discrimination parameter for an item set, yields the one-parameter (1-PL) model.

Though the aforementioned models are the most commonly used in practice, the current work employs an alternate function, the two-parameter normal ogive model (Lord & Novick, 1968; McDonald, 1999) in order to aid in estimation. This model utilizes similar parameters to the logistic family of models, but calculates the probability of success with respect to the normal distribution. It is given by:

$$P(X_{ij}=1|\theta_i,\alpha_j,\delta_j,c_j)=c_j+(1-c_j)\Phi(\alpha_j\theta_i+\delta_j), \quad (2)$$

where Φ represents the cumulative normal distribution function. Results from the three-parameter logistic and normal ogive IRFs closely resemble each other when $D = 1.7$ in *Equation 1.* According to Hambleton, Swaminathan, and Rogers

(1991), response probabilities between these two functional forms differ by less than .01 once the scaling factor has been applied.

Multidimensional IRT models. Often times a test, or particular test item requires multiple abilities in order to obtain a correct response. That is to say that there may be more than one latent characteristic underlying the item response(s). A mathematics word problem, where knowledge of the mathematical concepts as well as an ability to read the problem are required for success, is an example of such a situation. Several similarities exist between models applied in this type of scenario and those discussed in the previous section. There is still an assumption of monotonicity, or that the probability of success on an item increases when the level of the multiple abilities being measured increases. This allows the same functional form to be applied to both unidimensional and multidimensional models. The multidimensional family of item response models can be thought of as extensions of their unidimensional counterparts (for examples, see McDonald, 1997; Reckase, 1985; Reckase, 1997). Several new features apply to these models that were not necessary when modeling a single dimension. Of particular note is that each respondent is characterized by multiple person parameters instead of a single scalar parameter. These person parameters are often denoted θ_{im}, or the ability of person i on dimension m. Figure 2 provides a path diagram representative of a two-dimensional model. The curved line between the latent variables represents a relationship between them by way of a correlation or covariance. The dashed lines emanating from the latent characteristics to the

observed variables indicate items with cross-loadings, or items for which more than one ability influences the probability of a correct response.

When multiple dimensions best characterize a test, it may be ideal for each item to represent only one of these underlying traits. Tests that exhibit such a structure are said to be factorially simple. This is in contrast to complex structure where items may be dependent on, or "load on," multiple dimensions. Between these two structures lies approximate simple structure, where each item loads strongly on a single dimension and trivially (but still non-zero) on one, or more auxiliary dimensions. For the current work, the term "complex structure" will be used to indicate any scenario where an item exhibits a non-zero loading on more than one dimension.

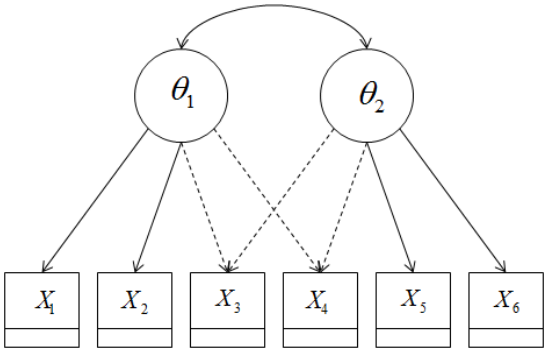

Figure 2. Six-item test with two correlated latent dimensions and dichotomously scored items.

The most commonly used models in multidimensional IRT are the compensatory models (Ackerman, 1989; Bolt & Lall, 2003). The 3-PL compensatory MIRT model is expressed as:

$$P(X_{ij}=1|\boldsymbol{\theta}_i,\boldsymbol{\alpha}_j,\delta_j,c_j)=c_j+(1-c_j)\frac{exp(\boldsymbol{\alpha}_j'\boldsymbol{\theta}_i+\delta_j)}{1+exp(\boldsymbol{\alpha}_j'\boldsymbol{\theta}_i+\delta_j)}, \quad (3)$$

where $\boldsymbol{\alpha}_j = (\alpha_{j1}, \ldots, \alpha_{jM})'$ 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 parameter, and δ_j represents a scalar related to item difficulty (Reckase, 1985; 1997). As was the case with the unidimensional models, there exists a hierarchical relationship between the multidimensional models such that fixing particular parameters yields a nested model. Fixing $c_j = 0$ yields the 2-PL MIRT model, and fixing all elements in the item discrimination vector ($\boldsymbol{\alpha}_j$) as equal yields the 1-PL MIRT model.

There also exists a multidimensional extension to the normal ogive model presented in the previous section, the 3-PL form of which is given by:

$$P(X_{ij} = 1 | \boldsymbol{\theta}_i, \boldsymbol{\alpha}_j, \delta_j, c_j) = c_j + (1 - c_j)\Phi(\boldsymbol{\alpha}_j'\boldsymbol{\theta}_i + \delta_j), \quad (4)$$

where Φ denotes the standard normal cumulative function (Bock & Aitkin, 1981; McDonald, 1967). This model may be constrained to yield the 2-PL and 1-PL forms that were possible with the logistic functions.

Conditional Independence in IRT

Local independence (LI), an instance of conditional independence, is a central assumption of item response models. Local independence stipulates that examinee responses to any pair of items are statistically independent when the parameters influencing their performance are held constant. That is to say that the responses are independent *conditional* on the model parameters. These parameters include the possibly vectored set of abilities $\boldsymbol{\theta}$, as well as the set of item parameters, denoted ω_j. This assumption is often represented formally as:

$$P(X_{ij},...,X_{iJ}|\boldsymbol{\theta}_i,\omega_j)=\prod_{j=1}^{J}P(X_{ij}|\boldsymbol{\theta}_i,\omega_j)\,. \quad (5)$$

Violation of the LI assumption is often referred to as local dependence (LD). Yen (1993) identified a host of potential sources of local dependence including external assistance, insufficient time to complete a task (i.e., speededness), fatigue, or a practice effect associated with exposure to multiple items of a similar type. Perhaps more importantly, evidence of local dependence may imply the existence of some unmodeled dimensionality. That is to say that some underlying characteristic may exist that is influencing examinee performance beyond what has been included in the model. This dimensionality may be of substantive interest to the researcher or may stand as nothing more than a "nuisance" dimension. Either way, a failure to account for said dimension may yield imprecise parameter estimates, which, in turn can influence the interpretation and use of test scores (Yen, 1993).

Weaker forms of LI. The assumption put forth by *Equation 5* may not always hold in practice. Satisfaction of this assumption, often referred to as strong local independence (SLI), requires not only that all bivariate dependencies be accounted for by the model parameters, but also that all higher-order dependencies be accounted for as well. Isolating these higher-order dependencies can be difficult in practice. Furthermore, if all bivariate dependencies are well modeled, higher-order dependencies, though possible, are unlikely (McDonald, 1994). Weak local independence (WLI; McDonald, 1994), also called pairwise

independence, focuses just on these bivariate dependencies. This assumption dictates that the following holds true:

$$cov(X_{ij}, X_{ij'} | \boldsymbol{\theta}_i, \omega_j) = 0 \text{ for all } \theta \text{ and } 1 \leq j < j' \leq J, \quad (6)$$

where *cov* denotes a covariance. When SLI holds, then WLI will also hold, mathematically speaking. When WLI is true, however, SLI may not necessarily hold if there exists some higher-order item dependency. McDonald (1994) argued that WLI is an empirically sufficient assumption in place of SLI as data sets that exhibit these higher-order dependencies may be rare in practice (Zhang & Stout, 1999b).

Stout (1987) advanced the notion of essential independence (EI), an assumption that is central to the DIMTEST procedure investigated in the current work. Essential independence is satisfied when the following holds:

$$\frac{\sum_{1 \leq j < j' \leq J} \left| cov(X_{ij}, X_{ij'} | \boldsymbol{\theta} = \boldsymbol{\theta}') \right|}{\binom{J}{2}} \rightarrow 0 \ , \quad (7)$$

for all $\boldsymbol{\theta}'$ as $J \rightarrow \infty$. EI differs from the previous two forms of independence (SLI and WLI) in that it is concerned with average independence as opposed to independence by item-pairs. Under this assumption, the average conditional covariance should be small and become smaller as the number of items, *J*, approaches infinity. Secondly, EI is only concerned with *dominant* dimensions, as opposed to all dimensions. The minimum number of dominant dimensions needed to satisfy *Equation 7* above is considered the essential dimensionality. If a single dominant dimension is able to satisfy the necessary conditions, then the set of

items is said to be essentially unidimensional (Nandakumar & Yu, 1996). As can be inferred from information to be presented in later sections, the DIMTEST procedure is predicated upon the notion of essential independence, whereas the SGDDM statistic utilizes the weak local independence assumption. This necessarily implies that the DIMTEST procedure carries with it the assumption of infinite, or at least sufficient test lengths.

Conditional Covariance Approaches to Dimensionality

Conditional covariance theory (CCT; Zhang & Stout, 1999a; 1999b) lies at the foundation of the DIMTEST and SGDDM methods, which are discussed in forthcoming sections, as well as HCA/CCPROX and DETECT, which are covered briefly in the section on subtest partitioning. A more rigorous discussion of the DETECT and HCA/CCPROX methods can be found in Zhang & Stout (1999b) and Stout et al. (1996). CCT was developed as a nonparametric alternative to parametric approaches to assessing dimensionality. While parametric approaches make certain assumptions with respect to the form of the item response function (IRF), CCT requires only that the function be monotonic. That is to say that the probability of a correct response should approach one as the possibly vectored latent characteristics approach infinity ($P(X_{ij}=1) \rightarrow 1$ as $\boldsymbol{\theta}_i \rightarrow \infty$). Zhang & Stout (1999a) used a generalized *m*-dimensional compensatory model in their presentation of CCT. This model is given by:

$$P(X_{ij}=1|\boldsymbol{\theta}_i) = H_j\left(\sum_{m=1}^{M}\alpha_{jm}\theta_m - \delta_j\right) , \tag{8}$$

where H_i is any non-decreasing (i.e., monotonic) function and all other notations take on their standard meanings.

Three features are central to CCT; the item, the unidimensional composite score for each dimension, and the total composite score for all dimensions (Stout et al., 1996). These features are related in that responses to the items combine to form item-weighted unidimensional composites for each dimension. These unidimensional composites, in turn, combine to form the dimensionally weighted total test composite. Stout and colleagues (1996) demonstrated this geometrically with the vector diagram presented in Figure 3. This diagram depicts a scenario wherein a set of items is characterized by two dimensions (θ_1 and θ_2). The total test composite is denoted θ_{TT} while the unidimensional composite scores are denoted by θ_{C1} and θ_{C2}. The individual items are represented as vectors clustered around their respective unidimensional composites.

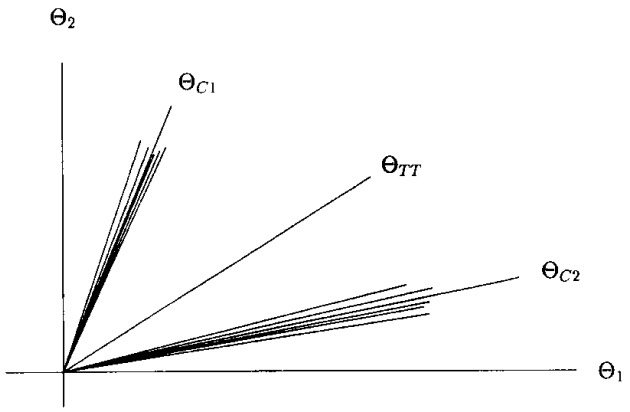

Figure 3. Geometric representation of a two-dimensional test (Stout et al., 1996).

Other key features of the diagram reveal the test structure, as well as additional characteristics of the item(s). The direction of the individual item vectors is often referred to as the direction of best measurement (Stout et al.,

1996), whereas the length of the vector indicates the discrimination of the item, with longer vectors representing larger magnitudes. For a test exhibiting simple structure and orthogonal (i.e., uncorrelated) dimensions, the vectors for all of the items measuring θ_{C1} and, thus, the θ_{C1} composite score, would align with the θ_1 axis. The same would hold for the items measuring θ_{C2} and the θ_2 axis. More commonly, tests exhibit more complex structures with partially overlapping (i.e., oblique) dimensions, as in the scenario depicted in Figure 3. In this case, the vectors for the unidimensional composites, denoted θ_{C1} and θ_{C2}, deviate from their respective axes (θ_1 / θ_2), indicating the presence of correlated dimensions. Furthermore, the item-specific vectors do not perfectly align with the unidimensional composites they are intended to represent. This is indicative of complex structure, or items that, in this case, measure one dimension best, but have non-zero loadings on a second dimension. Had these item vectors aligned with the unidimensional composites, simple structure would still have held despite the presence of a non-zero correlation between the dimensions.

Zhang and Stout (1999a) also put forth a relationship between the directionality of the items vectors and the degree of multidimensionality present in a set of items. Conditional on the total test composite, denoted θ_{TT} in Figure 3, any two items with directions of best measurement on the same side of the total score composite will exhibit positive conditional covariances; any two items with directions on opposite sides of θ_{TT} will exhibit negative conditional covariances; and if at least one of the item vectors lies on the total score composite, the conditional covariance between that item and all other items will be zero. With

respect to the scenario depicted in Figure 3, two items taken from the same cluster (either θ_{C1} or θ_{C2}) would be expected to exhibit a positive covariance, conditional on θ_{TT}, while two items taken from different clusters would be expected to exhibit a negative conditional covariance.

DIMTEST

DIMTEST (Nandakumar & Stout, 1993; Stout, 1987; Stout, Froelich, & Gao, 2001) is a commonly used method for assessing whether a single dimension is sufficient to model a set of item responses. Specifically, the DIMTEST approach is concerned with conducting a formal test of the essential unidimensionality assumption. It achieves this by splitting an item pool into two separate clusters, then evaluating the distinctness of the responses in each cluster. The first of these clusters, termed the assessment subtest (AT), is chosen such that the items contained within the partition are dimensionally similar (i.e., homogenous) to one another, but as dimensionally distinct from the remaining items as possible. The second cluster, the partitioning subtest (PT), consists of all items not used in AT and is used to cluster examinees based on their total PT subtest score. The separation of items into these two clusters can be, and has historically been done using a variety of approaches ranging from those stemming from the factor analytic tradition (see Stout, 1987) to clustering algorithms employing CCT-based assessments of dimensional distinctness (see Zhang & Stout, 1999b). These partitioning strategies can be approached in either an exploratory of confirmatory manner. The current work relies on an exploratory

partitioning approach, a brief description of which is provided following the presentation of the DIMTEST procedure.

The null and alternative hypotheses tested by DIMTEST are given by Stout et al. (1996). They are:

H_0: $AT \cup PT$ satisfies essential unidimensionality ($d = 1$)

H_A: $AT \cup PT$ fails to satisfy $d = 1$

Restated, the null hypothesis posits that the AT and PT partitions assess the same dominant underlying dimension, while the alternative implies that the items in the AT partition are best represented by a dimension that is distinct from that driving responses to the PT items. As will be seen in the next section, the distinctness of the dimensionality underlying these two partitions is the primary driver of the value of the statistic, T, utilized by DIMTEST to reach a decision as to the hypotheses in question. Formally, the null hypothesis of $d = 1$ is rejected if $T \geq Z_\alpha$, where Z_α is the critical value that separates the upper $100(1 - \alpha)$ percentile of the standard normal distribution at the α significance level (Nandakumar & Stout, 1993).

At its heart, and regardless of the bias correcting procedure being implemented, the DIMTEST statistic is essentially a standardized difference between total variability and unidimensional variability of a set of responses, conditional on total test score (Stout, Froelich, & Gao, 2001). Equation 1.10 in the aforementioned work by Stout and colleagues (2001) demonstrated that the difference in these two variance estimates is equivalent to the estimated covariance between the pairs of items in AT, again conditional on PT score. If this

difference is equal to zero, indicating that a single dimension adequately explains all of the variability in the examinee responses, then one could conclude that the AT and PT items measure the same single dimension. Small, yet statistically insignificant differences between the two variance estimates would imply essential unidimensionality, or that a single dominant dimension is sufficient to satisfy the assumption of local independence. Significantly large differences indicate that the two subtests represent, at minimum, two distinct dimensions, resulting in a rejection of the hypothesis of essential unidimensionality.

The process described in the previous paragraph can perhaps be more aptly described using graphic representations of scenarios likely to yield a rejection, or a failure to reject the DIMTEST hypothesis. Figures 4a and 4b, taken from Froelich and Habing (2008), depict vector diagrams of poor and good choices for an AT/PT partition, respectively. The items denoted AT_1, AT_2, etc. are those in the AT partition. The items in the PT partition are not shown but, rather, the θ_{PT} composite vector is shown in their stead. The length of each item vector represents the magnitude of the composite item discrimination, while the angle from the θ_1 axis indicates the composite item direction. The aforementioned DIMTEST procedure would be more likely to yield a rejection of the unidimensionality hypothesis under the scenario presented in Figure 4b than for Figure 4a.

(a) Poor Partition (b) Good Partition

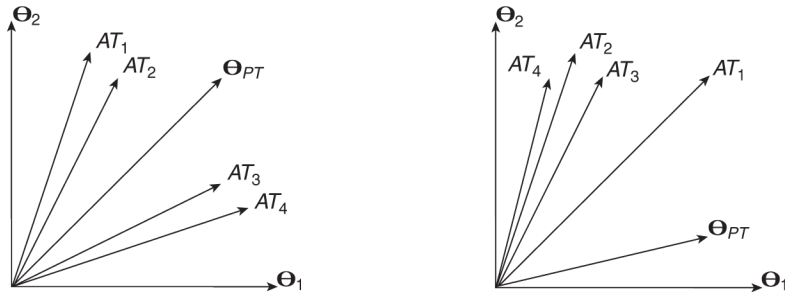

Figure 4. Examples of poor and good choices for AT and PT (Froelich & Habing, 2008).

The AT partition in Figure 4a does not comprise a set of dimensionally homogenous items as can be seen by the lack of any clustering of the item vectors. Rather, the items appear to exhibit composite directions that are distributed relatively uniformly through the latent variable space; some measure θ_1 best while others measure θ_2 best. Furthermore, the direction of the θ_{AT} composite vector, which is not shown in the diagram, would most likely fall along the θ_{PT} vector, indicating that the AT items do not measure a dimension that is distinct from that of the PT items.

Figure 4b, on the other hand, depicts a scenario in which the unidimensionality hypothesis evaluated by DIMTEST would likely be rejected. The θ_{AT} and θ_{PT} composite vectors are much more distinct that what was seen in Figure 4a. Additionally, the items contained within the AT partition are tightly clustered, indicating that they are relatively homogenous with respect to their direction of best measurement.

Calculating the DIMTEST statistic. After the AT/PT partitions have been chosen, examinees are separated into *k* subgroups based upon their score on the PT items. The next step, then, in arriving at the value of the DIMTEST statistic, as summarized by Stout et al. (2001), is to calculate the total score for examinee *i* in subgroup *k* as:

$$Y_{ik} = \sum_{j \in AT}^{J} X_{ijk} \ , \tag{9}$$

where X_{ijk} denotes the response (either a "1" or a "0") provided by examinee *i* in subgroup *k* to item *j*. The average total for the *I* examinees in subgroup *k* is then:

$$\overline{Y}_k = \frac{1}{I_k} \sum_{i=1}^{I_k} Y_{ik} \ . \tag{10}$$

Using the values obtained in *Equation 9* and *Equation 10*, the estimate of the variance for the examinee scores on the AT subtest, conditional on PT score, can be calculated as:

$$\hat{\sigma}_k^2 = \frac{1}{I_k} \sum_{i=1}^{I_k} (Y_{ik} - \overline{Y}_k)^2 \ . \tag{11}$$

In order to estimate the unidimensional variance for a particular subgroup, the difficulty for each item, *j*, within the subgroup, *k*, must first be estimated as:

$$\hat{p}_{jk} = \frac{1}{I_k} \sum_{i=1}^{I_K} X_{ijk} \ . \tag{12}$$

For dichotomously scored items, $\hat{p}_{jk}$ is essentially the proportion of examinees in a particular subgroup that got the item correct. Using this difficulty estimate, the unidimensional variance for the *k*th subgroup is given by:

$$\sigma_{U,k}^2 = \sum_{j=1}^{J} \hat{p}_{jk}(1-\hat{p}_{jk}) . \quad (13)$$

The difference between the total and unidimensional variance estimates for each subgroup is then the estimate of the conditional covariance amongst all item pairs for that subgroup. The logic here is that, if conditioning on total PT score is sufficient to satisfy the EI assumption, then all of the examinees within a particular subgroup should respond to a particular item in essentially the same manner. Should that be the case, the variance estimates should be both small, and similar to one another, yielding small difference between the two. This difference score is often denoted as $T_{L,k}$. In order to conduct a statistical test of the null hypothesis of unidimensionality, the difference between the total and unidimensional variances needs to be transformed to a standard metric. This is done by dividing the differences by the estimate of the variance of $T_{L,k}$, which is calculated as:

$$S_k^2 = \frac{(\hat{u}_{4,k} - \sigma_k^4) - \hat{\delta}_{4,k}}{J_k}, \quad (14)$$

where

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

and

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

Integrating the elements from *Equations 11*, *13*, and *14*, the DIMTEST statistic is given as:

$$T_L = \frac{\sum_{k=1}^{K} (\hat{\sigma}_k^2 - \hat{\sigma}_{U,k}^2)}{\sqrt{\sum_{k=1}^{K} S_k^2}}. \quad (15)$$

Stout (1987) showed that this statistic follows an approximate standard normal distribution under the null hypothesis for long tests and large examinee pools. The statistic tends to be positively biased, however, when short tests are used (Stout et al., 2001). This can lead to inflated Type I error rates. That is to say that the approach may suggest the presence of additional dimensions when the test is in fact unidimensional more often than is acceptable. Stout (1987) originally corrected for this bias via the use of a second assessment subtest, AT2. The value of the DIMTEST statistic was calculated for both AT1 and AT2 and a bias-corrected test statistic was given by:

$$T = \frac{T_L - T_B}{\sqrt{2}}, \quad (16)$$

where T_B is the value of the DIMTEST statistic obtained using the AT2 partition. This bias-corrected statistic was shown to perform well under a variety of testing conditions (see Nandakumar & Stout, 1993; Stout, 1987). When the AT1 items were of a similar difficulty or had large discrimination values, however, then the AT2 partition tended not to remove enough of the bias in the test statistic. Additionally, the need for a third item partition placed an unnecessary strain on the available item pool.

The current version of DIMTEST uses a resampling, or bootstrapping approach proposed by Stout, Froelich, and Gao (2001). Under this procedure, the value of T_L is calculated using the approach described above. The bias-correction

factor is estimated by first estimating item response functions (IRFs) for each item using the observed data. The method for estimating these IRFs is detailed in Stout et al. (2001). A series of new data sets are then generated using the IRFs. The DIMTEST statistic is calculated for each of these data sets. The average value of the statistic across all of the simulated data sets, denoted $\overline{T}_G$, is used to remove the bias present in T_L. The value of the new, bias-corrected test statistic is given by:

$$T = \frac{T_L - \overline{T}_G}{\sqrt{1+1/N}}, \tag{17}$$

where N is the number of data sets generated. This most recent instantiation of the DIMTEST procedure has been shown to exhibit both greater power and better control of Type I error rates than were demonstrated by previous versions (Finch & Habing, 2007; Froelich & Habing, 2008; Stout et al., 2001).

Item partitioning via ATFIND. As was mentioned earlier, the current version of DIMTEST offers both exploratory and confirmatory methods of partitioning items into the AT and PT clusters. Confirmatory methods involve the researcher separating the items manually, usually based upon some *a priori* theory or content expert feedback. Exploratory methods may still offer advantages to researchers, even in situations where strong *a priori* beliefs about the nature of the items are present. These advantages might include the opportunity to find agreement between the statistical partition and substantive beliefs or the uncovering of alternative partitions that may provide new insights into the structure of the test (Fay, 2012). The current work utilizes the exploratory

approach to partition items. This approach is generally referred to as ATFIND in the DIMTEST literature. ATFIND employs two separate CCT-based methods, HCA/CCPROX and DETECT, to propose and evaluate potential item groupings (for a more detailed description of these methods, see Froelich & Habing, 2008; Roussos, Stout, & Marden, 1998; Zhang & Stout, 1999b).

HCA/CCPROX, a procedure put forth by Roussos et al. (1998), is used to propose potential test partitions. Under this method, the proximity of each item to every other item in the test is determined using a conditional covariance-based approach (the CCPROX step). Next, an agglomerative hierarchical clustering procedure (HCA) is used to cluster items, or groups of items based upon their proximity. The process starts with J distinct clusters, where J is the number of items in the test, and is considered complete when all of the items are contained within a single cluster. At each step between these points, the two clusters that are the most similar are joined to form a single cluster.

The HCA/CCPROX method does not include any built-in functionality for evaluating the test partitions that it generates in terms of their dimensional distinctness. To that end, the DETECT index (Zhang & Stout, 1999b) is employed. This index is given by:

$$D(P,\theta) = \frac{2}{J(J-1)} \sum_{1 \le j < j' \le J} \delta_{j,j'} E[cov(X_j, X_{j'} \mid \theta_{TT})], \qquad (18)$$

where $\delta_{j,j'}$ (not to be confused with the δ_j used to indicate the difficulty of item j in earlier sections) takes on a value of 1 if items j and j' are in the same cluster, and a value of (-1) if they are in different clusters. This essentially penalizes the

value of the index when large, positive conditional covariances between items in separate clusters or large negatives for items in the same cluster are present This penalty is in keeping with the goal of DETECT, which is to find the partitioning of items that offers the largest deviation from dimensional similarity. The index, then, is maximized when items within a cluster exhibit strong positive relationships, conditional on the total score composite, and items in different clusters exhibit large negative relationships. Evaluated for any two cluster solution, the theoretical maximum, referred to as $DETECT_{max}$, would occur when the items in the two groups are as dimensionally distinct as possible.

Given the preceding descriptions of HCA/CCPROX and DETECT, the procedure for choosing the partitioning that best satisfies the requirements that (1) the items in AT be as homogenous as possible, and (2) the AT and PT clusters be as heterogeneous as possible, as summarized by Froelich and Habing (2008), is as follows:

1. Run HCA/CCPROX.
2. Each cluster for which $4 \leq j \in AT \leq J / 2$ is satisfied is considered a potential AT partition. The PT partition is then defined as the remaining test items.
3. Calculate the value of DETECT for each potential test partition from Step 2.
4. The AT/PT pairing with the largest DETECT value is selected as the AT and PT for use in DIMTEST.

As Step 2 implies, the AT subtest must contain at least four items. The DIMTEST program also stipulates that the PT subtest contain no less than 15 items. As such, the commercially available version of DIMTEST cannot be implemented for tests containing less than 19 total items. Further complicating matters is the fact that it is common-practice to utilize a separate subset of the examinee pool to conduct AT/PT partitioning than that used for calculating the DIMTEST statistic (Socha & DeMars, 2013). This may reduce DIMTEST's usefulness when only a very limited number of examinees are available.

Model-Based Covariance Approaches to Dimensionality

One potentially limiting characteristic of DIMTEST is that it focuses only on positive local dependence. This is a result of the procedure conditioning on θ_{PT} as opposed to a total score composite comprised of the entire set of items, often denoted θ_{TT}. The implication of this decision is that, since the AT items are dimensionally homogenous and as distinct as possible from the PT items, the covariances between the AT item pairs will tend towards positive values. As Roussos and Habing (2003) pointed out, however, the existence of positive local dependence implies the presence of negative local dependence. Failure to account for these negative dependencies may hamper the performance of a method aimed at assessing dimensionality. This failure may be particularly bothersome in situations where multidimensionality manifests itself as negative local dependence, such as in cases where a large portion of the item pool represent more than one dimension; that is to say that the item composite directions are relatively dispersed throughout the dimensional space (Levy & Svetina, 2011). In

cases such as these, using an approach that is sensitive to both positive and negative local dependence may be more appropriate.

One such metric is the model-based covariance, or MBC (Reckase, 1997), which is given by:

$$\text{MBC}_{jj'} = \frac{\sum_{i=1}^{I}(X_{ij} - E(X_{ij}|\boldsymbol{\theta}_i, \boldsymbol{\omega}_j))(X_{ij'} - E(X_{ij'}|\boldsymbol{\theta}_i, \boldsymbol{\omega}_{j'}))}{I}, \qquad (19)$$

where I is the number of examinees and $E(X_{ij}|\boldsymbol{\theta}_i, \boldsymbol{\omega}_j)$ is the expected value of the item response for examinee i to item j. As is clear from *Equation 19*, the value of $\text{MBC}_{jj'}$ can take on both positive as well as negative values. Values greater (less) than zero are indicative of positive (negative) local dependence, while a value of zero implies that the local independence assumption has been met. MBC conditions on the model-implied latent ability, $\boldsymbol{\theta}_i$, instead of a subtest score, as was the case with DIMTEST. MBC has been shown to exhibit acceptable power and control of Type I error rates in all but the most extreme testing conditions (Levy, Mislevy, & Sinharay, 2009).

Building off of the MBC metric, Levy and Svetina (2011) proposed the *generalized dimensionality discrepancy measure*, GDDM, defined as:

$$\text{GDDM} = \frac{\sum_{j \neq j'} \left| I^{-1} \sum_{i=1}^{I}(X_{ij} - E(X_{ij}|\boldsymbol{\theta}_i, \boldsymbol{\omega}_j))(X_{ij'} - E(X_{ij'}|\boldsymbol{\theta}_i, \boldsymbol{\omega}_{j'})) \right|}{J(J-1)}, \qquad (20)$$

where J is the number of items. GDDM is essentially the average absolute value of $\text{MBC}_{jj'}$ across all possible item pairs. Taking the absolute value of *Equation 19* allows both positive and negative local dependence to contribute to the value of GDDM (Levy & Svetina, 2011). GDDM can assume values ≥ 0, with equality

holding when McDonald's (1994) weak local independence assumption is met. Using a posterior predictive model checking (PPMC) framework, Levy and Svetina (2011) compared the performance of GDDM to a selection of other dimensionality assessment approaches, and showed that near nominal Type I error rates can be expected when using GDDM, even with relatively strongly correlated dimensions and fairly subtle multidimensionality are present. Power was satisfactory under most analysis conditions.

Interpreting the realized values of the MBC and GDDM statistics can be difficult as their scales are metric dependent. In order to alleviate this lack of interpretability, Levy, Xu, Yel, and Svetina (2012) introduced revised versions of these two measures, termed SMBC and SGDDM, with the *S* indicating standardization. The value of SMBC is given as:

$$\text{SMBC}_{jj'} = \frac{\dfrac{\sum_{i=1}^{I}(X_{ij} - E(X_{ij}|\boldsymbol{\theta}_i, \boldsymbol{\omega}_j))(X_{ij'} - E(X_{ij'}|\boldsymbol{\theta}_i, \boldsymbol{\omega}_{j'}))}{I}}{\sqrt{\dfrac{\sum_{i=1}^{I}(X_{ij} - E(X_{ij}|\boldsymbol{\theta}_i, \boldsymbol{\omega}_j))^2}{I}}\sqrt{\dfrac{\sum_{i=1}^{I}(X_{ij'} - E(X_{ij'}|\boldsymbol{\theta}_i, \boldsymbol{\omega}_{j'}))^2}{I}}}, \quad (21)$$

while SGDDM is calculated as:

$$\text{SGDDM} = \frac{\sum_{j>j'}\left|\dfrac{\dfrac{\sum_{i=1}^{I}(X_{ij} - E(X_{ij}|\boldsymbol{\theta}_i, \boldsymbol{\omega}_j))(X_{ij'} - E(X_{ij'}|\boldsymbol{\theta}_i, \boldsymbol{\omega}_{j'}))}{I}}{\sqrt{\dfrac{\sum_{i=1}^{I}(X_{ij} - E(X_{ij}|\boldsymbol{\theta}_i, \boldsymbol{\omega}_j))^2}{I}}\sqrt{\dfrac{\sum_{i=1}^{I}(X_{ij'} - E(X_{ij'}|\boldsymbol{\theta}_i, \boldsymbol{\omega}_{j'}))^2}{I}}}\right|}{J(J-1)/2} \quad (22)$$

All four metrics, MBC, SMBC, GDDM, and SGDDM will return a value of zero when local independence holds. Increases in the dependencies between item pairs will yield increasingly positive values in all four cases. The two standardized metrics offer interpretability in the same manner as the usual correlation coefficient. As is clear from *Equation 22*, SGDDM is the average SMBC value across all unique item pairs. Levy and colleagues (2012) conducted a PPMC-based simulation study in which SGDDM was shown to be sensitive to the presence of unmodeled dimensionality and to indicate adequate data-model fit when the dimensionality was correctly specified. It is also important to note that SGDDM, as well as GDDM can be implemented for tests of any length, whereas DIMTEST is only feasible for tests with at least 19 items. As Levy and Svetina (2011) pointed out, MBC or SMBC may be more appropriate for cases where one might have substantive interest in the sign of the coefficient, such as in the exploring the relationship between a single pair of items.

As it has only recently been proposed, SGDDM is a relatively uninvestigated metric, though its predecessors and the conditional covariance theory from which it draws have been the subject of a number of publications. To date, no study has explored the utility of GDDM or SGDDM for the specific task of assessing deviation from unidimensionality. Finally, SGDDM stems from a line of research that has sought to investigate the applicability of the PPMC framework to assessing dimensionality in item response data, though no feature of the metric relegates it to being used exclusively in conjunction with PPMC. An overview of the PPMC method is presented in the next section.

Bayesian Parameter Estimation

Bayesian approaches to statistical modeling have received an increasing amount of attention and implementation in recent decades. This increase has been made more prominent by the ever-increasing amount of computing resources available to researchers. The methods that lie within the Bayesian family have been the subject of any number of books and articles, each of which varies in terms of their technicality and focus on practical applications. The following sections provide a cursory, and by no means technical overview of basic Bayesian inference and parameter estimation. More detailed and, in many cases, more technical treatments of these topics can be found in Gill (2007), Gelman, Carlin, Stern, and Rubin (2003), and Fox (2010; particular to Bayesian IRT models), amongst others.

To operate in a Bayesian framework doesn't just imply the use of a particular modeling framework, but may also refer to a way in which one organizes their thinking. This organization typically consists of three central questions: (1) What did I believe at the onset? (2) What did I observe? and (3) What do I believe now, given my initial beliefs and what I observed? In practice, Bayesian modeling involves combining initial beliefs and uncertainties about the parameters of interest with the observed data to yield an updated set of beliefs about those parameters. As opposed to frequentist statistical traditions, unknown quantities, such as model parameters, are treated as random variables. These variables can be described via a posterior distribution, which is, again, a combination of the researcher's prior beliefs and the observed data. These prior

beliefs, termed the prior distribution, offer an overt representation of *a priori* uncertainty regarding the model parameters. Inferences about the parameters can be made using the posterior distribution, which can be presented in its entirety or summarized in any of the ways typically used to summarize distributions (e.g., mean, median, mode, variance, etc.). The posterior obtained from one set of observations can be used as the prior distribution for the next round of data collection, essentially allowing for a continuous updating of one's beliefs and the inferences that are drawn from them.

Bayes' theorem. Bayes' theorem provides the architecture for implementing the type of inference described above; combing prior beliefs with observations to yield an updated set of beliefs. More formally, Bayes' theorem offers a mechanism for determining the probability of an unknown set of parameters given a particular set of data. The statistical model that is chosen by the researcher (e.g., 3-PL IRT model) governs the probability of the observed data given the unknown parameters. This is often referred to as the likelihood component. A probability model for the observations conditional on the unknowns and some prior knowledge of how those unknowns might be distributed (i.e., prior distributions) provide sufficient information to determine the probability of the unknown parameters conditional on the data via Bayes' theorem, which is given as:

$$P(\boldsymbol{\Omega}|\mathbf{X}) = \frac{P(\mathbf{X}|\boldsymbol{\Omega})P(\boldsymbol{\Omega})}{P(\mathbf{X})}, \tag{23}$$

$\boldsymbol{\Omega}$ here denotes a collection of unknown parameters and **X** represents a collection of observed values. The denominator on the right side of *Equation 23* serves as a scaling factor, ensuring unit integration on the left side of the equation, per Kolmogorov's 2nd axiom of probability. This term can be excluded, which yields:

$$P(\boldsymbol{\Omega}|\mathbf{X}) \propto P(\mathbf{X}|\boldsymbol{\Omega})P(\boldsymbol{\Omega}) \tag{24}$$

That is to say that the probability of the unknown parameters, given the data, is *proportional* to the product of the prior probability of the unknowns and the likelihood from the data. This form is often applied in practice for two reasons: (1) it simplifies the necessary calculations and can drastically reduce computing time; (2) in terms of a probability density function, the rescaling only affects the Y-axis. Researchers are typically not interested in the density values (Y-axis), but rather in the scale of the parameter (X-axis) and the relative frequencies of the values for the parameter (the shape of the density function).

The role of prior information. Prior distributions for unknown parameters afford the opportunity to encode prior beliefs before observing data. Often times, the choice for the prior is informed by previous observations (e.g., results from a pilot study) or a synthesis of existing research (i.e., meta-analysis). In the absence of any meaningful background or contextual information, the prior can be specified in such a way as to indicate a high-degree of uncertainty, essentially allowing the information contained in the observed data to dictate the posterior. In terms of estimation, the prior distribution can drastically increase accuracy when only limited observations are available. The influence of the prior distribution on the posterior is mitigated as the number of observations increases.

This effect is often referred to as "swamping" the prior. Many of the objections and/or criticisms of the Bayesian approach to statistical inference are based upon the perceived subjectivity or arbitrary nature of the selection of the prior. However, as Fox (2010) points out:

> "The prior choice can be disputable but is not arbitrary because it represents the researcher's thought. In this light, other non-Bayesian statistical methods are arbitrary since they are equally good and there is no formal principle for choosing between them. (p. 16)"

The role of the posterior. The posterior distribution represents the distribution of values for the unknown parameters given the observed data and the prior. Each unknown entity is assigned a prior and, as such, will also have a posterior. The mechanics of generating the prior vary depending on the relationship between the distributional form of the prior and that of the likelihood. When the posterior is of a known form, it can be sampled from directly. In situations where the form of the posterior is unknown, it can first be approximated using Metropolis-Hastings, or other related approaches. Most software packages for conducting Bayesian analyses use Gibbs sampling (Casella & George, 1992; Gelfand & Smith, 1990; Smith & Roberts, 1993) for taking draws from the posterior distribution. Under this procedure, every unknown parameter is first assigned an initial value, usually drawn from the prior. These values are then updated iteratively by sampling from the full conditional distributions, or the distribution for each parameter conditional on all of the other variables. If certain regularity conditions hold then, in the limit, a draw from the full conditional

distribution is equivalent to a draw from the posterior distribution (Gill, 2007). The value for a particular parameter at time t+1 is taken conditional on the values of the other parameters at time t. This process continues until the desired number of draws has been taken. The estimated posterior is then composed of the collection of all of these draws. The posterior distribution can be summarized and presented in any of the usual ways, such as reporting of the mean, median, standard deviation, percentile cut values, or a central credibility interval, which is akin to a confidence interval in the frequentist context.

Bayesian vs. frequentist approaches to modeling. Bayesian methods do not stem from the frequentist traditions of hypothesis testing, which typically aim to compare the value of a sample-derived statistic to what would be expected under null conditions. Instead, the goal of Bayesian modeling can be thought of as approximating the population distribution for a parameter in light of prior beliefs and observed data. Despite this, approaches to conducting something akin to a traditional null hypothesis test do exist in the Bayesian context (see Raftery, 1996 for examples).

Bayesian approaches to inference and modeling offer a number of advantages over traditional null hypothesis significance testing methods (Gill, 2007). These may include:

- Parameters are treated as random and changing, not fixed values.
- They allow the researcher to encode his/her prior beliefs into the modeling process.

- The "answer" is a distribution, not a point-estimate. This means that notions of uncertainty about model outcomes are built into the process. It also allows for flexibility in the way the parameter estimates are presented.
- Allows for updating of beliefs in the event that new data is collected.
- Offers an easy method of handling missing data. In the Bayesian framework, missing values are treated as unknown parameters, meaning they are assigned a prior distribution and can be evaluated using a posterior distribution.

Posterior Predictive Model Checking.

Existing investigations of the GDDM and SGDDM statistics have utilized a posterior predictive model-checking (PPMC) framework. PPMC focuses on discrepancies between the observed data and replicate sets of model-implied, or model-*predicted* data. Discrepancies between some characteristic of the observed and replicate data sets may be indicative of data-model misfit. The replicate posterior, or poster predictive distribution, is given as:

$$P(\mathbf{X}_{rep}|\mathbf{X}) = \int_{\mathbf{\Omega}} P(\mathbf{X}_{rep}|\mathbf{X},\mathbf{\Omega})P(\mathbf{\Omega}|\mathbf{X})d\mathbf{\Omega} = \int_{\mathbf{\Omega}} P(\mathbf{X}_{rep}|\mathbf{\Omega})P(\mathbf{\Omega}|\mathbf{X})d\mathbf{\Omega}, \qquad (25)$$

where $\mathbf{\Omega} = (\boldsymbol{\theta}_i, \boldsymbol{\omega}_j)$ denotes the full collection of model parameters, $P(\mathbf{\Omega}|\mathbf{X})$ is the posterior distribution for the unknown model parameters, and $\mathbf{X}_{rep}$ is a set of replicate data (Levy et al., 2009). Discrepancy measures, such as SGDDM, are used to assess the discrepancy between the data and the model. Large differences between the *realized* values of the chosen discrepancy measure, denoted $D(\mathbf{X}, \mathbf{\Omega})$, and the *model-implied* values, $D(\mathbf{X}_{rep}, \mathbf{\Omega})$ are a potential indicator of data-model

misfit. Gelman, Meng, and Stern (1996) recommended the use of a posterior predictive *p* value (PPP) to summarize information in PPMC. PPP represents the degree of overlap between the distribution of the discrepancy measure derived from the observed data and that of the replicate data. PPP is calculated as:

$$\text{PPP} = P(D(\mathbf{X}_{rep},\boldsymbol{\Omega}) \geq D(\mathbf{X},\boldsymbol{\Omega}) \mid \mathbf{X}). \quad (26)$$

PPP values near .5 indicate relative alignment of the distribution for the realized and model-implied discrepancy measures. Values near zero (or unity) indicate that the realized values are consistently much larger (smaller) than those stemming from the posterior predictive distribution. This indicates that the model is systematically underpredicting (overpredicting) the unknown quantities.

PPMC provides a flexible platform for assessing data-model fit and conducting model criticism. As Levy and colleagues (2009) point out, the PPMC framework may offer a number of advantages over other model checking approaches. Specifically, PPMC does not necessarily rely on asymptotic theory, nor does it rely on measures with known sampling distributions. Furthermore, as Rubin (1984) points out, simple summary statistics can be used to monitor data-model fit regardless of the complexity of the models themselves.

Summary

Large-scale testing scenarios have been shown to be conducive to the success of the DIMTEST procedure, as well as most other dimensionality assessment approaches. These conditions may not be feasible for all researchers, however. Under smaller-scale testing scenarios, power and Type I error rate may be compromised. Furthermore, DIMTEST places restrictions on the minimum

number of items that can be used and exhausts a portion of the examinee pool in conducting subtests partitioning. As much as 75% of the examinee responses may be needed for partitioning under certain conditions (Socha & DeMars, 2013). The SGDDM via PPMC approach places no such restrictions on the number of items and is able to utilize all available examinee responses in achieving the goal of assessing underlying dimensionality. Additionally, the Bayesian modeling paradigm employed under the PPMC framework may be able to improve decision accuracy under small-scale testing conditions, particularly when the researcher has meaningful *a priori* beliefs about the parameters of interest that they would like to encode in the estimation process. Finally, a fundamental difference exists between to two methods of interest in terms the way in which they approach dimensionality testing. DIMTEST is explicitly presented as a formal hypothesis test, whereas SGDDM is a diagnostic tool. Under a typical hypothesis testing approach, both the statistical significance, via a *p* value, and the effect size are of importance. Although attempts have been made to formulate an effect size for DIMTEST (Seo & Roussos, 2010), it is not often employed in practice. Without information about the effect size, researchers are often left to rely on only the significance to make judgments about dimensionality. As with most frequentist hypothesis tests, significance can almost always be achieved given a large enough sample. To this effect, previous research has demonstrated that DIMTEST exhibits inflated Type I error rates under conditions of very large samples, particularly when combined with short tests (Fay, 2012; Finch & Habing, 2007; Socha & DeMars, 2013). Comparatively, SGDDM does not approach the

assessment of dimensionality in a traditional hypothesis testing fashion. Rather, it offers the researcher a tool to diagnose the degree of data-model fit (misfit). For those wishing to conduct something akin to a hypothesis test, the posterior predictive *p* value (PPP) generated by the SGDDM framework can be compared to the desired level of α in the manner of the usual upper-tail test. The goal of this study was to compare the performance of the DIMTEST and SGDDM approaches to assessing dimensionality by simulating a variety of small, moderate, and large-scale testing conditions. Other statistics closely related to SGDDM, such as MBC and SMBC were not included in the current work as previous literature has suggested that those metrics may perform very similarly to GDDM and SGDDM. DIMTEST was chosen as a point of comparison over other dimensionality assessment methods (e.g., nonlinear factor analytic approaches such as NOHARM; Fraser & McDonald, 1988; McDonald, 1997) as (1) it, like SGDDM, is rooted in Conditional Covariance Theory, the dominant paradigm in IRT dimensionality assessment, and (2) it is the most widely used and accepted method of assessing deviations from unidimensionality within the CCT tradition.

Chapter 2

REVIEW OF LITERATURE

This chapter consists of two separate sections, one each for the DIMTEST and SGDDM approaches to assessing dimensionality. These two sections are each organized in a similar fashion. In each, the research related to the performance of the approach under small-scale (i.e., relatively few subjects and/or a small number of items) testing conditions is reviewed. Additionally, empirical examples of the effect of the structure underlying the item responses (e.g., unidimensional, simple structure, complex structure, etc.) on the performance of each metric are presented. Finally, any examples of work that has been undertaken to compare the performance of either DIMTEST or SGDDM with other approaches are discussed. Although the focus of the current work is on the most recent instantiation of each of the approaches of interest, research that was undertaken using previous versions (e.g., DIMTEST using AT2/FAC, GDDM, etc.) is also presented.

DIMTEST Performance

As was discussed in the previous chapter, the DIMTEST method has undergone a series of refinements since its initial development (see Nandakumar & Stout, 1993; Stout, 1987; Stout, Froelich, & Gao, 2001). At each stage of this development, research has been undertaken to evaluate the performance of the method under a variety of circumstances. The following sections review the work conducted at these stages, with particular focus being placed on research pertaining to conditions where the sample size, test length, and/or test structure have been manipulated. Brief discussion of literature that has sought to gauge the

performance of DIMTEST relative to other metrics is also presented where applicable.

Sample size, test length, test structure and DIMTEST. A variety of studies have been undertaken to evaluate the performance of the DIMTEST method via Monte Carlo simulation. In many of these simulations, sample size, test length, and/or test structure were included as factors hypothesized to influence said performance. Following Stout's (1987) seminal work, Nandakumar and Stout (1993) offered refinements of the original procedure aimed at improving Type I error rates and power in cases of difficult items (large discrimination values), and the presence of guessing, as well as in automating the selection of the number of items to be included in the AT1 and AT2 partitions. These refinements were found to yield Type I error rates closer to the nominal value of $\alpha = .05$, as well as higher statistical power than were observed in Stout (1987) using the original procedure. Acceptable performance was noted in conditions with as little as 750 participants and 25 items, though power was found to suffer (i.e., observed power was $\leq .80$) under conditions combining low sample size with short tests as the correlation between dimensions increased from $\rho = .5$ to $\rho = .7$. Only one structure was used in the investigation. That structure consisted of two dimensions (θ_1, θ_2) with approximately one-third of items representing only θ_1, one-third representing θ_2, and the remaining items representing a mix of both dimensions.

Similar recommendations as to appropriate sample sizes and test lengths for use with the AT2/FAC instantiations of DIMTEST were put forth by Gessaroli

and De Champlain (1996). Their primary interest was in comparing the performance of DIMTEST with the performance of their $\chi^2_{G/D}$ statistic, a statistic that would also be used as a point of comparison later on by Levy and Svetina (2011) in their paper on the performance of GDDM. Gessaroli and De Champlain's results indicated that the DIMTEST with AT2/FAC procedure yielded acceptable Type I error rates under conditions where the sample size was at least $N = 1{,}000$ and there were at least 30 items. Performance tended to suffer with smaller samples or shorter tests, particularly when the item discriminations were "weak" or "moderate," which were defined as $\alpha_j \sim N(.72, .25)$ and $\alpha_j \sim N(1.07, .40)$, respectively. Power under the DIMTEST conditions tended to be most affected by test structure, represented, in this case, by the dominance of the second test dimension. Power tended to suffer the most with short tests (15 items[1], in this case) and a less influential second dimension (80% of items representing θ_1 and 20% representing θ_2). Increases in test length, the strength of θ_2, or the magnitude of the item discrimination values tended to result in power being increased beyond commonly accepted levels ($\geq .80$).

The second notable revision to the DIMTEST procedure was introduced by Stout, Froelich, and Gao (2001). This revision removed the need for the AT2 partition used to correct the positive bias in the DIMTEST statistic through a resampling procedure. This change effectively increased the proportion of available items that can be allocated to the AT1 (now just AT) and PT partitions

[1] The minimum number of items allowed by the commercially available version(s) of DIMTEST using ATFIND for subtest selection is 19. It is not clear how Gessaroli and De Champlain (1996), and later Finch and Habing (2007) were able to use only 15 items in their research.

and, thus, used for calculating the DIMTEST statistic. The same factor-analytic approach used in previous versions to choose these partitions was retained. Results from simulation studies presented in the same paper suggest Type I error rates near the nominal value of $\alpha = .05$ for tests of with 40 and 80 items and sample sizes of at least $N = 750$. Inflated Type I error rates (as high as 22 rejections out of 100) were noted using a test with only 25 items and an AT partition consisting of eight items. This may have been due to an insufficient amount of information being available in the PT partition, which is used in the bias-correcting resampling procedure. Exceptional power (.98 to 1.0) was noted in all conditions. While the results obtained by Stout et al. were not that dissimilar from those observed in studies using previous versions of DIMTEST (Nadakumar & Stout, 1993, for example), the removal of the need for a third item partition (AT2) provided additional flexibility, particularly in instances where defining two homogenous subsets of the original item pool proves difficult, such as with polytomously scored items.

The current version of DIMTEST employs the bootstrapping method for bias-correction and uses ATFIND which, as was discussed in the previous chapter, combines HCA/CCPROX and DETECT to select the items for the AT partition. The transition away from the factor-analytic (FAC) selection method was recommended by Froelich and Stout (2003), who found that the approach struggled to select an adequate group of dimensionally similar items for the AT partition, particularly when tests deviated from simple structure and when the correlation between dimensions were high ($\rho = .7$ was the largest correlation

investigated). Finch and Habing (2007) compared the most recent version of DIMTEST to NOHARM-based statistics (McDonald, 1967) designed to test the unidimensionality assumption. Their results indicated that DIMTEST may yield sufficient Type I error rates and power with as few as 15 items even when the correlation between dimensions is as high as 0.80. It is important to note, however, that the smallest sample size condition that they examined was 1,000 examinees.

Finally, Fay (2012) sought, amongst other things, to reexamine the sample size and test length minima put forth by earlier researchers (for examples, see Gessaroli & De Champlain, 1996; Pyo, 2000) and update them in light of the most recent version of the method. His findings suggested that DIMTEST may be able to maintain acceptable Type I error rates, if not rates slightly below the nominal value of $\alpha = .05$, with as few as 21 items and 250 participants. Fay's power analysis yielded statistical power in excess of 0.80 using DIMTEST with at least 27 items and 500 participants when the tests consisted mainly of highly discriminating items. Decreases in power were noted when increasing the correlation between dimensions beyond $\rho = .35$, decreasing the strength of the item discrimination parameters, or increasing the complexity of the test's structure. Fay recommended that, in absence of any *a priori* information about the features of the test, a conservative approach may be to use at least 750 examinees and 33 items.

SGDDM Performance

As the GDDM and SGDDM statistics constitute a relatively new line of research, very little literature exists that has investigated their performance in light

of test characteristics or as compared to other approaches. The existing body of literature, albeit sparse, is discussed in the following section. As was the case with the DIMTEST discussion, attention is given to the effects of sample size, test structure, and test length. A brief summary of the performance of SGDDM relative to other statistics is also given. Finally, some mention is made of the performance of Reckase's (1997) MBC metric, as it can be considered a related precursor to SGDDM.

Sample size, test length, test structure and SGDDM. As was mentioned in the previous chapter the GDDM statistic can be considered an extension or Reckase's (1997) model-based covariance (MBC) approach, in that it aims to capture the average absolute model-based covariance amongst item pairs. Levy, Mislevy, and Sinharay (2009) compared a variety of unidimensionality discrepancy approaches using a posterior predictive model checking (PPMC) approach. Their results indicated that MBC, as well as the related Q_3 (Yen, 1984), tends to be both uniformly distributed and exhibit near-nominal Type I error rates under null (i.e., unidimensional) conditions. These approaches also exhibited acceptable statistical power under multidimensional conditions. Not unlike work done using DIMTEST, their study put forth sample size, the correlation between dimensions, the magnitude of item discrimination parameters, and the number of items as factors that may influence the performance of any method aimed at assessing dimensionality.

Building on earlier work, Levy and Svetina (2011) presented the GDDM statistic. They compared GDDM to other approaches empirically by generating

data for 1,000 examinees from a 36-item, two-dimensional test exhibiting simple structure (dubbed "M0"), a two-dimensional test exhibiting approximate simple structure, and a variety of three-dimensional tests and fitting these datasets to M0. Using PPMC, they showed that GDDM was able to maintain acceptable Type I error rates when fitting these two- and three-dimensional models in all conditions other than those where a large proportion of the items represented multiple dimensions (25% of the items loaded on θ_1, 25% on θ_2, 25% on θ_1/θ_3, and 25% on θ_2/θ_3) and the correlation between these dimensions was high ($\rho = 0.5$). Type I error rates exceed the nominal rates by roughly 60% under this condition (e.g., .08 vs. $\alpha = .05$). Power was generally high ($\geq .68$) when fitting M0 to data generated from a model where a strong auxiliary dimension was present, but decreased notably when fitting M0 to data generated from models where either a very subtle third dimension was present or a small number of items cross-loaded on more than one dimension. This decrease in power was exacerbated by increases in the correlation between dimensions. In all conditions, however, GDDM did exhibit sensitivity to data-model misfit. It is important to note that the authors did not frame their discussion in a hypothesis-testing context, but rather as a diagnostic approach. The current work adopts a hypothesis testing approach to facilitate comparisons with DIMTEST, which aims to conduct a formal test of the null hypothesis of unidimensionality.

The standardized version of the GDDM statistic, SGDDM, was put forth by Levy, Xu, Yel, and Svetina (2012), and was built as an extension to the standardized model-based covariance (SMBC). Similar to the work of Levy and

Svetina (2011), SGDDM was tested empirically by generating data for 1,000 examinees on a series of 36-item tests. These tests were modeled as being either one-. two-, or three-dimensional and exhibited a variety of structures, including simple, complex, and testlet. The testlet exams contained primarily items that represented a single dimension, but with a small subset that represented multiple dimensions (see Rijmen, 2010 for further discussion). In all, seven separate test structure conditions were examined. As was the case with GDDM, SGDDM demonstrated sensitivity to both data-model fit and misfit. The proportion of extreme PPP values observed when fitting the correct model, akin to Type I error rate, ranged from .00 to .04, while the proportion of extreme PPP values observed when fitting a misspecified model, akin to statistical power, never deviated from 1.0. Sample size and test length were not manipulated in either Levy et al. (2012) or Levy and Svetina (2011), nor was the assessment of deviations from unidimensionality a central focus, thus providing the impetus for the current work.

Chapter 3

METHODOLOGY

The goal of this Monte Carlo simulation is to compare the performance of the DIMTEST and SGDDM statistics under both small and large-scale testing conditions. Factors previously shown to impact the performance of these statistics were manipulated. Those factors included sample size, test length, and dimensionality. Investigating the effect of dimensional correlation and simple versus complex structure was not a goal central to the study and, thus, those factors were held constant.

Data Generation

Item response data was generated via the two-parameter form of *Equation 4* (where $c_j = 0$) using R version 2.15.2 (R Team, 2008). The data was generated such that $X_{ij} = 1$ indicated a correct response and $X_{ij} = 0$ indicated an incorrect response. Under conditions where unidimensional data were modeled, the discrimination parameters along the second dimension (α_{j2}) were set to zero, as were the correlation between the dimensions.

Determination of sample size and test length values. An exploratory approach to determining the most informative sample size (N) and test length (J) values was used. This approach held three goals: (1) to find a combination of N and J where both the DIMTEST and SGDDM approaches exhibited satisfactory performance in terms of their ability to assess deviations from unidimensionality, (2) to find a combination of N and J where one approach clearly outperformed the other (if possible), and (3) to find a combination of N and J where neither

approach performed well. As will be discussed in the results section, a condition wherein $N = 750$ and $J = 24$ was used as a starting point and further conditions were defined based on the results of that, as well as subsequent conditions. Given that the minimum number of items allowed by the DIMTEST program is 19, only the performance of the SGDDM statistic will be investigated under any conditions that end up containing less than 19 items. The values used in the initial condition were chosen to be something of a combination of the minimum recommended values for use with DIMTEST (Fay, 2012) and the smallest values used in previous SGDDM analyses (Levy & Svetina, 2011; Levy et al., 2012).

Generation of person parameters. For each replication, person parameters were generated from a bivariate normal distribution such that $\boldsymbol{\theta} = (\theta_1, \theta_2) \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a variance-covariance matrix. The variance for each factor was set to one, while the off-diagonal elements of $\boldsymbol{\Sigma}$, representing correlations between factors, were set to a fixed value of $\rho = 0.3$, indicating a moderate relationship.

Generation of item parameters. All discrimination parameters were generated from a random truncated normal distribution for each replication within a condition such that $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_J) \sim N_{(0, \infty)}(1, 0.2)$. Item difficulty parameters (δ_j) were randomly generated from a normal distribution for each replication within a condition such that $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_J) \sim N(0, 0.7)$. Syntax for the generation of both item and person parameters is presented in Appendix A.

Strength of dependence of the item responses on the underlying dimensions may be an influencing factor on the performance of the procedures

relevant to this study (Levy, Mislevy, & Sinharay, 2009). Investigating that nature of that influence is not, however a goal of the current work and, as such, no conditions of systematically differing discrimination parameters (α_j) were specified. The values employed for these parameters in the current work represent what are often considered "moderate" choices in such studies. This was done to facilitate comparison to previous literature.

Test structure. Degree of multidimensionality in the form of the proportion of items representing an auxiliary dimension was manipulated via the method put forward by Froelich and Habing (2008). Under this approach, a 1 x *J* vector of item discrimination parameters is first generated via the mechanism described in the previous paragraph. The values in this vector are then transformed into a 2 x *J* matrix via the equations:

$$\alpha_{j1} = \alpha_j \cos(\beta_j) \text{ and } \alpha_{j2} = \alpha_j \sin(\beta_j) \quad (25)$$

where α_j is the initial parameter value, α_{j1} and α_{j2} are the resulting discrimination parameter values, and β_j is the angle between the item's direction of best measurement and the θ_1 axis. The current study included two test structure conditions: unidimensional and two-dimensional simple structure. Under the unidimensional condition, β_j was set to 0 (measuring θ_1 only), while the multidimensional simple structure condition included two-thirds of items with $\beta_j = 0$ (measuring θ_1 only), and one-third with $\beta_j = 90$ (measuring θ_2 only).

Number of replications. Due to computational limitations, the number of replications differed between the DIMTEST and SGDDM conditions. A variety of existing simulation studies have investigated the performance of DIMTEST

under conditions similar to those to be undertaken in the current work. The number of replications used in those studies has generally ranged from as low as 100 (e.g., Froelich & Habing, 2008; Nandakumar & Stout, 1993; Stout, Froelich, & Gao, 2001) to upwards of 500 (e.g., Fay, 2012; Finch & Habing, 2007). To be conservative, 1,000 replicate data sets were generated and analyzed for all conditions involving the use of DIMTEST.

As MCMC estimation via Gibbs sampling can be computationally intensive, a smaller number of replications are generally used in simulation studies employing these methods. Previous studies investigating the performance of PPMC and the GDDM/SGDDM statistics in the context of IRT have typically used 50 replications (e.g., Levy, Mislevy, & Sinharay, 2009; Levy & Svetina, 2011; Levy et al., 2012). In keeping with this work, the current study sought to use 50 replications for each condition as well. These replications were randomly selected from the 1,000 used in the DIMTEST conditions.

DIMTEST

DIMTEST version 2.1 (Nandakumar & Stout, 1993; Stout, 1987; Stout, Froelich, & Gao, 2001) was used to conduct the analyses of interest in the current study. All analyses were exploratory in nature and, as such, one-third of the sample in each condition was used to establish the AT partition, with the remaining two-thirds being used to calculate the DIMTEST statistic. As no accommodations were being made for the possibility of guessing on the part of the examinee in the data generation process, the *c*-parameter was set equal to zero for all conditions. The remaining parameters, specifically those pertaining to the

bias-correcting bootstrap procedure (seed number, evaluation points, and bootstrap replications) were left at their default values.

SGDDM

Mplus version 6.12 (Muthén & Muthén, 2007) was used to conduct MCMC estimation via Gibbs sampling under the SGDDM conditions. Three separate MCMC chains were used using software-supplied starting values. Convergence of the chains was monitored using the Brooks-Gelman-Rubin (BGR) diagnostic (Brooks & Gelman, 1998). The first five replications in each SGDDM condition were used to assess convergence. Once the number of MCMC iterations needed for the chains to converge had been determined, that value was then applied to the remaining replications. A BGR value of less than 1.05 was to be considered sufficient for chain convergence. Trace plots were also monitored to ensure sufficient mixing of the draws from the chains. Any evidence of serial dependence (i.e., non-trivial autocorrelations) was handled by thinning these draws. R version 2.15.2 was used to generate replicate datasets from the Mplus-generated MCMC draws, conduct thinning of the draws if necessary, and compute the SGDDM statistic presented in *Equation 22*.

Parameters. A variety of parameters settings can be customized when using Bayesian estimation and conducting posterior predictive model checking in Mplus. For the current study, initial values for the loadings and thresholds (analogous to item discrimination and difficulty parameters, respectively) for each of the indicators (J item response vectors, where J is the number of items) were supplied by the software. The posterior mean was used as a summary of the

posterior distribution. Finally, 300 model-implied replicate datasets were generated in R using the draws from the posterior distribution for use in PPMC procedures. All other settings were left at their default value (see Appendix B for sample syntax).

Prior distributions for model parameters. Prior distributions for the unknown model parameters need to be specified for each condition in which Bayesian parameter estimation will be used. In the context of the current study, these include the person (θ_{im}), item location (δ_j), and item discrimination (α_j) parameters. The latent person ability parameters were assigned standard normal prior distributions

$$\theta_{im} \sim N(0, 1).$$

The item location parameters were assigned diffuse normal prior distributions

$$\delta_j \sim N(0, 10).$$

As the current study was focused on assessing deviations from unidimensionality, the models being fit to the data assumed a single vector of item discrimination parameters, even though the data, under certain conditions, was generated using test structures that assume a 2 x J matrix of discrimination parameters. The values in that vector were assigned diffuse normal distributions censored with a lower-bound of zero

$$\alpha_j \sim N_{(0, \infty)}(1, 10).$$

Data Analysis

Figure 5 presents an overview of the DIMTEST and SGDDM simulation procedures. Empirical Type I error rates (α) and statistical power ($1 - \beta$, where β

is the Type II error rate) were used to assess performance under the unidimensional and multidimensional data conditions, respectively. For the DIMTEST condition, Type I error is defined as the proportion of analyses in which the null hypothesis (H_0) of unidimensionality is rejected when the data were generated via a unidimensional model. Power, conversely, is defined as the proportion of replications in which H_0 was rejected when the data were generated via a multidimensional model. Under the SGDDM conditions, H_0 is considered rejected when extreme values of the posterior predictive *p* value (PPP) are observed such that PPP $\geq (1 - \alpha)$, where α is equal to an acceptable rate of Type I error determined *a priori*. As was discussed earlier, the PPP values for each replication in this study were derived from a comparison of the SGDDM value based on a replicate data set with the SGDDM value based on a set of parameter draws from the posterior distributions based on the observed data. Each of these PPP values was based on 300 observations (i.e., 300 replicate data sets and 300 draws from the posterior distributions of the model parameters). This criterion represents, essentially, a one-tailed hypothesis test. The use of a one-tailed test is reasonable in this context as the SGDDM statistic is constructed as a non-directional measure reflecting the magnitude of the unmodeled associations. The empirical Type I error rate was compared to the nominal rate of $\alpha = .05$, which represents a standard commonly applied in social science research.

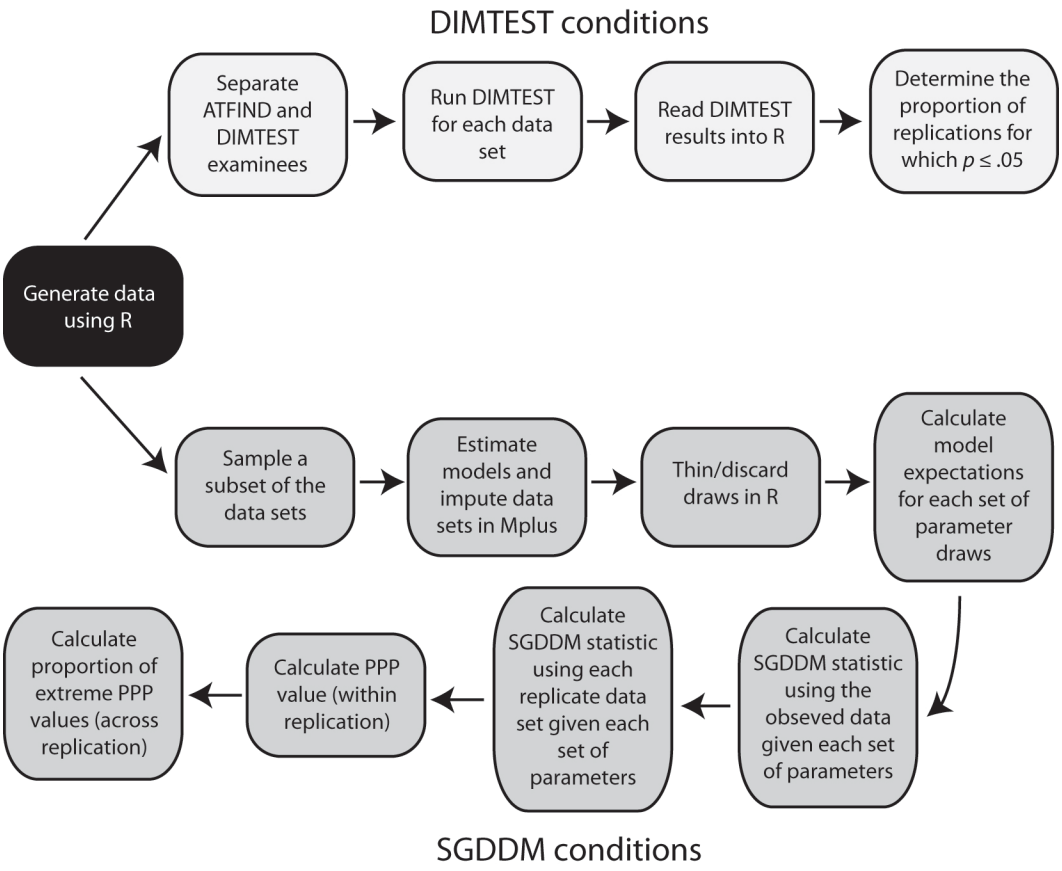

Figure 5. Flow chart of the DIMTEST and SGDDM simulation study procedures.

Hypotheses

Previous research in the realm of assessing a unidimensionality hypothesis, as well as features inherent to conditional covariance theory, suggest a number of factors that may influence performance. The hypotheses related to the variables of interest with respect to the relative performance of the DIMTEST and SGDDM approaches are presented below.

Sample size and test length. Neither sample size nor test length was hypothesized to influence performance under true unidimensional conditions for any of the values of these two variables utilized in the current study. Under conditions where the true structure was multidimensional, increases in either

sample size or the number of items should have resulted in higher statistical power.

Test structure and correlation between dimensions. The correlation between dimensions and the structure of the test were not manipulated in the current work. The constant dimensional correlation of $\rho = 0.3$ was not hypothesized to yield any differential effect on performance for any of the levels of the manipulated variables. All of the multidimensional conditions investigated utilized simple structure, wherein all items measured a single dimension. The proportion of items measuring θ_1 and θ_2 was held constant across experimental conditions. The use of simple structure should have yielded greater power than would be expected had more complex structures been applied.

For some researchers, thinking of models such as the two-dimensional type used in the current work as following a bifactor structure may be preferable. In a bifactor model the entire set of items are thought of as having some underlying trait in common, often termed a *general* factor, with a particular subset, or subsets of items sharing additional traits (*specific* factors). Figure 6 presents a depiction of a two-dimensional simple structure model with twelve items, while Figure 7 shows the same model as following a bifactor structure. The model in Figure 6 follows the structure of the data generation model used in the current work under the multidimensional conditions. These two representations are hierarchically related, in that constraining the loadings on the general factor, θ_g, to zero and freeing the factor correlation between θ_1 and θ_2 yields the model shown in Figure 6, and the parameter values from one can be translated to their

equivalent in the other using methods such as those discussed by Rijmen (2010) and Yung, Thissen, and McLeod (1999). Appendix C presents such a translation for a 12-item test using similar values for the item discrimination and factor correlation parameters as are used in the current work.

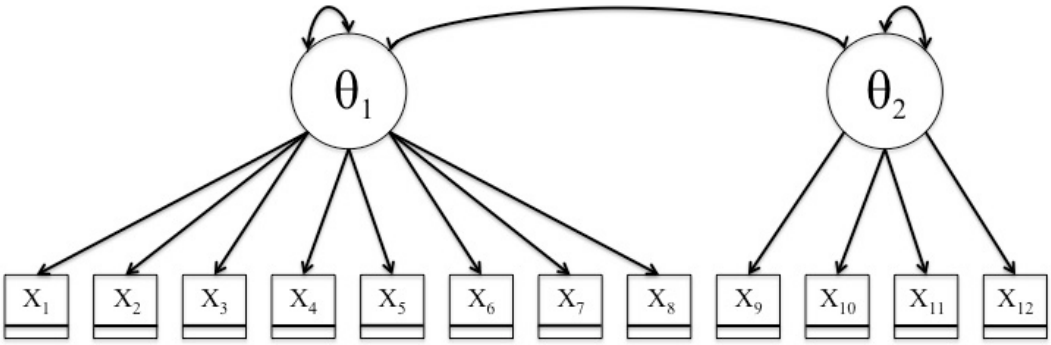

Figure 6. Representation of a 12-item exam following a two-dimensional structure.

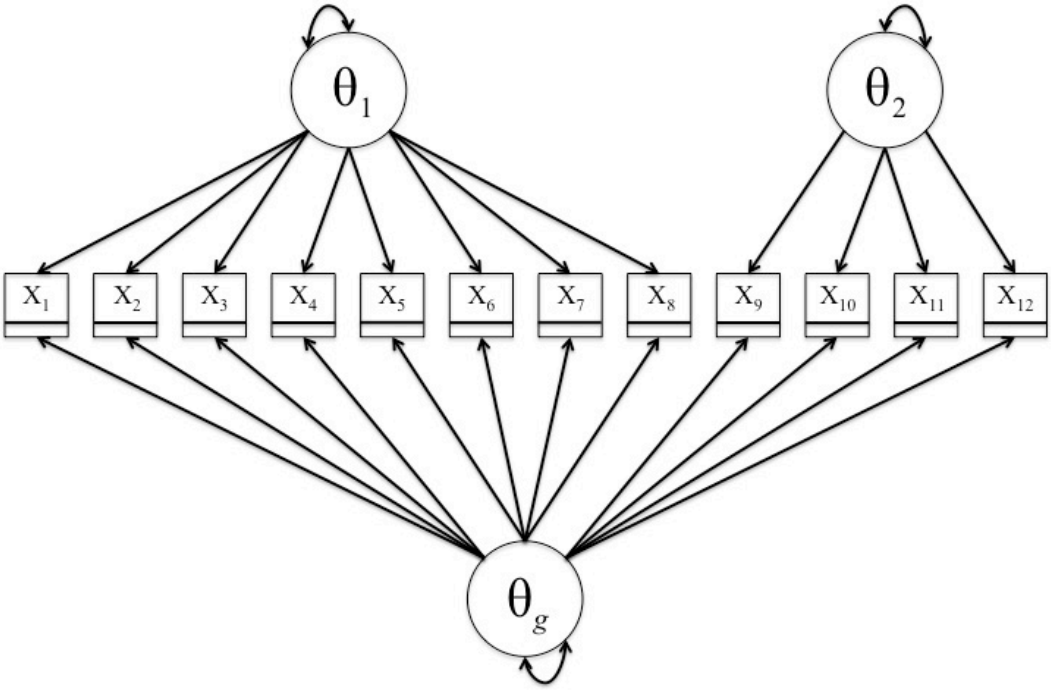

Figure 7. Representation of a 12-item exam following a bifactor structure.

DIMTEST vs. SGDDM. Both approaches were hypothesized to perform comparably, as well as favorably under conditions consisting of large sample sizes and relatively long tests. DIMTEST may be overly conservative in terms of Type I error rate under null conditions (Fay, 2012) whereas SGDDM has been shown to yield rejections at a near-nominal rate (Levy & Svetina, 2011; Levy et al., 2012). Decreases in power should have been exhibited under both methods as sample size and test length decreased. DIMTEST was unable to function when $J < 19$, therefore no comparison with SGDDM was possible under such conditions. DIMTEST has been shown to exhibit relatively poor power when short tests ($J \leq 21$) are combined with small sample sizes ($N = 250$) and moderately discriminating items (Fay, 2012). However SGDDM has not yet been evaluated under such conditions, therefore no *a priori* hypotheses were made with respect to comparative performance under these types of small-scale testing circumstances.

Summary

This chapter reviewed the procedures to be used in generating data from two separate test structures, conducting analyses using both the DIMTEST and SGDDM statistics, and evaluating the performance of those statistics. Manipulated variables included test structure, sample size, and test length. All other factors were either assigned fixed values (e.g., correlation between test dimensions), or given randomly generated values (e.g., person ability, item discrimination, and item location parameters). The number of replications used varied by condition with 1,000 replications being used to evaluate the DIMTEST statistic and 50 replications being used in the SGDDM conditions. Empirical Type

I rate and power were defined as the criteria by which the two statistics of interest will be evaluated.

Chapter 4

RESULTS

The goal of this Monte Carlo simulation was to compare the performance of the DIMTEST and SGDDM statistics under both small and large-scale testing conditions.

Exploratory Condition Search

An exploratory approach was used to determine the parameters (i.e., sample size and test length) of the experimental condition to be used. As was discussed in the previous chapter, this approach held three goals: (1) to find a combination of *N* and *J* where both the DIMTEST and SGDDM approaches exhibited satisfactory performance in terms of their ability to assess deviations from unidimensionality, (2) to find a combination of *N* and *J* where one approach clearly outperformed the other (if possible), and (3) to find a combination of *N* and *J* where neither approach performed well. The search began with an "anchor" condition chosen to represent a combination of parameters that was both of a smaller-scale than what is typically used in research concerning DIMTEST and SGDDM, yet for which both procedures might be expected to perform satisfactorily, thus satisfying goal (1) from above. This anchor condition utilized a sample size of 750 examinees and a test length of 24 items. As can be seen in Tables 1 and 2, and as will be discussed in a later section, both approaches yielded acceptable performance both when fitting the correct and incorrect models. It was determined that any conditions for which both *N* and *J* exceeded these initial values (750 and 24, respectively) would not be essential to the goals of the

study. One such condition was, however, added later in the process. Based on these initial results further conditions where $N = 250$ and $J = 12$ were added. Results of these conditions again proved somewhat similar where comparable, thus an $N = 100$ condition was added. This final sample size condition satisfied goal (3) from above, in that both approaches exhibited less-than-satisfactory performance under one of the $N = 100$ conditions. A condition where $J = 18$ was added in order to satisfy goal (2). Finally, a $J = 30$ condition was added as well to provide a second test length where SGDDM and DIMTEST could be compared.

In summary, three sample size conditions ($N = 100, 250, 750$) and four test length conditions ($J = 12, 18, 24, 30$) were used in the current work. This resulted in 24 total experimental conditions where SGDDM could be evaluated and 12 for DIMTEST. These 36 conditions were evaluated using both a properly specified, as well as a misspecified model, resulting in 72 total experimental conditions.

MCMC Estimation Parameters

The current work required that the number of burn-in iterations, as well as a thinning factor be determined for each experimental condition. The total number of iterations for each of the three MCMC chains also needed to be specified, and was calculated as

$$TI = (100 \times T) + B \ , \tag{27}$$

where TI stands for total iterations, T is the thinning factor, and B is the number of burn-in iterations required. The value of 100 represents the number of usable sets of parameter values needed from each chain in order to have the desired 300 total

sets for use in PPMC. For the current study, the number of burn-in iterations ranged from 1,000 to 4,000, with the largest values being seen in the conditions where model estimation proved most difficult (e.g., small samples combined with short tests and model misfit). A thinning factor of 30 was used for all conditions except for use of a factor of 50 when $N = 100$ and $J = 12$ or 24. The total iterations ranged from 4,000 to 9,000.

Unidimensional Data Conditions: Type I Error Rates

Table 1 below presents results obtained under the 36 conditions where the correct model was fit. Figure 8 below presents the same results in a graphical format. Each panel in Figure 8 corresponds to one combination of sample size and test length. The sample size values are indicated in the bar at the top of each panel while the test lengths are denoted by the four hash marks within each panel. The proportion of p or PPP values at or below .05 appears on the vertical axis. The dashed line cutting across each panel indicates the commonly used $\alpha = .05$ nominal rate. The results of the DIMTEST and SGDDM conditions are presented separately and are indicated by the solid lines within each panel. Results for the DIMTEST conditions are marked with a "+" sign, while SGDDM results are marked with a "Δ." The results obtained for the $J = 24$ and $J = 30$ conditions allow for direct comparison of the two procedures as the analyses for each were conducted using common data sets, although the number of replications differed. As the commercially available version of DIMTEST using ATFIND does not allow for less than 19 items, no results are presented using that approach for the J

=18 and $J = 12$ conditions. A not applicable ("NA") indicator is used to indicate these conditions in Table 1.

Table 1

Proportion of Extreme PPP/p Values when Using Data from a Unidimensional Model in the Analysis Conditions.

		PPMC using SGDDM	DIMTEST
Sample Size (N)	Test Length (J)	Proportion of PPP ≤ .05 (50 replications)	Proportion of $p \leq .05$ (1,000 replications)
100	12	**.060**	N/A
	18	.040	N/A
	24	.020	.025
	30	.040	.018
250	12	.020	N/A
	18	.040	N/A
	24	.040	.014
	30	.020	.008
750	12	.000	N/A
	18	.020	N/A
	24	.020	.011
	30	**.080**	.010

Note. Values greater than .05 are indicated in bold.

The DIMTEST approach tended to be conservative under unidimensional conditions, that is to say that the Type I error rate tended to be below, and, in some conditions, well below the nominal value of $\alpha = .05$. Slight increases in Type I error rate were seen when moving from tests with 30 items to tests with 24 items or when decreasing sample size, but still remained below the nominal level.

The proportion of extreme PPP values under the SGDDM conditions tended to be much closer to the nominal rate than was seen with DIMTEST under

most conditions. These rates exhibited less stability than those resulting from the use of DIMTEST due to the substantially smaller number of replications used. The results seen at the limits of the experimental conditions, in particular, stand out. The proportion of extreme PPP values when $N = 100$ and $J = 12$ was slightly above the nominal rate at .06.

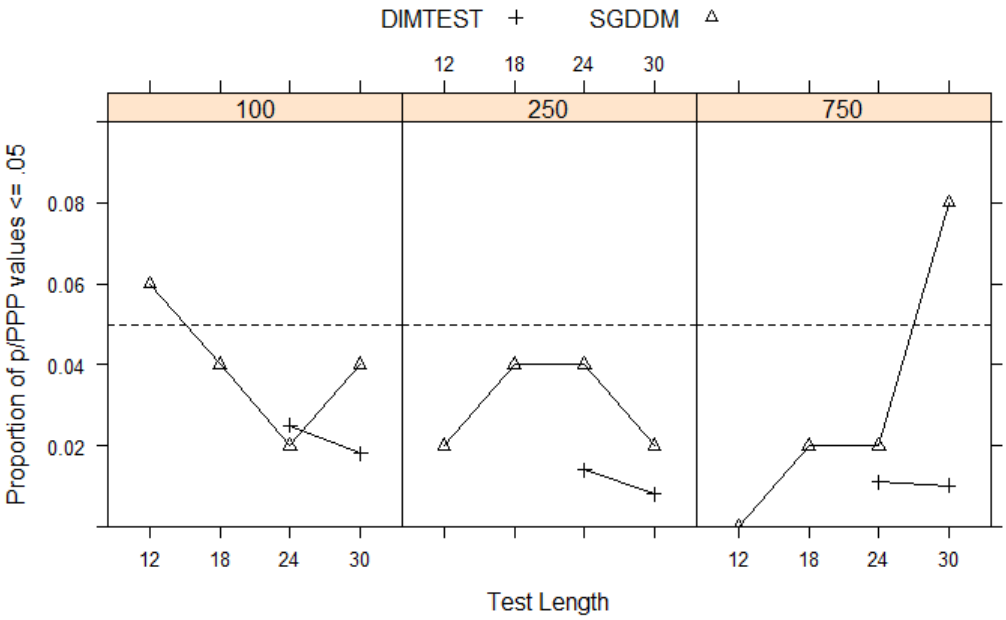

Figure 8. Results of the analysis conditions where the correct model was fit.

Multidimensional Data Conditions: Estimation of Power

Table 2 in below, as well as Figure 9 below present results obtained under the 36 conditions where a misspecified model was fit. The presentation of the panels is Figure 9 is the same as that of Figure 8 above with the exception being that the dashed line cutting across the panels now indicates a rejection rate of .80, a commonly applied criterion in power analyses. It should be noted that some of the SGDDM conditions used less than the intended 50 replications due to model estimation issues in Mplus. The actual number of replications used in these

conditions is noted under Table 2. These issues were largely a result of the way in which the prior distributions for the item discrimination parameters are handled in the Mplus language. Specifically, Mplus does not allow for the explicit use of a censored prior distribution. Instead, the researcher must add constraints using the "MODEL CONSTRAINT" subcommand (see Appendix B). In the case of the current work, this entailed constraining the estimates to be positive. An unfortunate side effect of this is that the software will occasionally fail to resolve the parameter estimates when the posterior distribution for one or more of the parameters has a lot of its mass concentrated near zero. While these issues may have increased the impact of any chance characteristics of the data, they are unlikely to have affected the overall pattern of results. Other software options exist for conducting PPMC (e.g., WinBUGS; Lunn, Thomas, Best, & Spiegelhalter, 2000), though these options tend to require more time and user input for automation.

Table 2

Proportion of Extreme PPP/p Values when Using Data from a Two-dimensional Simple Structure Model in the Analysis Conditions.

Sample Size (N)	Test Length (J)	PPMC using SGDDM Proportion of PPP ≤ .05 (50 replications)	DIMTEST Proportion of $p \leq .05$ (1,000 replications)
100	12	**.684**[a]	N/A
	18	.917[b]	N/A
	24	1.00[c]	**.729**
	30	.976[d]	**.778**
250	12	.980	N/A
	18	1.00	N/A
	24	1.00	.998
	30	1.00	1.00
750	12	1.00	N/A
	18	1.00	N/A
	24	1.00	1.00
	30	1.00	1.00

Note. Results based on 38[a], 36[b], or 41[c, d] replications. Values less than .80 are indicated in bold.

Both approaches fared well regardless of test length when $N = 250$ or 750; the proportion of extreme p/PPP values seen under these conditions were well above .80. A ceiling effect at 1.0 made the results of the two approaches almost indistinguishable under conditions where a direct comparison was possible (i.e., $J \geq 19$). The performance of the two approaches diverged when the sample size was lowered to $N = 100$. DIMTEST exhibited power below .80 under both test length conditions under which it was examined. Furthermore, a downward trend was present when the number of items was decreased from $J = 30$ to $J = 24$. The

SGDDM statistic, conversely, demonstrated satisfactory performance (proportion of extreme PPP values ≥ .80) under both of the aforementioned test length conditions, as well as in the $J = 18$ condition. It was only under the $J = 12$ condition that the proportion of extreme PPP values fell below the acceptable criterion. It is of note that, though power, from a null hypothesis testing perspective, was only .68 under this condition, the mean PPP value for the 38 replications was .047. From a diagnostic perspective, a PPP value such as this might be construed as evidence of data-model misfit by some researchers.

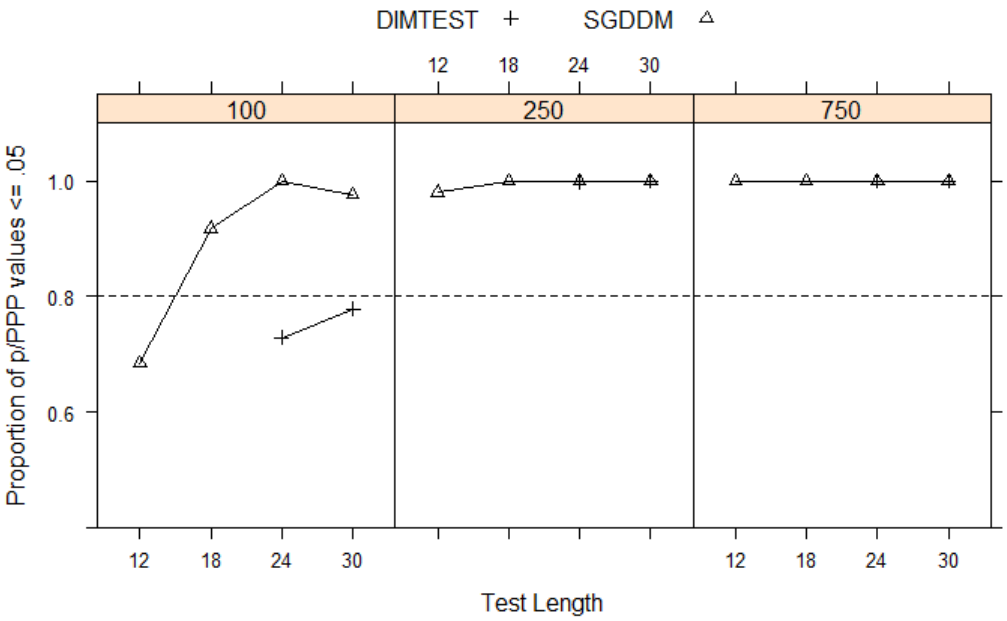

Figure 9. Results of the analysis conditions where a misspecified model was fit.

Summary

This chapter has presented results for the use of unidimensional and multidimensional data in the analysis conditions. Type I error rates, power, and proportion of extreme PPP values were the primary foci of the current work. In general, the results suggest that both approaches are capable of identifying data-

model fit and misfit in all but the most extreme cases when using the data structures employed in the current work (i.e., simple structure with moderate item discrimination and factor correlation parameters).

Chapter 5

DISCUSSION AND CONCLUSIONS

Each of the manipulated variables had demonstrated an impact on Type I error rate and power. These outcomes are discussed separately for the unidimensional and multidimensional conditions. Recommendations for practitioners and researchers are provided in light of the findings. Finally, the limitations of the current study, as well as suggestions for further research, are presented.

Interpretation of Results

Unidimensional conditions. Neither of the manipulated factors (sample size and test length) were hypothesized to impact the performance of either SGDDM or DIMTEST under unidimensional conditions with the exception of DIMTEST not being available when $J \leq 19$. This hypothesis was largely supported by the results in Table 1. The proportion of extreme *p*/PPP values were below the nominal rate of $\alpha = .05$ for most every combination of sample size and test length. Many of these values under the SGDDM conditions, however, were too close to α to conclude that the "true" rate is significantly different from .05, given the relatively small and finite number of replications used. Only two of 18 conditions yielded rates of extreme results greater than .05: SGDDM when $N = 100/J = 12$ (.06) and SGDDM when $N = 750/J = 30$ (.08). The former could theoretically represent a real effect indicative of the limits of the SGDDM statistic to capture data-model fit, given the sample size, test length, and structure used in that condition. Further work using smaller sample sizes and/or shorter tests could

be conducted to evaluate this effect. The latter result, on the other hand, was likely a result of sampling variability, as no theoretical reason exists to suggest inflation of this value given the conditions. Previous studies involving SGDDM have used larger sample sizes and longer tests than the current work, yet have exhibited Type I error rates at or below the nominal rate. Repeating this condition using more replications could potentially correct this issue. These results were in keeping with previous studies that used comparable sets of conditions (e.g., Fay, 2012; Levy et al., 2012).

In general, both approaches seem well suited to indicating data-model fit (rejecting the null hypothesis, in the case of DIMTEST) when the data follow unidimensional structure, given the conditions used in the current study. If anything, both approaches err on the side of caution in that they tend to be slightly conservative from a hypothesis testing perspective, with DIMTEST being the more conservative of the two. This notion is in line with the *a priori* hypotheses presented in Chapter 3. Of note is the fact that the rejection rates for the DIMTEST conditions did not seem to approach the nominal rate as sample size and test length were increased as one would have expected based on theory (Stout, 1987).

Multidimensional conditions. As was hypothesized, evidence for data-model misfit, or multidimensionality in this case, became more apparent with larger samples and/or longer tests (Table 2). These findings are consistent with previous research using DIMTEST (Fay, 2012; Finch & Habing, 2007; Froelich & Habing, 2008) and SGDDM (Levy & Svetina, 2011; Levy et al., 2012). They

are also consistent with Zhang and Stout's (1999a) conditional covariance theory. Both approaches exhibited satisfactory (i.e., ≥ .80) power for all test lengths when a sample size of at least 250 was used. Increases in test length for conditions with $N = 250$ or 750 were met with little to no additional effect due to a ceiling of 1.0 on power. This effect may be mitigated in future research my decreasing the nominal rate to $\alpha = .01$. It was only under conditions where $N = 100$ that a difference in results between the two approaches was noted. The SGDDM statistic demonstrated a greater ability to assess data-model misfit under conditions where the two approaches could be compared. Potential sources of DIMTEST's relatively poor performance when using such a small sample size could be the way in which the software handles AT/PT partitioning as well the procedure's use of total score subgroups as a means of conditioning on examinee ability. As has been previously mentioned, DIMTEST exhausts one-third of the total examinee pool in the AT/PT partitioning process. When the total examinee pool consists of only 100 participants, this leaves only 67 cases for calculating the DIMTEST statistic. Perhaps more importantly is the fact that the DIMTEST procedure tests the assumption of essential unidimensionality given in *Equation 7*. This assumption is concerned with item pair covariances conditional on ability. DIMTEST uses examinee total score on the PT items as a proxy for ability. As calculating a covariance requires a sample of at least $N = 2$, all total score subgroups with less than two cases are excluded from the analysis. It follows, then, that decreases in sample sizes are met with an increased probability of any particular ability subgroup being excluded. In the current work, for example, as

much as 18% of the 67 available cases excluded in the smallest sample size conditions due to this issue. This left as few as 54 examinees for calculating the DIMTEST statistic. In comparison, the largest sample size conditions saw only around 4% of the examinee pool being discarded due to insufficient ability subgroup counts, leaving ~480 valid cases.

Recommendations

The primary goal of this study was evaluate and compare the performance of DIMTEST and SGDDM under small-scale testing conditions. Previous research has indicated that DIMTEST may be able to maintain reasonable Type I error rates and be reasonably powered with as few as 250 examinees under certain conditions (e.g., simple structure, moderate discrimination and factor correlation parameter values), while no research has been conducted on the performance of SGDDM under small-scale conditions.

Overall, the results of the current work suggest that as little as 250 examinees with tests as short as 12 items (19 for DIMTEST) may be sufficient for assessing deviations from unidimensionality assuming the researcher is confident that any multidimensionality would exhibit factorially simple structure. Using samples as small as 100 examinees may be appropriate for use with SGDDM when combined with tests consisting of at least 18 items. If employing SGDDM as a diagnostic tool rather than a means to evaluate a unidimensionality hypothesis via traditional significance criteria, then as few as 12 items may be enough to yield a reliable assessment of data model fit (misfit).

Limitations and Opportunities for Further Research

The obvious caveat to the recommendations above is that they only hold under the assumption that multidimensionality manifest itself in a factorially simple structure. In practice, it may be rare for this to be the case and, perhaps more importantly, it may be impossible for a researcher to evaluate this assumption with any confidence before submitting their data to the DIMTEST or SGDDM processes. A more conservative recommendation as to sample size and test length minima may be one which makes no such test structure assumptions. A limitation to the current work in this respect is that no factorially complex structures were examined. Furthermore, previous research has suggested other item characteristics that may affect one's ability to reliably assess dimensionality. These characteristics include the magnitude of factor correlations, the strength of item discrimination parameters, and the skewness/kurtosis of the item difficulty distributions used during data generation. None of these factors were manipulated in the current work and, as such, the recommendations may not be generalizable to anything but a fairly narrow subset of testing conditions.

In light of these limitations, an obvious extension to the current work would be the inclusion of factorially complex test structures and a more expansive list of item characteristics as manipulated variables. Further extensions might include examining alternative approaches to assessing dimensionality as points of comparison. For example, nonlinear factor analytic approaches such as NOHARM have been compared to DIMTEST (Finch & Habing, 2007), but not to

SGDDM, though Levy and Svetina (2011) did compare GDDM to two statistics based on NOHARM modeling. Other approaches might include the evaluation of fit statistics obtained using a structural equation modeling (SEM) framework or an IRT-specific modeling program such as IRTPro (Cai, Thissen, & du Toit, 2011). Finally, a logical extension of the current work would be to expand the dimensionality assumption to include structures other than one of a unidimensional nature and/or to allow for items with more than two response categories (i.e., polytomous). Levy et al. (2012) have examined SGDDM's ability to assess the fit of data to more dimensionally complex models, such as that of a three dimensional model or a testlet model, but not under small-scale testing conditions. While DIMTEST is limited to the assumption of a unidimensional structure, the other approaches listed above (NOHARM, SEM, IRTPro) are not.

Summary

Similar to previous studies using DIMTEST and SGDDM, as well as other model-based covariance approaches to assessing dimensionality, performance with respect to assessing data-model fit improved with increases in either sample size or test length. From a hypothesis testing perspective, both approaches may be overly conservative in terms of Type I error rate with most of the conditions yielding proportions of extreme *p*/PPP values well below the nominal rate of .05. When viewed as a diagnostic tool, however, the SGDDM approach clearly demonstrated a high-degree of data-model fit in all 12 conditions where the correct model was fit. Similarly, a clear indication of data-model misfit was present using SGDDM when the unidimensional model was fit to the two-

dimensional data with all but one condition yielding a proportion of extreme PPP values above the .80 power threshold commonly used by social scientists under a frequentist null hypothesis-testing framework. Even in the most extreme case (N = 100 and J = 12), SGDDM still suggested misfit, though the proportion of extreme PPP values was below .80. Power was also satisfactory when using DIMTEST in conditions where the sample size was at least N = 250. Power tended to suffer, however, when using DIMTEST under the smallest sample size condition (N = 100).

In light of these results, it is recommended that sample sizes and test lengths of at least N = 250 and J = 19, respectively be used with DIMTEST and that values of at least N = 100 and J = 18 be used with SGDDM. This recommendation, however, assumes that the researcher be reasonably confident that any potential multidimensionality come only in factorially simple forms.

REFERENCES

Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement, 13*(2), 113-127.

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

Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using markov chain monte carlo. *Applied Psychological Measurement, 27*(6), 395-414.

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics, 7*(4), 434-455.

Cai, L., Thissen, D., & du Toit, S.H.C. (2011). IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]. Lincolnwood, IL: Scientific Software International.

Casella, G., & George, E. I. (1992). Explaining the gibbs sampler. *The American Statistician, 46*(3), 167-174.

De Ayala, R. J. (2009). *The theory and practice of item response theory.* Guilford Press.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists* Lawrence Erlbaum.

Fay, D. (2012). *Sample size and test length minima for DIMTEST with conditional covariance-based subtest selection.* (Unpublished M.A.). Arizona State University,

Finch, H., & Habing, B. (2007). Performance of DIMTEST-and NOHARM-based statistics for testing unidimensionality. *Applied Psychological Measurement, 31*(4), 292-307.

Fox, J. (2010). *Bayesian item response modeling: Theory and applications* Springer.

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

Froelich, A. G., & Habing, B. (2008). Conditional covariance-based subtest selection for DIMTEST. *Applied Psychological Measurement, 32*(2), 138-155.

Froelich, A. G., & Stout, W. (2003). A new bias correction method for the DIMTEST procedure. *Unpublished Manuscript.Retrieved may, 15*, 2005.

Gelfand, A. E., & Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association, 85*(410), 398-409.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis* Chapman & Hall/CRC.

Gelman, A., Meng, X., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6*, 733-759.

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.

Gill, J. (2007). *Bayesian methods: A social and behavioral sciences approach* (2nd ed.) CRC press.

Habing, B., & Roussos, L. A. (2003). On the need for negative local item dependence. *Psychometrika, 68*(3), 435-451.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* Sage Publications, Incorporated.

Kline, R. B. (2010). *Principles and practice of structural equation modeling* Guilford press.

Levy, R., Mislevy, R. J., & Sinharay, S. (2009). Posterior predictive model checking for multidimensionality in item response theory. *Applied Psychological Measurement, 33*(7), 519-537.

Levy, R., & Svetina, D. (2011). A generalized dimensionality discrepancy measure for dimensionality assessment in multidimensional item response theory. *British Journal of Mathematical and Statistical Psychology, 64*(2), 208-232.

Levy, R., Xu, Y., Yel, N., & Svetina, D. (2012). A standardized generalized dimensionality discrepancy measure and a standardized model-based covariance for dimensionality assessment for multidimensional models. *Unpublished Manuscript,*

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores.

Lunn, D.J., Thomas, A., Best, N., Spiegelhalter, D. (2000). WinBUGS – a Bayesian modeling framework: concepts, structure, and extensibility. *Statistics and Computing, 10,* 325-337.

McDonald, R. P. (1967). Nonlinear factor analysis. *Psychometric Monographs,*

McDonald, R. P. (1997). Normal-ogive multidimensional model. *Handbook of Modern Item Response Theory,* 257-269.

McDonald, R. P. (1999). *Test theory: A unified treatment* Lawrence Erlbaum.

McDonald, R. (1994). Testing for approximate dimensionality. *Modern Theories in Measurement: Problems and Issues,* , 63-86.

Muthén, L., & Muthén, B. (2007). Mplus. *User's Guide, Ed, 3*

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

Nandakumar, R., & Yu, F. (1996). Empirical validation of DIMTEST on nonnormal ability distributions. *Journal of Educational Measurement, 33*(3), 355-368.

Pyo, K. (2000). Assessing dimensionality of a set of language test data. Paper presented at the *Annual Meeting of the American Educational Research Association, New Orleans, LA. Appendices,*

Raftery, A. E. (1996). Hypothesis testing and model selection via posterior simulation. *Markov Chain Monte Carlo in Practice,* , 163-188.

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. *Handbook of Modern Item Response Theory,* , 271-286.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9*(4), 401-412.

Rijmen, F. (2010). Formal relations and an empirical comparison among the Bi-Factor, the testlet, and a Second-Order multidimensional IRT model. *Journal of Educational Measurement, 47*(3), 361-372.

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), 1-30.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applies statistician. *The Annals of Statistics,* , 1151-1172.

Seo, M., & Roussos, L. A. (2010). Formulation of a DIMTEST effect size measure (DESM) and evaluation of the DESM estimator bias. *Journal of Educational Measurement, 47*(4), 413-431.

Smith, A. F., & Roberts, G. O. (1993). Bayesian computation via the gibbs sampler and related markov chain monte carlo methods. *Journal of the Royal Statistical Society.Series B (Methodological),* , 3-23.

Socha, A., & DeMars, C. E. (2013). An investigation of sample size splitting on ATFIND and DIMTEST. *Educational and Psychological Measurement,*

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*(4), 589-617.

Stout, W., Froelich, A. G., & Gao, F. (2001). Using resampling methods to produce an improved DIMTEST procedure. *Essays on Item Response Theory,* , 357-375.

Stout, W., Habing, B., Douglas, J., Kim, H. R., Roussos, L., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement, 20*(4), 331-354.

Team, R. (2008). Development core. *R: A Language and Environment for Statistical Computing (Vienna: R Foundation for Statistical Computing),*

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*(2), 125-145.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3), 187-213.

Yung, Y., Thissen, D., & McLeod, L.D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika, 64*(2), pp. 113-128.

Zhang, J., & Stout, W. (1999a). Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika, 64*(2), 129-152.

Zhang, J., & Stout, W. (1999b). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64*(2), 213-249.

APPENDIX A

DATA GENERATION CODE

```
comp.MIRT.2PL.nogive.p <- function(theta.i, a.j, d.j){
        if(length(theta.i) != length(a.j)) print("Warning: Theta and Discrimination
        vectors of different size")
        p <- pnorm(sum(a.j*theta.i)+d.j, mean=0, sd=1, lower.tail=TRUE,
        log.p=FALSE)
        p
}

library(MASS)
library(msm)

for(which.rep in 1:n.reps){
        theta.true <- mvrnorm(n=N, mu=kappa.true, Sigma=phi.true)
        dim.angle <- runif(J, min=0, max=90)
        a.true.rand <- alpha.structure

        for(j in 1:J){
                if (M==1){
                        a.true.rand[j,1] = rtnorm(1, mean=1, sd=0.2, lower=0,
                        upper=Inf)
                }
                if (M==2){
                        if (alpha.structure[j,1]==1 & alpha.structure[j,2]==1){
                                a.true.rand[j,1] = rtnorm(1, mean=1, sd=0.2,
                                lower=0, upper=Inf)
                                a.true.rand[j,2] = (sin(dim.angle[j]) *
                                a.true.rand[j,1])
                                a.true.rand[j,1] = (cos(dim.angle[j]) *
                                a.true.rand[j,1])
                        }
                        if (alpha.structure[j,1]==1 & alpha.structure[j,2]==0){
                                a.true.rand[j,1] = rtnorm(1, mean=1, sd=0.2,
                                lower=0, upper=Inf)
                        }
                        if (alpha.structure[j,1]==0 & alpha.structure[j,2]==1){
                                a.true.rand[j,2] = rtnorm(1, mean=1, sd=0.2,
                                lower=0, upper=Inf)
                        }
                }
        }
        d.true.rand <- rnorm(J, mean=0, sd=0.7)
        X <- matrix(0, nrow=N, ncol=J)



        for(i in 1:N){
```

```
                for(j in 1:J){
                        p.i.j <- comp.MIRT.2PL.nogive.p(theta.i=theta.true[i, ],
                        a.j=a.true.rand[j, ], d.j=d.true.rand[j])
                        if(p.i.j >= runif(1)) X[i,j]=1
                }
        }

write.table(X, file= paste(condition.folder, "uni_n100_j12/Data/", "data.", "rep.",
which.rep, ".dat", sep=""), sep="", col.names=FALSE, row.names=FALSE)

write.table(theta.true, file=paste(condition.folder, "uni_n100_j12/Parameters/",
"theta.", "rep.", which.rep, ".dat", sep=""), sep=" ", col.names=FALSE,
row.names=FALSE)

write.table(d.true.rand, file=paste(condition.folder, "uni_n100_j12/Parameters/",
"d.", "rep.", which.rep, ".dat", sep=""), sep=" ", col.names=FALSE,
row.names=FALSE)

write.table(a.true.rand, file=paste(condition.folder, "uni_n100_j12/Parameters/",
"a.", "rep.", which.rep, ".dat", sep=""), sep=" ", col.names=FALSE,
row.names=FALSE)
}

#END
```

APPENDIX B

MODEL ESTIMATION SYNTAX

```
TITLE:
Fit a unidimensional model in Mplus

DATA:
FILE IS data.rep.1.dat;

VARIABLE:
NAMES ARE  x1 - x12;
USEVARIABLES  x1 - x12;
CATEGORICAL ARE  x1 - x12;

ANALYSIS:
ESTIMATOR = BAYES;
CHAINS = 1;
FBITERATIONS =  5000;
POINT = MEAN;

MODEL:
f1 by
x1*(f1x1)
x2*(f1x2)
x3*(f1x3)
x4*(f1x4)
x5*(f1x5)
x6*(f1x6)
x7*(f1x7)
x8*(f1x8)
x9*(f1x9)
x10*(f1x10)
x11*(f1x11)
x12*(f1x12);
[f1@0];
f1@1;
[x1$1*](d1x1);
[x2$1*](d1x2);
[x3$1*](d1x3);
[x4$1*](d1x4);
[x5$1*](d1x5);
[x6$1*](d1x6);
[x7$1*](d1x7);
[x8$1*](d1x8);
[x9$1*](d1x9);
[x10$1*](d1x10);
[x11$1*](d1x11);
[x12$1*](d1x12);
MODEL PRIORS:
```

```
f1x1~N(1,10);
f1x2~N(1,10);
f1x3~N(1,10);
f1x4~N(1,10);
f1x5~N(1,10);
f1x6~N(1,10);
f1x7~N(1,10);
f1x8~N(1,10);
f1x9~N(1,10);
f1x10~N(1,10);
f1x11~N(1,10);
f1x12~N(1,10);
d1x1~N(0,10);
d1x2~N(0,10);
d1x3~N(0,10);
d1x4~N(0,10);
d1x5~N(0,10);
d1x6~N(0,10);
d1x7~N(0,10);
d1x8~N(0,10);
d1x9~N(0,10);
d1x10~N(0,10);
d1x11~N(0,10);
d1x12~N(0,10);

MODEL CONSTRAINT:
f1x1>0;
f1x2>0;
f1x3>0;
f1x4>0;
f1x5>0;
f1x6>0;
f1x7>0;
f1x8>0;
f1x9>0;
f1x10>0;
f1x11>0;
f1x12>0;

DATA IMPUTATION:
      IMPUTE = ALL (c);
      PLAUSIBLE = latent.rep.1.chain.1.out;
      SAVE = fit.rep.1.chain.1.impute.*.out;
      NDATASETS = 100;
OUTPUT:
TECH1 TECH8;
```

```
PLOT:
TYPE = PLOT2;
```

APPENDIX C

BIFACTOR PARAMETERS

Table 3

Translation of unstandardized parameter values from a two-dimensional simple structure to a bifactor model for a 12-item exam.

	2DSS			Bifactor			
Parameter	θ_1	θ_2	δ	θ_g	θ_1	θ_2	δ
σ^2_θ	1	1	--	1	1	1	--
α_1	1	0	0	.55	.84	0	0
α_2	1	0	0	.55	.84	0	0
α_3	1	0	0	.55	.84	0	0
α_4	1	0	0	.55	.84	0	0
α_5	1	0	0	.55	.84	0	0
α_6	1	0	0	.55	.84	0	0
α_7	1	0	0	.55	.84	0	0
α_8	1	0	0	.55	.84	0	0
α_9	0	1	0	.55	0	.84	0
α_{10}	0	1	0	.55	0	.84	0
α_{11}	0	1	0	.55	0	.84	0
α_{12}	0	1	0	.55	0	.84	0
$\rho_{\theta_1, \theta_2}$	0.3			0			

Note. σ^2_θ denotes a factor variance, α an item discrimination, δ an item difficulty, and $\rho_{\theta_1, \theta_2}$ the correlation between two factors. The loadings on θ_g are constrained to be equal for model identification purposes.