INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI®

Université d'Ottawa · University of Ottawa

ASSESSING THE PERFORMANCE OF THE APPROXIMATE CHI-SQUARE AND STOUT'S T STATISTICS WITH DIFFERENT TEST STRUCTURES

by

Xiao L. Pang

Faculty of Education

Dissertation presented to the School of Graduate Studies and Research as partial fulfillment of the Ph.D. Degree in Education

University of Ottawa

ACKNOWLEGEMENTS

This study could not have been completed without the timely intervention of several individuals. I would like to express my heartfelt thanks to Dr. Marvin Boss, my supervisor, for his encouragement, support, criticism, patience with me throughout the years, and guiding me to the right direction. His sense of responsibility as an educator and carefulness are deeply appreciated.

I would also like to express my heartfelt thanks to Dr. Marc Gessaroli for being my lecturer first and then my previous supervisor, and all the technical support, valuable comments, and patience in helping me with my questions throughout the years. His understanding, generosity, and high level of psychometric expertise are most impressive and unforgettable.

Heartfelt thanks also go to Dr. Weimin Li for his support, valuable inputs, and spiritual encouragement.

Finally, I owe a great deal to my husband, Hongqi, and my daughter, YaYa, for their support, understanding, spiritual encouragement, and sacrifice made for me throughout the years. Now, it's time to make up!

Table of Contents

List of Tables

List of Figures

Abstract

The assessment of dimensionality underlying the responses to a set of test items is a fundamental issue in applying correct IRT models to examinee ability estimation and test result interpretation. Thus, a search of effective ways of assessing dimensionality and valid criteria for dimension decision making has been of interest.

Currently, three assessment methods have been shown to be particularly promising: the original Stout's T (Stout's T_1), the refined Stout's T (Stout's T_2) (Nandakumar, 1987, 1993) based on Stout's essential unidimensional assumption (Stout, 1987), and the Approximate χ^2 (De Champlain & Gessaroli, 1992) derived from McDonald's nonlinear factor analysis based on the weak principle of local independence (McDonald, 1981).

However, the three indices have only been tested under limited research conditions. For example, the Approximate χ^2 has only been compared to Stout's T_2 with unidimensional and two dimensional simple test structures (De Champlain, 1992; Gessorali & De Champlain, 1996; Gessorali, De Champlain, & Folske, 1997). A similar comparison was made among the Approximate χ^2, Stout's T_1, and Stout's T_2 by Breithaupt (1995) . Stout's T_1 was assessed with complex test structure where the importance of the two dimensions was set to be equal (Stout, 1987). Similar studies were carried out with Stout's T_2 (Nandakumar, 1994; Nandakumar & Stout, 1993) and the Approximate χ^2 (De Champlain & Gessaroli, 1998). Clearly, there is more to learn about the utility of the three indices.

The purpose of this study was to assess and compare the Type I error rates and power of the Approximate χ^2, Stout's T_1, and Stout's T_2 in assessing dimensionality of a set of item responses. The variables used in the Type I error study were test length (L) (40 and 80 items),

sample size (N) (500, 1,000, and 2,000), and item discrimination (a) (.7, 1.0, and .14). A 2x3x3 design was created. For each cell of the design, 100 replications were carried out. In the power study, in addition to the three variables used in the Type I error study, different test structure (two dimensional simple test structure and two dimensional complex test structure) and dimension correlation (r=.0, .4, .57, and .7) were used. For both simple and complex test structure, a dimension ratio of 3:1 was set. Similarly, 100 replications were carried out for each combination of the conditions. A total of 14,400 data sets were generated.

With regards to the Type I error rate, the Approximate χ^2 had perfect Type I error control with zero rejections of the unidimensional assumption across all conditions simulated in this study.

Higher Type I error rates were observed for Stout's T_1 and Stout's T_2 than the Approximate χ^2. However, in most conditions, the Type I error rates were lower than the nominal level of .05. For both indices, long test length and small sample size resulted in higher Type I error rates. The inflated Type I errors (greater than the nominal level of .05) were observed for a few cases. Slightly higher Type I error rates were found for Stout's T_2 than for Stout's T_1 under most conditions.

Based on the results of the power study, with two dimensional simple test structure, the Approximate χ^2 had excellent power in most cases except for N=500, a=.7 and 1.0, and r=.7. The mean power of the Approximate χ^2 was .966. Test length had no effect on the power of the Approximate χ^2.

Excellent power with two dimensional simple test structure was found for Stout's T_1 and Stout's T_2. Except for the cell of a=.7, N=500, r=.7, and L=40, power for Stout's T_1 and Stout's T_2

was greater than .89. The mean power was .99 for both indices.

With complex test structure, the performance of the Approximate χ^2 was found to be quite different. In most cases, the Approximate χ^2 either performed poorly or completely failed to identify the right test structure. The average power was only .23. Except for r=.0, increasing sample size and discrimination resulted in no or little improvement in the power of the Approximate χ^2. The only cell where the improvement in power for the Approximate χ^2 was shown was the cell of N=2,000 and a=1.4. Under this condition, excellent or good power across all levels of correlation was observed. The average power for this cell was .98. Overall, for 15 of 72 occasions good or excellent power was obtained.

The power of Stout's T_1 and Stout's T_2 was less affected by complex test structure. However, in many cases the power of the two indices was not satisfactory. This situation became more pronounced as test length decreased and correlation increased. The mean power for Stout's T_1 was .43 and for Stout's T_2 was .45. For r≥.4 at L=40 and r≥.57 at L=80, increasing sample size and discrimination did not produce satisfactory power in a majority of cases. Under these conditions, the power of both indices was very poor. In contrast to the Approximate χ^2, under the condition of a=1.4, r≥.57, and N≥1.000, the power of Stout's T_1 and Stout's T_2 was found either no improvement or was negatively related to sample size. As sample size increased, more interaction effects were present for both indices. In most cases, a similar interaction pattern was observed for both indices. Slightly higher power was found for Stout's T_2 in most cases. Overall, for 21 of 72 occasions Stout's T_1 and Stout's T_2 had good or excellent power.

In conclusion, according to the results obtained in this study, each index possesses certain advantages and drawbacks. The Approximate χ^2 had good Type I error of zero over all

conditions and excellent power with two dimensional simple test structure. Stout's T_1 and Stout's T_2, on the other hand, had higher Type I error rates than the Approximate χ^2, ranging from zero to 12, excellent power with two dimensional simple test structure, and better power than the Approximate χ^2 with two dimensional complex test structure. However, Stout's T_1 and Stout's T_2 have to be used with great caution given the unsatisfactory power shown with two dimensional complex test structure in many cases.

It should be pointed out that these results should be interpreted with extreme caution. The parameters identified for items and for tests limit generalizations that may be made. However, the results do lead one to suggest that continued study with both real and simulated data be carried out to determine the utility of these indices.

Chapter I

INTRODUCTION

It is well known that a crucial assumption underlying most item response models (IRM) is that item response probabilities are a function of a single trait. In real testing situations, however, this unidimensional assumption may never be strictly met. Researchers have shown that real test data often may not be well modeled by locally independent unidimensional models (Ackerman, 1987; Blais & Laurier, 1995; Gessaroli, 1995; Reckase, 1979; Thissen, Steinberg, & Mooney, 1989; Yen, 1984, 1985, 1993). For example, an achievement math test that includes verbal questions may require reading comprehension in addition to knowledge of mathematics. A reading test on distinct subjects may include independent clusters after the general reading ability has been partialed out, indicating the existence of other dimensions (Gessaroli, 1995; Thissen, Steinberg, & Mooney, 1989). In these cases, the interpretation of the test results based on a unidimensional model may not be appropriate.

Effects produced using a unidimensional model with multidimensional data have been reported. In general, item parameters were poorly recovered, especially when several equally important abilities were required to correctly answer an item (Ackerman, 1987; Ansley & Forsyth, 1985; Drasgow & Parsons, 1983; Reckase, 1979). For example, Reckase (1979) assessed the influence of multidimensionality on the performance of unidimensionally based programs such as Logist. He found that when two or more dominant dimensions were present in the data, LOGIST estimated ability poorly by tending to track only one component of ability. Drasgow and Parsons (1983), using Logist, have demonstrated that the estimation of both ability and item structure was distorted consistently as the dominance of the general latent factor

decreased. Yen (1984, 1985) found that a systematic underestimation of ability often occurs with items involving more than one skill when analyzed based on a unidimensional logistic model. Furthermore, analyzing multidimensional data based on a unidimensional model can invalidate the applications of IRT-based techniques in test equating, item calibrating, and differential item functioning investigations (Camilli, 1992; Oshima & Miller 1992).

Other concerns about the misinterpretation of test results when the assumption of unidimensionality is violated are the misclassification of mastery or non-mastery status, incorrect assignment of examinees selected for special programs, and misdiagnosis of learning disabilities (Breithaupt, 1995; De Champlain, 1992; Gessaroli, 1995; Hattie, 1984). Thus, it is a necessary task to determine the correct dimensional test structure in the application of IRT-based techniques.

This problem has led to a search for effective ways of assessing dimensionality and valid criteria for decision making. Numerous indices have been developed. Unfortunately, for most of the indices, use has been shown to be problematic. Still others, though promising, are purely descriptive and offer no concrete criterion for the determination of the number of dimensions underlying a set of item responses (Hambleton & Rovinelli, 1986; Hattie, 1984,1985; Nandakumar, 1994). Currently, three assessment methods have been shown to be particularly promising: the original Stout's T (Stout's T_1) (Stout, 1987) and the refined Stout's T (Stout's T_2) (Nandakumar, 1987, 1993) based on Stout's essential unidimensionality (Stout, 1987), and the Approximate χ^2 (De Champlain & Gessaroli, 1992) derived from McDonald's nonlinear factor analysis (NFA) based on the weak principle of local independence (McDonald, 1981).

All three indices are quite effective in identifying unidimensional test structure and the

nature of certain multidimensional test structures in various conditions (De Champlain, 1992; Breithaupt, 1995; Gessaroli, 1995; Gessaroli & De Champlain, 1996; De Champlain & Gessaroli, 1998; Gessaroli, De Champlain, & Folske, 1997; Nandakumar, 1991, 1994; Nandakumar & Stout, 1993). Because multidimensional tests may be structured very differently, the potential of these three statistics in assessing test dimensionality under a wide range of test structures has not been fully explored and compared. In this study, the three indices are assessed under various experimental conditions. Particularly, two kinds of test structure frequently found in multidimensional tests, simple test structure and complex test structure, are to be studied. Simple test structures are those wherein each item measures only one dimension but all items do not measure the same dimension. Complex test structures are those wherein the probability of correctly answering some items is determined by more than one ability. Most of the earlier work assessing dimensionality using the three indices has been with simple test structure. Thus, an interesting question relates to how sensitive the three indices are to complex test structure.

The purpose of this study is to examine the power and Type I error of the Approximate χ^2, Stout's T_1, and Stout's T_2 in rejecting the assumption of unidimensionality under different conditions. Variables manipulated in this study are, in the unidimensional case, test length, sample size, and item discrimination and in the two dimensional case, type of test structure (two-dimensional simple and two dimensional complex test structures), test length, sample size, item discrimination, and correlation between traits.

To support the study, a detailed review of relevant literature is presented in Chapter II. In Chapter III, the methodology is presented. Results are reported in Chapter IV, discussion is given in Chapter V, and conclusions and suggestions for future research are presented in Chapter VI.

Chapter II

REVIEW OF LITERATURE

In this chapter, a review of the methods used to assess test dimensionality is presented. First, early indices used in test dimension assessment are briefly discussed. Indices include those based on answer patterns, reliability, component analysis, and linear factor analysis. Second, two broad categories of recently developed indices are discussed in detail. The first category covers indices based on nonlinear factor analytical models. The second category includes indices based on nonparametric analytical models. Third, a detailed review of studies related to four promising indices (the Approximate χ^2, an Approximate Likelihood Ratio χ^2, Stout's T_1, Stout's T_2) is given. Finally, the results provided by the studies are summarized and research questions for the present study are presented.

Early Methods and Indices Used in Test Dimension Assessment

Early methods and indices that were used to assess test dimensionality have been extensively studied, evaluated, and reviewed (De Champlain, 1992; Gessaroli, 1994; Gessaroli & De Champlain, 1996; Hambleton & Rovinelli, 1986; Hattie, 1984, 1985; McDonald, 1995; Nandakumar, 1991, 1994; Nandakumar & Stout, 1993; Stout, 1987, 1990). These indices can be grouped into three categories: indices based on answer patterns, indices based on reliability, and indices based on component and linear factor analysis.

Indices based on answer patterns have been shown to be poor indicators of test dimensionality due to the inappropriate assumptions underlying these indices and ineffective applications. In practical applications, indices based on answer patterns can not distinguish a unidimensional test from a multidimensional test (Hattie, 1984; 1985).

Indices based on reliability have also been dismissed by researchers (Hattie, 1984; 1985; McDonald, 1995) for their lack of rationale. For example, with coefficient alpha (Cronbach, 1951) it is assumed that a necessary condition for a unidimensional test is that the matrix of inter-item correlations is of unit rank. Yet, researchers have found that it is possible for the matrix of a set of items that measure more than one factor to be of unit rank (Novick & Lewis, 1967). In addition, coefficient alpha increases as the number of items increases (Green, 1977). Thus, it can not be used as an indicator of test dimensionality.

Indices based on principal components and linear factor analyses of phi correlation matrices or tetrachoric correlation matrices have been widely studied. These indices involve using a criterion such as the magnitude of the first eigenvalue, the ratio of the first to the second eigen value (Bentler, 1972; Hambleton, 1980; Hutten, 1980; Lord, 1980), the percentage of variance accounted for by the first factor (Cattell & Tsujioka, 1964; Hambleton & Traub, 1973; Hutten, 1979; Reckase, 1979), the number of eigenvalues greater than one (Armor, 1974; Birnbaum & Tatsuoka, 1982; Laforge, 1965), the sum of residuals (McDonald, 1981), the correlation of raw and factor scores (Dubois, 1970), chi-square based on linear factor analysis (Bock & Lieberman, 1970), goodness of fit based on the ratio of the amount of variance associated with one factor to total test variance (Tucher & Lewis, 1973), Omega (Armor, 1974; Carmines & Zeller, 1979; Greene & Carmines, 1980; Heise & Bohrnstedt, 1970; Smith, 1974), and communalities (Green, Lissitz, & Mulaik, 1977; Hattie & Hansford, 1982; Watkins & Hattie, 1980). Researchers have demonstrated that the use of indices based on principal components and linear factor analysis are generally problematic in assessing test dimensionality (Berger & Knol, 1990; Hambleton & Rovinelli, 1986; Hattie, 1984; Zwick & Velicer, 1986). First, in most cases,

they have been found to overestimate the actual number of dimensions underlying a set of item responses. For example, a linear factor analysis of phi correlation matrices often resulted in the identification of spurious factors that were originally termed as difficulty factors . Second, using a linear factor analysis of a tetrachoric correlation matrix often results in non-Gramian correlation matrices (negative values in the matrix) and Heywood cases (communalities equal to or greater than one) (Hattie, 1984; Nandakumar, 1991). Third, the analysis of tetrachoric correlations between two dichotomous variables requires a stringent assumption of normality for the latent trait variable underlying each of the dichotomous items. This is not a requirement for dimensionality defined in terms of latent traits. Violation of this assumption can result in a distortion of the true degree of association between the two dichotomous variables. Fourth, the presence of guessing can affect the analysis of tetrachoric correlations, producing spurious factors (Carroll, 1945). Lord (1980) has suggested that the tetrachoric correlation matrix should not be used when guessing is present. Finally, a large sample size is required to obtain fairly accurate estimates of tetrachoric correlations (Gessaroli, 1995; Hulin, Drasgow, & Parsons, 1983). For the above reasons, there has been a decline in using principal component and linear factor analysis methods to assess test dimensionality.

Indices Based on Nonlinear Factor Analysis

Nonlinear factor analytical models have theoretically and empirically been proved to be among the few most promising methods for assessing dimensionality underlying dichotomous item responses. The basis for the promise of this approach is the close relationship between nonlinear factor analytical models and item response models. Specifically, item response models have been mathematically proven to be nonlinear counterparts of the Spearman common factor

models (McDonald, 1967, 1982, 1989; Takane and De Leeuw, 1987). McDonald (1967) has shown that if the latent traits have normal distributions, the two-parameter normal ogive model can be fitted to binary data by the analysis of covariance structures. In the following section, the assumptions underlying nonlinear factor analytical models are discussed.

The Assumptions Underlying Nonlinear Factor Analytical Models

Nonlinear factor analytical models are defined by either a strong (full information) form or a weak (bivariate information) form of the principle of local independence. The strong principle of local independence states that for specific latent traits $\underline{\theta}$ (or factors), the probability of obtaining correct responses to any item is independent of the probability of correctly responding to any other item. With the weak principle of local independence (McDonald, 1981), it is assumed that at any fixed points of the latent traits the residual item covariances for any two items after a factor analysis has been carried out should equal zero. That is, for any items U_i and U_j,

$$Cov(U_i, U_j \mid \Theta) = 0, i \neq j. \qquad (1)$$

In application, based on the two different principles of local independence, the test of nonlinear factor model fit is approached by two essentially different methods: full information methods based on the strong principle of local independence assumption and the limited information method based on McDonald's weak principle of local independence (McDonald, 1967). Theoretically sound, the full-information analysis method (Bock & Aitkin, 1981; Bock &

Lieberman, 1970; Mckinley, 1989) has been shown to be computationally inefficient. Due to the stringent assumption of strong local independence, this method uses all the information in item responses. This results in 2^p observed, distinct response patterns that are large relative to the sample size. In this case, to obtain fairly accurate parameter estimation even for a moderate test length, an extremely large sample size is required (Berger & Knol, 1990; Gessaroli, 1995; Knol & Berger, 1991). In addition, statistics such as likelihood-ratio χ^2 and likelihood-ratio χ^2 difference test based on the full information methods have been shown not very powerful in detecting the correct number of abilities underlying a set of test items (Berger & Knol, 1990). De Champlain and Gessaroli (1998) found that likelihood-ratio χ^2 difference test suffered from an inflated Type I error rate. Its power was highly influenced by the independent variables manipulated such as test length, sample size, and dimension correlation.

On the other hand, the limited information method (Christoffersson, 1975; Fraser & McDonald, R. P., 1988; McDonald, 1985; Muthen, 1978) based on the weak principle of local independence uses pairwise relationships between the items. Researchers have shown that using both full information and limited information methods assessing model fit to the same data set has yielded a similar amount of information (Bock & Lieberman, 1970; Christoffersson, 1975; Knol & Berger, 1990; McDonald, 1995;). McDonald (1995) concludes that the information contained in the higher joint moments of the binary responses does not significantly increase model misfit. Given the above, it seems that the weak principle of local independence and limited information methods are preferred from the practical point of view. Several indices and procedures based on the limited information methods have been developed. An overview of these indices follows.

Christoffersson and Muthen's GLS Procedure

Christoffersson (1975) fitted two parameter logistical models to a set of binary data using the Generalized Least Squares method (GLS) to minimize the weighted sum of squares of a covariance matrix of residuals, $\varepsilon'\Sigma_\varepsilon^{-1}\varepsilon$. $\varepsilon'\Sigma_\varepsilon^{-1}\varepsilon$ is estimated by $\varepsilon'S_\varepsilon^{-1}\varepsilon$ when Σ_ε is unknown. The estimator is defined by minimizing

$$F(h, \Lambda, \Phi) = \varepsilon' S^{-1} \varepsilon = (P - P^*)' S^{-1} (P - P^*); h' = (h_1, h_2, \ldots, h_m) \qquad (2)$$

which is regarded as a function of the unknown parameters h, Λ, and Φ, where h_i is the threshold level, Λ is the loadings, and Φ is the factor correlation. The minimized $F(h, \Lambda, \Phi)$ is asymptotically distributed as chi-square with 1/2M(M+1)-df, the effective number of estimated parameters. Thus a theoretical chi-square statistic can be calculated based on GLS to test the number of significant factors that best fit the data set.

The GLS approach has been found to be promising in assessing number of dimensions underlying binary data. However, due to the nature of using a weight matrix in GLS, the size of the covariance matrix S and the time to compute S increases rapidly with the increase of number of items. As a result the maximum number of items that can be analyzed is limited. This has greatly restricted the use of Christoffersson and Muthen's method.

Akaike's Information Criterion (AIC) Statistic

AIC (Akaike, 1987), based on maximum likelihood estimation, has been used as a criterion for dimension assessment (Berger & Knol, 1990; Gessaroli, 1995; McDonald, 1989). AIC is of the form

$$AIC(H) = \chi 2_{df} - 2(d.f.) \qquad (3)$$

Theoretically, AIC may be used as a measure of expected squared distance, in a suitable metric, between the maximum likelihood estimates of the parameters in a k-dimensional model and parameters describing the true unrestricted distribution from which the sample size is drawn (McDonald, 1995). AIC is interpreted as a "badness-of-fit" index. In the context of dimensionality, the optimum model would have the smallest AIC value; that is, the higher the value, the less adequate the fit of a given model. Gessaroli (1995) used AIC to assess successive factor models that were fitted to a four paragraph based reading comprehension test. AIC was found to decrease constantly as the number of factors obtained increased. However, the result has to be interpreted cautiously for four reasons: 1) Gessaroli (1995) used real data sets with no a priori knowledge of the exact number of underlying dimensions; 2) AIC is heavily dependent upon sample size (McDonald, 1989); 3) it has been found that, for a sufficiently small sample size, the optimum value of AIC can only be obtained by the unidimensional model, and for a sufficiently large sample size, it can only be obtained by the saturated model. McDonald (1995) concluded that it would appear that AIC cannot possibly be recommended for use with real data; 4) AIC is of a descriptive nature. It does not provide a significance test of the discrepancy between each successive factor model.

The Unbiased Relative Fit Index (h_k)

Another index that has been suggested for use in dimension assessment is h_k (Bentler, 1990; McDonald & Marsh, 1990). h_k is written as

$$h_k = 1 - (d_k / d_0) \qquad (4)$$

That is, to a null model,

where

d_k is the estimated noncentrality parameter for the k-dimensional model;

d_0 is the same quantity computed for the model of zero dimensions.

h_k measures the goodness of fit of a proposed model relative to a null model with the purpose of identifying the most parsimonious model that fits the data.

In the case of item response models, the null model is either the hypothesis of mutual statistical independence in full information methods or the hypothesis of zero item covariances in bivariate information methods. Using the limited information method , d_0 could be obtained in principle with the factor loading matrix null. According to McDonald (1995), a value of $h_k \geq .9$ indicates an acceptable fit. Gessaroli (1995) used h_k to assess the fit of the nonlinear factor analytical models to a set of responses of 2000 examinees to a 40 item reading comprehension test. For the 1, 2, 3, and 4 factor models specified in the study, h_k values were all greater than .9, though each higher level of factor model showed slight improvement over the lower level factor models. If a value of .9 as a criterion is acceptable for model selection and according to the principle of parsimony for model selection, the 1-factor model instead of other models should be chosen. However, the reading test did not seem to be unidimensional since it included four paragraphs each based on different subject matter. As with other descriptive statistics, model selection using h_k seems to be mainly a subjective process.

The Mean Absolute Residual and Standardized Mean Absolute Residual

Researchers have suggested that the mean residual covariances or the mean item residual correlations (r_{ij}) obtained after a nonlinear factor model is fitted to a set of item responses be employed as a criterion for dimension assessment (Hattie, 1984; McDonald, 1995). Empirical studies have demonstrated that these indices are related to the number of dimensions underlying a set of item responses (Gessaroli, 1995; Hambleton & Rovinelli 1986; Hattie, 1984; Nandakumar 1994).

Similarly, these indices are of a descriptive nature. They do not offer a criterion for determining at what point these goodness-of-fit statistics are sufficiently small for a test to be unidimensional. Hambleton and Rovinelli (1986) have also emphasized that a problem in the use of nonlinear factor analysis is the appropriate number of factors and polynomial terms to retain in a solution. Motivated by this need, Gessaroli and De Champlain (1992, 1996) proposed the Approximate χ^2 and Gessaroli, De Champlain, and Folske (1997) suggested an Approximate Likelihood Ratio χ^2 be used as significance tests of the results obtained after fitting a nonlinear factor model to a set of item responses.

The Approximate χ^2 Based on McDonald's Nonlinear Factor Analytical Model

The Approximate χ^2, based on McDonald's weak principle of local independence, assesses the null hypothesis that the sum of the off-diagonal elements in a matrix of residual correlations is equal to zero after fitting an m-factor nonlinear model to a set of item responses. The advantage of the use of the Approximate χ^2 with a nonlinear factor analytical model is that it offers a statistical significance test of model fit. The Approximate χ^2 is computed based on the results obtained after fitting a m-factor model to a set of item responses.

The calculation of the Approximate χ^2 statistic implemented in NOHARM II (Fraser, 1988) involves the following five computational steps after fitting a one or two factor model.

1. For all pairs of items, determine the proportion of examinees who correctly answered item i, item j, as well as both items. These quantities are referred to as $p_i^{(o)}$, $p_j^{(o)}$, and $p_{ij}^{(o)}$, respectively.

2. Based on the results of the q parameter model, for all pairs of items determine the expected as well as residual joint-proportions of examinees who correctly answered items i and j. The estimates of the residual joint-proportions are provided by the computer program NOHARM II (Fraser, 1988) and are referred to as $p_{ij}^{(r)}$.

3. Calculate the estimated residual correlation ($r_{ij}^{(r)}$) for each pair of dichotomous items with the following formula:

$$r_{ij}^{(r)} = \frac{p_{ij}^{(r)}}{\sqrt{p_i^{(o)}(1-p_i^{(o)})\,p_j^{(o)}(1-p_j^{(o)})}} \qquad (5)$$

4. Transform each of the estimated residual correlations to a Fisher z ($z_{ij}^{(r)}$) using

$$z_{ij}^{(r)} = .5\log_e(1+r_{ij}^{(r)}) - .5\log_e(1-r_{ij}^{(r)}) \qquad (6)$$

5. Calculate an approximate χ^2 statistic defined as

$$\chi^2 = (N-3)\sum_{i=2}^{n}\sum_{j=1}^{i-1} z_{ij}^{2(r)} \qquad (7)$$

where

$z_{ij}^{2(r)}$ is the square of the Fisher z corresponding to the residual correlation between items i and j, (i, j=1,....,n) and N is the number of subjects in the sample.

This statistic is approximately distributed as a central χ^2 with df = .5n(n-1)-t, where n is the number of items and t is the total number of independent parameters estimated. The χ^2 value is compared with the critical value corresponding to the degrees of freedom. A significant value means that there exists a large discrepancy between the expected item residual covariance matrix and the observed item residual covariance matrix. The null hypothesis of the model fit is rejected and a more appropriate model should be tested.

According to De Champlain (1992), De Champlain and Gessaroli (1998), Gessaroli (1995), Gessaroli and De Champlain (1996), the Approximate χ^2 has several advantages: 1) The assessment of dimensionality is based on a model that is directly related to IRT (nonlinear factor analysis); 2) the statistic has the desirable trait of being based on the discrepancy function, i.e. the discrepancy between the observed and fitted item-covariance matrix, which is consistent with the unweighted least-squares estimation procedure; 3) because of the unweighted-least squares estimation procedure, this statistic is derived from a model that has no severe limitation on the number of items or dimensions that can be analyzed; (4) this statistic involves actual hypothesis testing and is not merely descriptive; (5) it can be used to assess departure from unidimensionality as well as the successive fit of more complex models to a set of item responses; and (6) it is less inflated given the number of parameters estimated with the model as

compared with other indices such as AIC, the Likelihood-ratio χ^2, the χ^2 Likelihood-ratio difference test based on full information method, and LISRE8 χ^2 (Jörekog & Sörbom, 1993) based on structural equation modeling procedure.

However, there are also some limitations with the Approximate χ^2. First, it lacks a strong theoretical foundation, since it is computed from the results of unweighted least-squares estimation. This limitation, however, does not seem to seriously affect its application. In fact, χ^2 based on unweighted least-squares estimation has been found to be equivalent to a χ^2 obtained from GLS estimation (Browne, 1977). Second, it is well known that a problem with any chi-square test of model fit is that the best fitting model as shown by the chi-square value is a function of sample size. A similar limitation may also exist with the Approximate χ^2.

In short, with obvious advantages over other indices, the Approximate χ^2 appears to be one of the most important indices and its application under wide and complex testing conditions is worth further investigation.

The Approximate Likelihood Ratio χ^2

Another index used to test model fit after fitting an m-factor nonlinear model to a set of item responses is an Approximate Likelihood Ratio χ^2 (Gessaroli, De Champlain, & Folske, 1997). The Approximate Likelihood Ratio χ^2 can be calculated for each pair of items j and k. It is calculated as

$$G^2_{jk} = 2\sum_{l\ 0}^{i}\sum_{m\ 0}^{l} p_{lm} \ln \frac{\hat{p}_{lm}}{\hat{p}_{lm}}, \qquad (8)$$

and is asymptotically distributed as χ^2 with one degree of freedom. Under the null hypothesis each G^2_{jk} should be independent; thus a test of the null hypothesis is given by

$$\chi^2 = \sum_{j\ k} G^2_{jk} \qquad (9)$$

It could be tested against a chi-square distribution with n(n-1)/2-t degrees of freedom where n is the number of items and t is the number of estimated parameters (Gessaroli, De Champlain, & Folske, 1997).

According to Gessaroli, De Champlain, and Folske (1997), the difference between the Approximate χ^2 and the Approximate Likelihood Ratio χ^2 is that the former is obtained after first transforming the residual joint proportions to Fisher z values and finally correlation coefficients while the latter is based directly on the residual joint proportions obtained after fitting a m-factor nonlinear factor model to a set of item responses. The fact that the Approximate Likelihood Ratio χ^2 is calculated directly from the results obtained from the residual joint proportions obtained from a NOHARM solution is an advantage over the Approximate χ^2, because while transforming the residual joint proportions to Fisher z values the Approximate χ^2 may lose some information .

Indices Based on Nonparametric Analytical Methods

A number of indices and procedures based on nonparametric analytical methods have been developed and used in the assessment of test dimensionality. The most popular and

representative ones that involve statistical significance tests are Rosenbaum and Holland's procedure, Stout's T_1, and Stout's T_2. In this section, the three methods are discussed in detail. In addition, a few recently developed descriptive statistics are also briefly reviewed.

Rosenbaum and Holland's Procedure

Rosenbaum (1984) and Holland and Rosenbaum (1986) proposed a method (HR) based on conditional association between pairs of items for assessing dimensionality. The basic concept underlying HR is that if the items are locally independent and unidimensional and the item characteristic curves are monotonic, then the items are conditionally positively associated. The conditional association for each pair of items is tested using the Mantel-Haenszel (MH) test (Mantel & Haenszel, 1959) after the number-right score on the remaining items is fixed. The computed Z value is compared to the lower tail of the standard normal distribution. A statistically significant Z implies that the pair of items in question are not conditionally positively associated, after the score on the remaining items is fixed, and thus are not unidimensional. The M-H statistic is computed for all N(N-1)/2 pairs of items, where N is the total number of items in a test. A large number of item pairs with significant negative correlations conditional on N-2 total scores indicates that the test is multidimensional.

Ben-Simon and Cohen (1990), Nandakumar (1994), and Zwick (1987) found that the HR approach was a very conservative test, i.e. the number of significant negative partial associations for unidimensional data sets was far below the expected 5% level. Ben-Simon and Cohen (1990) found that HR erroneously misclassified nearly half of the multidimensional items they analyzed as unidimensional. Nandakumar (1994) found that HR showed good power in rejecting unidimensionality when the correlation between traits was low (r=.3). It failed to reject

unidimensionality when correlation between traits was high (r=.7). The reason for the HR losing power, according to Zwick (1987), is that the conditioning score used in computing the covariances is not perfectly correlated with the latent trait variable; thus, an appropriate choice of the conditioning score seems necessary to improve the power of the HR procedure. However, an appropriate choice of conditioning score appears to be difficult due to the use of the covariances of all item pairs in the HR procedure.

Stout's T Statistics (T_1 and T_2)

Among the nonparametric analytical indices, Stout's T_1 (the original Stout's T) (Stout, 1987) and Stout's T_2 (the refined Stout's T) (Nandakumar, 1987; Nandakumar & Stout, 1993) are two dimension assessment indices that have received much attention from researchers. In the following subsections, the assumption underlying the two indices is presented first and then a detailed review of Stout's T_1 and Stout's T_2 follows.

The Assumption of Essential Local Independence (Stout, 1987) The assumption of essential local independence and essential unidimensionality (Stout, 1987) underlying Stout's T_1 and Stout's T_2 implies that for a given subset of item responses from an item pool, the average absolute conditional (on Θ) covariances of responses to item pairs approach zero as the length of the subset increases. It is formally defined as:

$$\lim_{N\to\infty}\binom{N}{2}^{-1}\sum_{l\le j}\sum_{k\le N}\left|Cov(U_j,U_k)\,|\,\Theta=\theta\right|\to 0 \qquad (10)$$

Essential unidimensionality and essential local independence are said to be the weaker forms of traditional IRT assumptions of unidimensionality and local independence (Nandakumar,

1994). Essential local independence requires that as test length goes to infinity, the average covariance, $|Cov.(U_i,U_k|\Theta=\theta)|$, over item pairs is small in magnitude for all θ. With the strong principle of local independence in IRT, all major and minor abilities influencing item responses have to be considered when assessing local independence, whereas in essential dimensionality it is sufficient to consider only the influence of dominant dimensions.

Because of the nature of inherent multidimensionality that is frequently found in real test data (Humphreys, 1985, 1986; Reckase, Ackerman, & Carlson, 1988; Traub, 1983; Yen, 1984, 1985, 1993), the stringent traditional IRT assumptions of unidimensionality and local independence are rarely met. Thus, Stout (1990) argues that the traditional IRT assumptions of unidimensionality and local independence can be replaced by the weaker but more practical and psychologically more appropriate assumptions of essential unidimensionality and essential local independence, respectively.

Stout's T_1 Stout's T_1 statistic, developed by Stout (1987), tests the hypothesis of essential unidimensionality through assessing the likelihood that the given set of item responses comes from an essentially unidimensional item pool (Stout, 1987). A set of test data is said to be essentially unidimensional if the conditional covariances at any fixed level of θ are sufficiently small at some statistical significance level. Computational procedures are implemented in the computer program DIMTEST (Nandakumar & Stout, 1993). The general procedure for testing the null hypothesis of essential unidimensionality is briefly described below:

The n test items are split into three subtests: Assessment 1 subtest (AT1), Assessment 2 subtest (AT2), and a partitioning subtest (PT). AT1 and AT2 each include approximately one fourth of the items. PT consists of the remaining items. Items in AT1 are selected so that they

are dimensionally homogeneous and at the same time dimensionally different from the items in PT. Items in AT2 are selected to have a difficulty distribution similar to the items in AT1 to offset the statistical bias in AT1 arising due to short test length and extremely high or low difficulty level of AT1 items. The subtest PT is used to group examinees into K subgroups, which is used as the conditioning score in the computation. AT1, AT2, and PT would be of a similar dimension only if the null hypothesis is true.

The AT item selection can be done using either expert opinion or exploratory factor analysis. Examinees are grouped according to their PT scores. Item responses of AT1 are used to calculate two variance estimates: the usual variance estimate (σ^2_k), and the unidimensional variance estimate ($\sigma^2_{U.k}$) within each subgroup. Their difference is standardized and summed across subgroups to arrive at the statistic T_{1a}.

$$T_{1a} = \frac{1}{K^{1/2}} \sum_{k=1}^{K} \left[\frac{\hat{\sigma}^2_k - \hat{\sigma}^2_{u.k}}{S_k} \right] \qquad (11)$$

where

σ^2_k is the variance estimate of the AT1 subtest among examinees in the subgroup k and $\sigma_{u.k}^2$ is the estimate of the unidimensional variance computed by summing the item variances of subtest AT1. The standard error of estimate for subgroup k, S_k, is given by

$$S_k^2 = \left[(\hat{\mu}_{4.k} - \hat{\sigma}_k^4) + \hat{\delta}_{4.k} / M_4 + 2\sqrt{(\hat{\mu}_{4.k} - \hat{\sigma}_k^4)\hat{\delta}_{4.k} / M^4} \right] / N_k \qquad (12)$$

where

$$\hat{\mu}_{4,k} = \sum_{j=1}^{N_k} (Y_j^{(k)} - \overline{Y}^{(k)})^4 / N_k$$

and

$$\hat{\delta}_{4,k} = \sum_{i=1}^{M} (1 - \hat{p}_i^{(k)})(1 - 2\,\hat{p}_i^{(k)})^2$$

Similarly, items in AT2 are used to compute two variance estimates, σ^2_k and $\sigma_{u.k}{}^2$, and their difference is standardized and summed across subgroups to arrive at the statistic T_{1b}. Finally, Stout's T_1 statistic is given by:

$$T_1 = (T_{1a} - T_{1b})/\sqrt{2} \qquad (13)$$

If the test is essentially unidimensional, the differences between the usual variance estimate and the unidimensional estimate computed from AT1 would be small. If the test is multidimensional, the differences between the two variance estimates would be large and T_1 would be greater than Z (1.96) at a significance level of .05. The null hypothesis of essential unidimensionality is then rejected.

Nandakumar (1987) found that Stout's T_1 statistic did not perform well in the presence of guessing combined with high discriminations. Thus, an improved version, Stout's T_2, was developed. In the following section, Stout's T_2 is described.

Stout's T_2 There are three major improvements in Stout's T_2 over Stout's T_1.

1) The corrections for high-discrimination bias. According to Nandakumar (1987) and Stout and Nandakumar (1993), the presence of guessing and high discrimination items caused misclassification of examinees in the formation of subgroups. This inflated Stout's T_1. Nandakumar and Stout (1993) stated that including easy items in PT scores can effectively reduce the bias in classifying examinees. The correction for high-discrimination bias involves using the Wilcoxon Rank Test to rank the items according to their difficulty values. The very easy items from AT1 are selected and replaced with more difficult but dimensionally homogeneous items. This results in easy items in PT.

2) Automation of the length of AT1 and AT2. For Stout's T_1, the sizes of subtests are specified by the user a priori. Nandakumar and Stout (1987) developed an algorithm to automatically determine the size of assessment subtests. The purpose of this refinement is to provide a more convenient procedure for the determination of the sizes of the subtests. Exploratory factor analysis is carried out on the item responses. Items are selected according to the magnitude of item loadings on the second extracted factor. According to Nandakumar (1987) and Nandakumar and Stout (1993), the ratio of the size of AT1 to the total test length = 1/4 and the minimum factor loading on the second factor = .15 give the most reliable variance estimates for the automating test length procedure and yield a low Type I error rate and better power.

3) Correction of the standard error of estimate. As mentioned above, Stout's T_{1a} is obtained by summing the difference between the usual variance estimate and the unidimensional variance estimate across subgroups and then normalizing it. Stout (1987) suggested that under normal conditions when essential unidimensionality holds, T_{1a} and the final statistic T_1 should be normally distributed as the numbers of examinees and items approach infinity. When essential

unidimensionality is violated, T_{1a} and T_1 should have asymptotical power.

However, Stout (1987) showed that for test length and examinee population size typically encountered in practice, the statistic T_1 falsely rejected the hypothesis of unidimensionality more frequently than the nominal error rate. Thus, to correct this deficiency, two modifications were made. First, enlarge S_k to S_k' so that the values of T on average would be smaller, thereby reducing the rate of occurrence of Type I error to close to or below the nominal level. S_k' is obtained by

$$(S_k')^2 \equiv [(\hat{u}_{4,k} - \hat{\sigma}_4^k) + \frac{\delta_{4,k}}{M^4}]/N_k . \qquad (14)$$

Second, normalize each difference between σ^2_k and $\sigma^2_{U,k}$ by its estimated standard error and then sum instead of first summing and then normalizing. Nandakumar (1987) and Nandakumar and Stout (1993) found that the statistic $T_2=(T'_{2a}-T'_{2b})\sqrt{2}$ constructed with S_k' yielded an appropriate Type I error close to the nominal level and with greater power than Stout's T_1.

Other Indices Based on Stout's Essential Unidimensionality

A number of descriptive statistics based on Stout's essential unidimensionality have been developed recently. These indices include proximity measures used with Agglomerative Hierarchical Cluster Analysis (HCA) (Roussos, 1995; Roussos, Stout, & Marden, 1998). Specifically, these authors recommended three proximity measures: Pccor, a measure based on patterns of positive and negative conditional correlation, Pccov, a measure based on conditional item pair covariance, and P_{MH}, a measure based on the Mantel-Haenszel log-odds ratio. Other indices are Dimensionality Evaluation to Enumerate Contributing Traits (DETECT) (Stout,

Habing, Kim, Roussos, & Zhang, 1996) and Δ, an Unbiased Index of the Amount of Multidimensionality (Gao & Stout, 1997). Developed from the same assumption, each index has a different purpose. The purpose of the indices used with HCA is to discover multidimensionality structures by simultaneously identifying the independent item clusters that combine to produce simple structure. DETECT is especially designed to assess the amount of multidimensionality present in a test that includes several dimensionally different item clusters and allows the determination of whether the test exhibits approximate simple structure. While DETECT measures the amount of multidimensionality in the entire test, Δ is designed to determine the magnitude of departure from essential unidimensionality, i.e. how dimensionally disparate a set of items is relative to the direction of the best measurement presented by PT in Stout's T_1 or Stout's T_2 procedures (Zhang & Stout, 1996) after rejection of the unidimensional hypothesis on the test.

These indices are informative as measures of the amount of multidimensionality and the structure type present in the test. However, they do not offer statistical tests of the importance of the amount of multidimensionality present in the test and may not be used as a criterion for decision making. However, as Roussos, Stout , and Marden (1998) pointed out, these indices can be valuable aids to existing procedures such as Stout's T statistics and nonlinear factor analysis.

Summary of the Indices for Dimension Assessment

In the above sections, procedures and indices that have been proposed for test dimension assessment were presented. A number of earlier used indices such as those based on answer patterns, on reliability, and on principal component and linear factor analysis are not recommended for use by researchers because of their lack of rationale and ineffectiveness in

assessing dimensionality (De Champlain, 1992; Gessaroli & De Champlain, 1996; Hattie, 1984, 1985; McDonald, 1995).

Two groups of indices have been suggested by Hattie (1984, 1985) and McDonald (1995). They are indices based on item residual covariance/correlation after fitting a parametric model to a set of test data and indices that do not require estimation of the model parameters, particularly those based on conditional covariance nonparametric multidimensionality assessment methods. However, most of the indices are of a descriptive nature. The major drawback of these indices is that they do not offer a criterion for decision making, thus limiting their use to auxiliary approaches to those methods involving statistical tests. These indices include AIC (Akaike, 1987), The Unbiased Relative Fit Index (h_k) (McDonald, 1995), The Mean Absolute Residual and Standardized Mean Absolute Residual (Gessaroli, 1995; Hattie 1984, 1985; Hambleton & Rovinelli, 1986; McDonald, 1995; Nandakumar, 1994), HCA proximity measures (Roussos, 1995; Roussos, Stout, & Marden, 1998), DETECT (Stout, Habing, Kim, Roussos, & Zhang, 1996), and Δ an Unbiased Index of the Amount of Multidimensionality (Gao & Stout, 1997). It appears that one of the important tasks is to set up some reliable criterion for determining the number of test dimensions. Therefore, no attempt was made in the present study to assess the descriptive indices described above.

There are a number of indices or procedures that provide a statistical significance test of dimensionality. Among these indices are Christoffersson and Muthen's GLS Procedure (Christoffesson, 1975; Muthen, 1978), the HR Procedure (Holland & Rosenbaum, 1986; Rosenbaum, 1984), the Approximate χ^2 based on McDonald's nonlinear factor analytical models (De Champlain, 1992; Gessaroli & De Champlain, 1996), the Approximate Likelihood Ratio χ^2

(Gessaroli, De Champlain, & Folske, 1997), Stout's T_1 (Stout, 1987), and Stout's T_2 (Nandakumar, 1987; Stout & Nandakumar, 1993) based on nonparametric multidimensionality assessment of conditional covariance. Christoffersson and Muthen's GLS procedure is less favorable as its use is limited to a small number of items due to the nature of the GLS procedure. The HR procedure has been criticized for being too conservative for assessing unidimensionality and less powerful with multidimensional tests with high correlation of dimensions. The four most promising indices that offer a statistical significance test of dimensionality appear to be the Approximate χ^2, Stout's T_1, Stout's T_2, and the Approximate Likelihood Ratio χ^2. A special section is devoted to the detailed review of the studies related to the four indices.

Studies Related to Stout's T_1, Stout's T_2, the Approximate χ^2, and the Approximate Likelihood Ratio χ^2

Stout (1987) carried out a Monte Carlo study to evaluate Type I error and the power of Stout's T_1 statistic in assessing test dimensionality. In the unidimensional case, four variables were used: 1) test type, item responses were simulated to be similar to five tests: SAT Verbal (SATV), ACT Math Usage (ACTM), ACT English Usage (ACTE), ASVAB Auto and Shop Information (ASVAB AS), and ASVAB Arithmetic Reasoning (ASVAB AR). The length of the 5 tests varied. The minimum test length used was 25 items and the maximum test length was 50 items. Variables studied included mean discrimination values ranging from 0.72 to 1.46 and SDs ranging from 0.25 to 0.7; 2) sample size (750 and 2,000); 3) the size (M) of Assessment Subtest 1 (AT1) and Assessment Subtest 2 (AT2) that are used to compute Stout's T_1 (16% and 24% of the total test items respectively were used as M); and 4) Item Response Functions (item response functions were simulated using the three parameter logistic model and piecewise linear models).

Guessing parameter was set at .20 uniformly. Stout found that for most simulated tests when dimension (d) = 1, Type I error rates were kept below the nominal level regardless of sample size, test type, and model used to simulated data sets.

In the two dimensional case, the additional variables studied were dimension correlation (.5 and .7) and guessing (0 and .20). For the five test types, the discrimination parameters (a_{1i} and a_{2i}) of the two dimensions for each item were generated independently. Of the five two dimensional tests, two tests involved simple test structure simulated using a piecewise linear model and three had complex test structure simulated with a 3-parameter logistic model. For the simple structure tests, two levels of dimension dominance were used: 50% of items were related to dimension 1 and 50% to dimension 2 or 75% of items were related to dimension 1 and 25% items to dimension 2. For the complex structure tests, each test was taken to consist of pure items dependent on dimension 1, pure items dependent on dimension 2 alone and complex items dependent on both dimensions. Each pure item cluster and complex item cluster within a test included one third of the total test items. For both unidimensional and multidimensional studies, for each combination, 100 replications were carried out. However, test length was not examined.

In the multidimensional case, power was found to increase as dimension correlation decreased and sample size increased. Dimension dominance was also found to affect the rejection rates. Higher rejections were found with 50:50 dimension dominance than 75:25 dimension dominance. The findings also showed that Stout's T_1 was less powerful when guessing combined with high discrimination values was used. For example, for one of the two dimensional complex test structures with a mean discrimination value of 1.46 and a guessing parameter .2, the correct rejections were .67. When guessing was removed, the correct rejections

increased to 94. Since the simple structure tests and complex structure tests were not simulated under equal conditions the findings for the two test structures were not directly comparable.

Nandakumar and Stout (1993) reported two Monte Carlo studies. The first study was to examine the effect of guessing combined with high discrimination on the Type I error rate of Stout's T_1. These authors divided an 80 item SATV into two sets. The first set consisted of items having discrimination parameters greater than 1.0 (high discrimination) and the second set consisted of items having discrimination parameters equal to or less than 1.0 (low discrimination). Three sample sizes were used (750, 1,000, and 2,000). Stout's T_1 was applied separately to the two data sets. One hundred replications were carried out. Findings showed that the number of rejections for the test with low discrimination was less than .05. However, the rejection rate for the test with high discrimination far exceeded the nominal level. The rejection rate was .28 for sample size 750, .46 for sample size 1,000, and .58 for sample size 2,000.

The purpose of the second Monte Carlo study was to compare Type I error and the power of Stout's T_1 and Stout's T_2. One hundred replications were carried out. Both simulated and real data sets were used. For the simulated data sets, similar experimental conditions used in Stout (1987) were replicated for both unidimensional and two dimensional data.

With the simulated data sets, for the unidimensional case, Nandakumar and Stout (1993) found that the Type I error rate of Stout's T_2 was slightly higher than Stout's T_1 but still close to the nominal level. When guessing was coupled with high discrimination values (SATV), the Type I error rate dropped from .28 to .08 for the sample of 750 and from .58 to .07 for the sample of 2,000 compared with the findings from the preliminary study by Nandakumar and Stout (1993).

In the multidimensional case, Stout's T_2 had better power than Stout's T_1 for every test type, sample size, and level of correlation. On average, power went up from .67 for Stout's T_1 to .88 for Stout's T_2 per 100 replications when r=.5 and sample size of 750 were used, from .92 for Stout's T_1 to .99 for Stout's T_2 for the case of r=.5 with 2,000 examinees, and from .36 for Stout's T_1 to .54 for Stout's T_2 for the case of r=.7 with 2,000 examinees. Both sample size and correlation between traits were found to affect the power of Stout's T_2. Larger sample size (2000) was associated with more rejections and high correlation (.7) was related to fewer rejections. Discrimination and test length effects were not assessed.

In addition, four actual data sets were used. These tests included Arithmetic Reasoning tests for grades 10 and 12 (AR10 and AR12). Each test had 30 items. The other two data sets were ACT mathematics Form B and C (F29B and F29C). Each set had 40 items. With the real data sets, out of 100 replications the rejections were .06 for AR10, .13 for AR12, .86 for F29B, and .82 for F29C. Therefore, Nandakumar and Stout (1993) concluded that AR10 and AR12 were essentially unidimensional while F29B and F29C appeared to be multidimensional.

Based on this study, Stout's T_2 appeared to be more robust than Stout's T_1 to high-discrimination parameters coupled with guessing and more powerful in identifying the multidimensional nature of the data sets. However, the only variables used in this study were sample size and dimension correlation. Thus, no information was provided concerning the effects of test length, discrimination, and different kinds of complex test structures on the performance of Stout's T_1 and Stout's T_2. For example, in the multidimensional case, only one kind of complex test structure was used. Within the structure each dimension was set as equally important, which is not common in real testing situations.

Nandakumar (1994) compared the Type I error rate and power of Stout's T_2 to the HR approach and nonlinear factor analysis based on a single replication. Both real and simulated data sets were used. Three unidimensional and four two dimensional data sets, all with 2,000 examinees, were generated. The only variable used in the unidimensional case was test length (25, 40, and 50 items). In the multidimensional case two variables were used : test length (25 and 50 items) and dimension correlation (.3 and .7). Eight different real data sets were used: four of them were expected to be unidimensional and the other four were expected to be two dimensional.

Nandakumar (1994) found that all three methods correctly confirmed unidimensionality but differed in their ability to detect the lack of unidimensionality. Stout's T_2 showed better power than the other two methods. HR procedure and nonlinear factor analysis showed good power, provided the correlation between abilities was low (r=.3).

In this study, factors such as discrimination level and guessing that might affect the performance of the three procedures were not examined. In addition, test structure is an important factor that should have been considered. The results from this study were based on a single replication. The possibility of the result due to chance may not be excluded.

Hattie, Krakowski, Rogers, and Swaminathan (1996) carried out a simulation study to evaluate factors that affect the computation and rejection rates of Stout's T_2. Both unidimensional and multidimensional data were simulated. The unidimensional data set included 35 items from a unidimensional domain with discrimination all equal to 1.0. The variables examined for the unidimensional study were: three methods for ATI item selection (principal component analysis, refined tetrachorics, and nonlinear factor analysis), model used to simulate the data sets

(compensatory model and partially compensatory model), difficulty (-1 to 1 and -2 to 2), and guessing (0 and .15).

In the multidimensional study, both two dimensional and three dimensional data were simulated. The two dimensional data set included 18 items from a first dimension and 17 items from a second dimension. The three dimensional data set included 12 items from a first dimension, 11 items from a second dimension, and 11 items from a third dimension. For both two dimensional data and three dimensional data sets, item discrimination values were set at 1.0 uniformly. In addition to the variables used in the unidimensional study, an additional variable examined was dimension correlation (.1, .3, and .5). However, these correlation levels were nested within each of the two dimensional and three dimensional structures. For example, for d=2, correlation was set at .1, .3, and .5; and for d=3, the same correlation levels were used. For each data set, 15 replications were carried out. The responses of 1,000 examinees were generated for each replication.

These authors found that for the unidimensional data, principal component analysis and refined tetrachorics tended to select items for AT1 with more extreme b values under all conditions. The nonlinear factor procedure based on maximum likelihood, on the other hand, tended to select items that provided best fit. For the two dimensional data, principal component analysis and refined tetrachorics were more likely to select items from a single dimension than the nonlinear factor analysis. The number of items selected across the three item selection methods was not affected by the varying levels of b and c, the models used to simulate the examinees' responses, and the number of test dimensions. However, the percentage of the selected AT1 items that belong to the same dimension decreased with an increase in the range of

b. higher dimension correlation, and with data simulated using a partially compensatory model.

Data simulated using a partially compensatory model resulted in a large number of nonpositive definite matrices. This situation was more pronounced when the guessing rate of .15 was used. Under this condition, .96 or more matrices were nonpositive definite. The compensatory model produced only 4% nonpositive definite matrices. According to these authors, extremely difficult items can also lead to a high incidence of nonpositive definite matrices.

They found that the effect of nonpositive definiteness on the power of Stout's T_2 was dramatic. With the two and three dimensional data sets, the average value of T_2 was .52 when the matrices were not positive definite and 2.64 with positive definite. This means T_2 incorrectly identified multidimensionality as essential unidimensionality when nonpositive definite matrices were used.

ANOVA was carried out to assess the effect of the factors on T_2 values for the three methods for ATI item selection (principal component analysis, refined tetrachorics, and nonlinear factor analysis) . It was found that for principal component analysis the largest amount of variance was due to model used to simulate the data. The next important factor was dimension. However, since dimension correlation was nested within the number of dimensions and treated as a single factor, the effect of dimension may be confounded with correlation. In addition, ANOVA findings also showed that the interactions among most independent variables were significant. Two way interactions were those between dimension by model, dimension by c, and b by model; three way interactions were those between dimension by b by c, dimension by b by model, and dimension by c by model. A similar main effect was found for the other two

analysis methods. However, fewer interaction effects among the independent variables were found. Nonlinear factor analysis showed the least interaction effects with only dimension by model and b by model being significant effects. The post-hoc analysis of the interaction effects were not reported in the study. These authors concluded that T_2 was sensitive to data simulation model and the number of dimensions. They indicated that T_2 was not appropriate for the data simulated using a partially compensatory model.

When the rejection rates were examined, Stout's T_2 was found to have a higher Type I error rate than the nominal level (alpha = .05) under most conditions. The average rates were .15. The power was between .85 and .89 and from .76 to .78 respectively for two dimensional and three dimensional data across three AT1 selection methods. Dimension correlation, data generation models, and guessing each had an impact on the power of Stout's T_2. Higher power was associated with a decrease in dimension correlation and guessing and the use of the data simulated with the compensatory model. When the data based on the compensatory model were used, the range of b parameters did not affect the power of Stout's T_2. However, with the partially compensatory model, the use of lower b (-1, 1) resulted in much lower power for Stout's T_2 with three dimensional data being more affected.

These authors concluded that Stout's T_2 was only applicable for identifying dimensionality of data simulated using a compensatory model. When the matrix was not positive definite, Stout's T_2 was weakened and should not be used. The incidence of nonpositive definite matrices was high when the data were simulated using a partially compensatory model or when everyone answered all items correctly or when examinees use much guessing. In other words, nonpositive definite matrices are most likely to occur when both extremely easy and difficult

items were present.

Gessaroli and De Champlain (1992, 1996) carried out a Monte Carlo study comparing the Type I error and power of the Approximate χ^2 and Stout's T_2 with simulated unidimensional test structure and two dimensional simple test structures. Variables used in the unidimensional case were test length (15, 30, and 45 items), sample size (500 and 1000), and discrimination strength specified by three levels of discrimination (low discrimination: μ_a=.72 and σ_a=.06; moderate discrimination: μ_a = 1.07 and σ_a=16; and high discrimination: μ_a= 1.46 and σ_a=.26).

In the two dimensional case, simple test structure was simulated with test length, discrimination, and sample size as in the unidimensional case. The additional variables were dimension dominance defined by number of items related to each dimension: 50:50, half of the items related to dimension 1 and half to dimension 2; and 80: 20, 80% of items loaded on the first dimension and 20% on the second dimension. Correlation between dimensions was set at 0 and no guessing parameter was used. One hundred replications were carried out for both unidimensional and multidimensional data.

These authors found that the Approximate χ^2 had better control of Type I error rate than Stout's T_2 under all conditions simulated with a maximum Type I error rate of .04 while Stout's T_2 had a maximum Type I error rate of .08. Test length, sample size, and test type had no effect on Type I error rate of the two indices.

With the two dimensional data sets, both indices showed excellent power except for the condition of short test length (15 items). At this level, sample size, discrimination strength, and dimension dominance had no strong effect on the power of the Approximate χ^2 while the power of Stout's T_2 was found to be affected. With a sample size of 500 and low discrimination

strength. Stout's T_2 showed a low rejection rate (19 out of 100 replications) for dimension dominance (80:20). When dimension dominance (50:50) was used, the rejections of Stout's T_2 increased to 69. While under the former condition, the Approximate χ^2 showed 55 rejections and under the latter, 99 rejections.

Gessaroli and De Champlain (1992; 1996) showed that the Approximate χ^2 was as powerful as Stout's T_2 with multidimensional simple structure data sets when large sample size (1000) and longer test length (over 30 items) were used and more powerful when small sample size, short test length, low discrimination strength, and strong dominance on dimension 1 were used. However, these findings may not be generalizable to wider conditions and further study is needed. For example, some related questions are: Is the power of the Approximate χ^2 affected by multidimensional complex test structures? How does it compare with Stout's T_1 and Stout's T_2 under the conditions of multidimensional complex structures? So far, no relevant information is available.

Breithaupt (1995) did a similar study assessing the performance of the Approximate χ^2, Stout's T_1, and Stout's T_2 with unidimensional and two dimensional simple structure. In the unidimensional assessment, the responses of 1000 examinees were simulated with varied test length (30 and 45 items) and guessing (ranging from .00 to .25). The unidimensional data sets were simulated with μ_a=.72 and σ_a,=.025 and μ_b=0 and σ_b=.96. In the two dimensional assessment, the additional variable used was dimension correlation (0, .5, and .7). The first dimension was set as dominant with an item ratio of 80:20. For both unidimensional and multidimensional assessments, one hundred replications were carried out.

With unidimensional data, the Approximate χ^2, Stout's T_1, and Stout's T_2 statistics

yielded acceptably low Type I error rates (below .05) under all conditions. Test length and guessing did not significantly influence the performance of the three statistics.

With multidimensional data sets, each of the factors (test length, correlation, and guessing) was found to affect the performance of the Approximate χ^2, Stout's T_1, and Stout's T_2. Longer tests were related to higher rejection rates. Higher correlations and guessing were associated with fewer rejections. For all three indices, dimension correlation produced the strongest effect. As an example for the no guessing condition, on average, for the Approximate χ^2, the rejection rate for r=0 was 100%, for r=.5 was 93.5%, and for r=.7 was 18.5%; for Stout's T_1, the rejection rates for r=0, r=.5, and r=.7 were 95.5%, 85%, and 51.5% respectively. For Stout's T_2, the rejection rates for r=0, r=.5, and r=.7 were 96%, 86.5%, and 55.5% respectively.

According to Breithaupt, Stout's T_1 and Stout's T_2 did not differ substantially from each other in terms of the rejection rates for all conditions. This finding is somewhat inconsistent with the findings provided by Nandakumar and Stout (1993). They found that the Stout's T_2 was especially robust against the structure involving high discriminations combined with guessing and was more powerful than Stout's T_1. The disagreement in the findings is likely due to different test structures used by the authors of the two studies. Nandakumar and Stout (1993) used complex structure wherein the two dimensions were set equally important whereas Breithaupt (1995) used simple structure with the first dimension set as dominant (80:20). Another possible cause was the relatively low discrimination level (mean=.72) that was used in Breithaupt's study. If a high discrimination level were used, Stout's T_1 and Stout's T_2 might yield different results. Further study comparing the two Stout's indices and the Approximate χ^2 under wider conditions of test structure and item discrimination is warranted.

Gessaroli, De Champlain, and Folske (1997) carried out a Monte Carlo study to compare the Type I error rate and the power of an Approximate Likelihood Ratio χ^2, the Approximate χ^2, and Stout's T_2. These authors used three test lengths (15, 45, and 75 items) and two sample sizes (1000 and 5000) for both unidimensional and multidimensional simple test structure data sets. For the unidimensional data set, five levels of item difficulty were used: -1.8, -1.0, 0, 1.0, and 1.8. Discrimination parameters of all items were set at .80. This pattern of parameter input was replicated 3, 9, and 15 times for the 15, 45, and 75 item test length conditions, respectively.

For the two dimensional data sets, a simple structure was used with 80% of items related to the first dimension and 20% related to the second dimension. The correlation between dimensions was set at .7. Item discrimination structure for dimension 1 and dimension 2 was defined as a_1=0.8 and a_2=0.0 for 80% of the items while a_1=0.0 and a_2=0.8 for the remaining items of the test. One hundred replications were carried out.

With the unidimensional data sets, the Approximate Likelihood Ratio χ^2 and the Approximate χ^2 had similar Type I error control. Both seemed to be quite conservative in incorrectly rejecting the assumption of unidimensionality (maximum one rejection for the Approximate Likelihood Ratio χ^2 and zero rejections for the Approximate χ^2). Stout's T_2 was not assessed with 15 item tests. Similar Type I error rates were observed. Test length and sample size had no effect on the Type I error rate of the three indices.

With the two dimensional data sets, the Approximate Likelihood Ratio χ^2 was found to be more powerful than the Approximate χ^2 with a sample size of 1000 and showed similar power with the sample size of 5000. However, test length affected the two indices differently. With sample size of 1,000, for the Approximate χ^2, power of .10 was observed for the 15 item test.

When test length increased to 45 and 75 items, power decreased to .02 and .03 respectively. The Approximate Likelihood Ratio χ^2 showed higher power with .31 for the 15 item test, .48 for the 45 item test, and .79 for the 75 item test. With a sample size of 5,000, perfect power of 1.0 was obtained for both indices across the three test lengths. Again, Stout's T_2 was only assessed with the 45 item test and 75 item test due to program error and only the sample size of 5,000 was assessed with the 75 item test. For the 45 item test, sample size influenced the power of Stout's T_2. The power was .44 for a sample size of 1,000 and 1.0 for a sample size of 5,000. For large sample size, power was not affected by test length. Similar power (1.0) was obtained for 75 item test used with sample size of 5,000. Thus, these authors concluded that the Approximate Likelihood Ratio χ^2 was as powerful as Stout's T_2 and showed much improvement over the Approximate χ^2 with highly correlated data.

As a newly proposed index, the Approximate Likelihood Ratio χ^2 has not been widely studied. Its potential as a dimension assessment method has not been fully explored. For example, in this study, the experimental conditions were limited to only one level of discrimination and dimension correlation. More work should be carried out to assess the use of the Approximate Likelihood Ratio χ^2 under varied conditions.

De Champlain and Gessaroli (1998) examined empirical Type I error rates and power of three dimensional assessment procedures with short tests and small sample sizes. The three procedures used were the Approximate χ^2 (De Champlain & Gessaroli, 1996), a chi-square goodness of fit statistic (Browne, 1977) found in LISREL8, and a marginal maximum likelihood procedure (Bock & Aitken, 1981) found in TESTFACT (Wilson, Wood, & Gibbons, 1991).

In the unidimensional case, dichotomous item responses were simulated using a three-

parameter logistic IRT model. Variables manipulated were test length (20 and 40 items) and sample size (250, 500, and 1,000 examinees). The 40 item tests were composed of two 20 item tests with identical item parameters randomly selected from one form of the Law School Admission Test (LSAT). Discrimination values ranged from .34 to 1.15 and item difficulty values ranged from -1.71 to 1.54. The maximum value for guessing was .22. One hundred replications were carried out for each combination of test length and sample size.

In the multidimensional case, in addition to the variables used in the unidimensional study, two different test structures (simple and complex) were simulated. For both structures, correlation was set at .0 and .7. With the simple test structure, 25% of the items were related to the first dimension and 75% to the second dimension with mean discrimination values of .81 and .72 respectively. Item difficulty values and guessing were similar to those used in the unidimensional case. The complex test structure was simulated with 25% of the items measuring only the first dimension, 25% of the items measuring only the second dimension, and 50 % of the items measuring both dimensions. For each of the items measuring both dimensions, item discrimination was set the same for each dimension. The mean discrimination parameter for these items was 1.2. For items measuring only the first dimension, the mean item discrimination was .81 and for the items measuring only the second dimension the mean item discrimination was .72. When the items discriminations (including both unique and complex items) were averaged for each dimension, the mean discrimination for the first dimension was .78 and for the second dimension was .75. That is, both dimensions were set approximately equally important. Similarly, one hundred replications were carried out for each cell of 2 (test length) x 3 (sample size) x 2 (test structure) x2 (dimension correlation) design.

The results showed that the Type I error rate for the Approximate χ^2 tended to be below or near the nominal alpha level (.05). Sample sizes 250 and 500 had no effect on the Type I error rate for both test lengths. Under these conditions, the maximum Type I error rate was .01. For sample size 1,000 the Type I error rate was .05 for the 20 item test and .07 for the 40 item test.

Type I error rate of LISREL8 χ^2 was severely inflated. For the cell of 20 items and 1,000 examinees, Type I error rate was .68. The Type I error rate for Llikelihood Ratio χ^2 test was also severely inflated and strongly affected by all the variables. For example, for 20 item tests, Type I error rate was .58 for sample size 250, .41 for sample size 500, and .17 for sample size 1,000. For 40 item tests, increasing sample size showed little improvement. Type I error rate was .79, .77, and .77 for sample sizes of 250, 500, and 1,000 respectively.

With the two dimensional data sets, the Approximate χ^2 showed extremely good power over all conditions with a minimum power of .99. LISREL8 χ^2 showed excellent power (1.0) over all conditions for data sets of 20 items and 1,000 examinees. The data sets of 40 items and other sample sizes were not tested with this index. With the simple structure, TESTFACT showed excellent power (1.0) over all conditions for zero dimension correlation. For both test structures, with high dimension correlation, power for TESTFACT increased as sample size and test length increased. For the simple structure, power ranged from .77 to .94 for 20 item tests and .97 to .99 for 40 item tests while for the complex structure, the range was from .62 to 1.0 for 20 item tests and .94 to .99 for 40 item tests.

According to De Champlain and Gessaroli (1998), the good power of TESTFACT and LISREL8 is somewhat attributable to the inflated Type I error rates. It appears that the Approximate χ^2 has more advantages over the other two indices with empirical Type I error rates

that were near the nominal alpha level in all conditions. However, the conditions used in this study were restricted. Most item discrimination parameters were between .7 to .8. The complex structure was simulated such that the two dimensions showed equal importance. All this restricted the generalization of the results. It would be interesting to assess the effects on these indices of other levels of item parameters and test structure where item parameters are distributed differently between the two dimensions.

Summary of Relevant Studies

The performance of the Approximate χ^2, has been assessed under varied conditions by a number of researchers (Breithaupt, 1995; De Champlain, 1992; De Champlain & Gessaroli, 1998; Gessaroli & De Champlain, 1996; Gessaroli, De Champlain, & Folske, 1997). The findings showed that this index had a Type I error rate below or near the nominal level (.05). The results concerning the power of this index were different. Most researchers have shown that the Approximate χ^2 demonstrates good power in correctly rejecting the unidimensional assumption with two dimensional simple test structure with a dimension correlation of .5 or below.

Recently, De Champlain and Gessaroli (1998) have shown that the Approximate χ^2 showed good power not only with two dimensional simple structure but also with complex structure of two equally important dimensions. However, most of these studies were carried out under limited conditions such as short test length and small sample size. The complex structure used in this study was similar to the one used by Nandakumar (1994), Nandakumar & Stout (1993), and Stout (1987), in terms of dimension importance to assess Stout's T_1 and Stout's T_2. More investigation must be carried out in order to better understand the performance of the Approximate χ^2 and its possible utility in more complex testing situations. One of the situations

that has not been extensively studied is two dimensional complex test structure where the importance of the second dimension gradually increases. This kind of structure is frequently found in math tests involving both purely computational questions and different kinds of problem solving questions.

The Approximate Likelihood χ^2 is another index suggested by Gessaroli, De Champlain, and Folske (1997) and used to test the residual correlation matrix obtained after fitting an m-factor nonlinear factor analytic model to a set of item responses using a polynomial approximation to a normal ogive. This index was said to be as powerful as Stout's T_2 and have stronger power than the Approximate χ^2 when used with two dimensional simple test structure with high dimension correlation (r=.7) (Gessaroli, De Champlain, & Folske, 1997). Only one study was carried out on this index (Gessaroli, De Champlain, & Folske, 1997). Clearly, more work needs to be done to understand the nature and performance of this index. Unfortunately, this index was not included in the present study due to the fact that no information about it was available until the data analysis for the present study was completed.

Studies concerning Stout's T_1 and Stout's T_2 showed different results. Stout's T_2 has been reported to be more robust to high discrimination coupled with guessing and more powerful in rejecting the unidimensional assumption than Stout's T_1 (Nandakumar & Stout, 1993). The findings, however, are not consistent with Breithaupt (1995), who found that Stout's T_1 and Stout's T_2 did not differ greatly in terms of either Type I error rate or power based on her 100 replications with two dimensional simple structure.

Nandakumar (1993) reported that both Stout's T_1 and Stout's T_2 had good control of Type I error rate. However, Hattie, Krakowski, Rogers, and Swaminathan (1996) showed that Stout's

T_2 had higher Type I error than the nominal level under most conditions. Other researchers have indicated that test length did not affect Type I error rate of Stout's T_2 (Breithaupt, 1995; Gessaroli & De Champlain, 1996) and Stout's T_1 (Breithaupt, 1995). These authors used a minimum test length of 15 items and a maximum test length of 45 items. The test lengths used by Nandakumar, (1994), Nandakumar and Stout, (1993), and Stout (1987) were also short and restricted to 50 items or less. No sufficient information is available concerning the effect of a test length longer than 50 items on Type I error rate of Stout's T_1 and Stout's T_2.

Most studies have shown that the power of Stout's T_1 and Stout's T_2 was affected by sample size, test length, guessing, and dimension correlation. A decrease in power was associated with a decrease in sample size and test length and an increase in guessing and dimension correlation. However, one of the limitations of most studies is that the test structure was restricted to two dimensional simple test structure (Breithaupt, 1995; Gessaroli & De Champlain, 1996; Gessaroli, De Champlain, & Folske, 1997; Hattie, Krakowski, Rogers, & Swaminathan, 1996). The only two dimensional complex test structure that has been used is one wherein both traits had equal importance (De Champlain & Gessaroli, 1998; Nandakumar, 1994; Nandakumar & Stout, 1993; Stout, 1987). Other types of complex test structure that are frequently found in practice should be examined together with possible factors that might affect the power of the Approximate χ^2, Stout's T_1, and Stout's T_2. For example, as mentioned before, one of the interesting test structures is one frequently found in a math usage test that includes items that mainly measure math skill but require gradually increased verbal skill as more words are used in the items.

Hattie, Krakowski, Rogers, and Swaminathan (1996) reported that Stout's T_2 was

sensitive to whether the multidimensional data arose from a compensatory model or a partially compensatory model and guessing. The tetrachoric correlations based on the data generated from a partially compensatory model or the data simulated with much guessing resulted in too many non-Gramian matrices. Stout's T_2 failed in these cases and hence was not appropriate.

In addition, information about the effect of specific levels of discrimination on the performance of Stout's T_1 and Stout's T_2 is not available. Clearly, further comparison of the performance of Stout's T_1 and Stout's T_2 under wider conditions are warranted.

There are only two studies where comparisons of the performance of the Approximate χ^2 and Stout's T_2 (Gessaroli & De Champlain, 1996; Gessaroli, De Champlain & Folske, 1997) were made. In addition, Breithaupt (1995) compared Stout's T_1 and Stout's T_2 to the Approximate χ^2. Stout's T_1 and Stout's T_2 were reported to have better power than the Approximate χ^2 with test data involving high dimension correlation (.7) while the Approximate χ^2 showed better power than Stout's T_1 and Stout's T_2 with test data having fewer items (15) and small sample size (500). However, the comparative studies were only carried out with unidimensional test structure and two dimensional simple test structure. Stout's T_1 and Stout's T_2 have been reported to be able to identify the correct number of dimensions with complex structure (Nandakumar, 1994; Nandakumar & Stout, 1993; Stout, 1987). De Champlain and Gessaroli (1998) have reported that the Approximate χ^2 showed higher power for two dimensional complex test structure. However, no study has been done to compare the performance of the three indices with complex test structure. To generalize the use of the three indices, further study should focus on comparing the three indices under the conditions of type of test structure, wider ranges of dimension correlation, sample size, and test length. As well, these indices should be assessed at specific

item discrimination levels.

Purpose of the Study

The purpose of this study, therefore, was to assess and compare the effectiveness of the Approximate χ^2, Stout's T_1, and Stout's T_2 statistics in dimensional assessment of unidimensional test structure, two-dimensional simple structure, and two dimensional complex test structure. Specifically, the following research questions were addressed:

1) What are the simple and interactive effects of test length, sample size, and item discrimination on the Type I error rate of the Approximate χ^2, Stout's T_1, and Stout's T_2?

2) What are the simple and interactive effects of test structure, sample size, test length, item discrimination, and correlation between dimensions on the power of the Approximate χ^2, Stout's T_1, and Stout's T_2 ?

Methodology for this study is presented in the next chapter.

Chapter III

METHODOLOGY

A Monte Carlo study was carried out to assess the Type I error rate and the power of the Approximate χ^2, Stout's T_1, and Stout's T_2 when applied in the context of different test structures. In this chapter, the model used to simulate data sets, the variables manipulated in the unidimensional and multidimensional assessment, data generation, and procedures for data analysis are presented.

The Model Used for Data Simulation

The two-dimensional compensatory model (Reckase, 1985), implemented in the computer program MD4F (Roussos, 1995), was used to simulate the data in the study. There were several reasons for the choice of a compensatory model: 1) A compensatory model allows high proficiency on one dimension to compensate for low proficiency on other dimensions in response to an item; 2) It has frequently been used to generate item responses and has been shown to adequately reflect the interaction between a person and an item (Ackerman, 1994; Breithaupt, 1995; De Champlain, 1992; Gessaroli & De Champlain, 1996; Gessaroli, De Champlain, & Folske, 1997; Hattie, Krakowski, Rogers, & Swaminathan, 1996; Nandakumar, 1994; Nandakumar & Stout, 1993; Reckase, 1985); 3) the data generated using a compensatory model were appropriate for Stout's T_1 and Stout's T_2 while other models may not be (for example, a partially compensatory model) (Hattie, Krakowski, Rogers, & Swaminathan, 1996).

In practice, multidimensional tests may include two kinds of test structure: simple test structure or complex test structure. Typical simple structure tests are those with items related to

distinct subject concepts. For example, a math test may include items measuring geometry ability and items related to algebra ability. Geometry and algebra are two distinct math abilities. One kind of complex test structure can be best represented by a math usage test that includes both pure calculation and problem solving items. In this case, for problem solving items, the requirement of a verbal ability (the second dimension) to respond to the items correctly gradually increases as the items involve more words. In this study, two kinds of test structures (simple and complex) were simulated. The simple structure test was simulated to be similar to a math test wherein 75% of the items in the test measured algebra ability and 25% of the items measured geometry ability. The complex structure test was simulated to be similar to a math usage test wherein 25% of the items in the test measuring pure computation ability and 75% of the items measuring problem solving ability. The two-dimensional compensatory model (Reckase, 1985) is thought to be able to model these testing scenarios appropriately.

The two dimensional compensatory model is formally given as

where

$$P(x_{ij}=1 \mid a_i, d_i, \alpha_j, \theta_j) = \frac{e^{(\sum_{k=1}^{n} a_{ik}\theta_j \cos\alpha_{jk} + d_i)}}{1 + e^{(\sum_{k=1}^{n} a_{ik}\theta_j \cos\alpha_{jk} + d_i)}} \qquad (15)$$

$P(x_{ij}=a_i, d_i, \alpha_j, \theta_j)$ is the probability of a correct response to item i by person j;

$\mathbf{a_i}$ is a vector of discrimination parameters;

d_i is a scalar parameter that is related to the parameter (MDIF) D;

α_j is an angle.

θ_j is a vector of ability parameters;

a_{ik} is the k_{th} element of $\mathbf{a_i}$;

$\theta_j \cos\alpha_{jk}$

is an element of θ_j, and α_{jk} gives the angle from the kth dimension to the point where the item response surface (IRS) has the maximum slope.

Multidimensional difficulty (MDIF) D_i is defined as:

$$D_i = \frac{-d_i}{[\sum_{k=1}^{n}(a_{ik})^2]^{0.5}}. \qquad (16)$$

where

D_i represents the distance between the origin of the m-dimensional ability space and the point in the space where the item information is maximum. The line joining this point to the origin is at an angle of α_{jk} to the kth ability dimension where

$$\cos\alpha_{ik} = a_{ik} / [\sum_{k=1}^{n} a_{ik}^2]^{0.5}. \qquad (17)$$

The multidimensional discrimination (MDISC) as a function of the slope of the IRS is defined at the steepest point in the direction from the origin of the θ_j . MDISC for the two-dimensional case is defined as (Reckase, 1985):

$$MDISC = \sqrt{a_{j1}^2 + a_{j2}^2}. \qquad (18)$$

The MDISC statistic is an overall measure of the capability of an item to distinguish among individuals who are in different locations in the θ space. If an item measures a single dimension, θ_1, then $a_{j1}>0$ and $a_{j2}=0$. The MDISC, then, is equal to the unidimensional discrimination parameter (Reckase, 1985). In the unidimensional case, the model would be equivalent to the traditional two parameter logistic model (Birnbaum, 1968). In the two dimensional cases, both a_{j1} and $a_{j2}>0$.

Variables Used in the Study

Unidimensional Study

The purpose of the unidimensional study was to assess Type I error of each of the three indices: the Approximate χ^2, Stout's T_1, and Stout's T_2. The variables manipulated in the unidimensional study were test length, sample size, and discrimination. They are described below:

Test Length Test length was set at 40 and 80 items. The number of items chosen reflects two test lengths that are frequently found in achievement tests. The test length used in most studies was no more than 50 items for the Approximate χ^2, Stout's T_1, and Stout's T_2 (Breithaupt, 1995; De Champlain, 1992; Gessaroli & De Champlain, 1996; Hattie, Krakowski, Rogers, & Swaminathan, 1996; Nandakumar, 1994; Nandakumar & Stout, 1993; Stout, 1987). The test length used in this study therefore provides additional information regarding the effect

on these indices of test length greater than 50 items.

Sample Size Three sample sizes (500, 1,000, and 2,000) were examined. The ability for examinees was randomly generated from a unit normal distribution. Previous researchers have indicated that Stout's T_1 and Stout's T_2 perform well with a sample size of 2,000 (Nandakumar, 1994; Nandakumar & Stout, 1993; Stout, 1987). Little is known, however, about the performance of the Approximate χ^2 with a sample size of 2,000 and how it compares with the performance of Stout's statistics. The Approximate χ^2 is based on a parametric estimation procedure. It may require a larger sample size than Stout's nonparametric statistics to obtain accurate parameter estimation. So far, little is known about the optimal sample size that produces accurate estimation results for each of these indices in assessing test dimensionality. One of the characteristics of a χ^2 test is its sensitivity to sample size. It is expected that the Type I error and the power of the Approximate χ^2 may be affected by larger sample sizes.

Item Discrimination Three levels of MDISC were selected (.7, 1.00, and 1.40). These values have been found frequently in real test data and have been used by earlier researchers to simulate test item responses (Ackerman, 1994; Breithaupt, 1995; De Champlain, 1992; Hattie, Krakowski, Rogers, & Swaminathan, 1996; Kim, Zhang, & Stout, 1995; Nandakumar, 1994; Nandakumar & Stout, 1993; Reckase, 1985; Reckase & Mckinley, 1991; Stout, Habing, Douglas, Kim, Roussos, & Zhang, 1996; Yen, 1984). The initial values for the unidimensional test structure are listed in Table 1, Table 2, and Table 3, where only the values for the first ten items are listed because the remaining items are repetitions of the first ten.

Item difficulty was fixed for the unidimensional test structure. There are ten levels of item difficulty ranging from -1.6 to 1.6 in gradually increasing increments. The range of these values

was chosen because these values are frequently found in the literature.

According to Hattie, Krakowski, Rogers, and Swaminathan (1996), the involvement of guessing leads to non-positive definite matrices in Stout's T_1 and Stout's T_2. Consequently, the tetrachorics can not be accurately estimated. Guessing is not used in the present study.

Table 1

Initial Item Parameters for Unidimensional Test Structure (MDISC=0.7)

D_i	d_i	Item	MDISC	a_{i1}	a_{i2}
1.60	-1.12	1	0.70	0.70	0.00
1.10	-0.77	2	0.70	0.70	0.00
0.65	-0.46	3	0.70	0.70	0.00
0.30	-0.21	4	0.70	0.70	0.00
0.05	-0.04	5	0.70	0.70	0.00
-0.05	0.04	6	0.70	0.70	0.00
-0.30	0.21	7	0.70	0.70	0.00
-0.65	0.46	8	0.70	0.70	0.00
-1.10	0.77	9	0.70	0.70	0.00
-1.60	1.12	10	0.70	0.70	0.00

Table 2

Initial Item Parameters for Unidimensional Test Structure (MDISC=1.0)

D_i	d_i	Item	MDISC	a_{i1}	a_{i2}
1.60	-1.60	1	1.00	1.00	0.00
1.10	-1.10	2	1.00	1.00	0.00
0.65	-0.65	3	1.00	1.00	0.00
0.30	-0.30	4	1.00	1.00	0.00
0.05	-0.05	5	1.00	1.00	0.00
-0.05	0.05	6	1.00	1.00	0.00
-0.30	0.30	7	1.00	1.00	0.00
-0.65	0.65	8	1.00	1.00	0.00
-1.10	1.10	9	1.00	1.00	0.00
-1.60	1.60	10	1.00	1.00	0.00

Table 3

Initial Item Parameters for Unidimensional Test Structure (MDISC=1.4)

D_i	d_i	Item	MDISC	a_{i1}	a_{i2}
1.60	-2.24	1	1.40	1.40	0.00
1.10	-1.54	2	1.40	1.40	0.00
0.65	-0.91	3	1.40	1.40	0.00
0.30	-0.42	4	1.40	1.40	0.00
0.05	-0.07	5	1.40	1.40	0.00
-0.05	0.07	6	1.40	1.40	0.00
-0.30	0.42	7	1.40	1.40	0.00
-0.65	0.91	8	1.40	1.40	0.00
-1.10	1.54	9	1.40	1.40	0.00
-1.60	2.24	10	1.40	1.40	0.00

Thus, there are 2x3x3 combinations (test length: 40 and 80, by sample size: 500, 1,000, and 2,000, by item discrimination: 0.7, 1.0, and 1.4) for item parameters. For each of the combinations, one hundred replications were carried out. Altogether, 1,800 data sets were simulated for the unidimensional study.

Multidimensional study

In the multidimensional case, in addition to the variables (test length, discrimination, and sample size) used in the unidimensional case, two more variables were considered: test structure and correlation between dimensions.

Test Structure Two types of test structure were simulated: 1) two-dimensional simple test structure. The simple structure items are simulated to be similar to a math test where each item measures only one dimension: geometry or algebra; 2) Two-dimensional complex test structure. This structure includes items that measure a unique dimension and mixed items that measure both dimensions. They are simulated to represent a typical math usage test where the magnitude

of the second (verbal) dimension measured by the items gradually increases as success on items relies more and more on verbal ability.

In order to make the two test structures comparable, the total probability of answering the items in the two types of structure should be the same. This is achieved by setting the ratio of the first dimension to the second dimension measured in both structures the same, 3:1. As mentioned previously, in the simple structure test, 75% of the items in the test measure dimension one, 25% measure dimension two. In the complex structure, 25% of the items in the test measure θ_1 and 75% of the items measure composites of θ_1 and θ_2. Of these items, 25% of the items require predominantly θ_1, 25% of the items require a decreased amount of θ_1, and 25% of the items require equal amounts of both abilities. Thus, for the items measuring θ_1 and θ_2 composites, α_{jk} ranges from 24 to 45 degrees in approximately 10 degree increments. The initial values for the two structure simulations with 40 items are presented in Tables 4, 5, and 6 for MDISC of .70, 1.0, and 1.4. The 80 item tests include two sets of 40 items.

Correlation Between Dimensions Correlations between dimensions were set to be .0, .4, .57, and .7. These values represent approximately equal intervals of variance held in common by the two dimensions. Researchers have reported different findings concerning the effect of the correlation between dimensions on item parameter estimation and model fit. Drasgow and Parsons (1983) indicated that multidimensional data when $r > .5$ can be analyzed by a unidimensional logistic model. Batley and Boss (1993) found that when $r = .7$, there was a great tendency to collapse space and it was difficult to distinguish between the dimensions. Breithaupt (1995) showed that Stout's T_1, Stout's T_2, and the Approximate χ^2 worked well with data when $r \leq .5$ but lost power when $r = .7$. Dimension correlation is a factor that affects the performance of

the three indices. This influence may vary with different types of test structure. It was expected that when correlation between dimensions was used with simple test structure the influence would be weaker than the influence produced with complex test structure because it is likely that when an item measures two dimensions there already exists some relationship between the two dimensions. When correlation is added, the relationship between the two dimensions is strengthened, which may make it more difficult to distinguish the two dimensions. However, this needs to be empirically verified. In this study, the comparison of the three indices was carried out between two dimensional simple test structure and two dimensional complex test structure; For both types of structure, the total probability of answering items correctly was set the same. This design should provide more accurate information concerning the effect of dimension correlation on the three indices. The initial values for the multidimensional test structures are listed in Table 4, Table 5, and Table 6.

The above variables resulted in a 2x2x3x3x4 design: test structure (simple and complex), by test length (40 and 80), by sample size (500, 1,000, and 2,000), by item discrimination (0.7, 1.0, and 1.4), by correlation (0, 0.4, 0.57, and 0.7). For each of these combinations, 100 replications were carried out. Thus, 14,400 multidimensional data sets were analyzed.

Table 4

Initial Item Parameters for Two Dimensional Simple and Complex Structures (MDISC=0.7)

D_i	d_i	Item	MDISC	a_{i1}	a_{i2}	α_{i1}	MDISC	a_{i1}	a_{i2}
			(Simple Structure)				(Complex Structure)		
1.60	-1.12	1	0.70	0.70	0.00	0	0.70	0.70	0.00
1.10	-0.77	2	0.70	0.70	0.00	0	0.70	0.70	0.00
0.65	-0.46	3	0.70	0.70	0.00	0	0.70	0.70	0.00
0.30	-0.21	4	0.70	0.70	0.00	0	0.70	0.70	0.00
0.05	-0.04	5	0.70	0.70	0.00	0	0.70	0.70	0.00
-0.05	0.04	6	0.70	0.70	0.00	0	0.70	0.70	0.00
-0.30	0.21	7	0.70	0.70	0.00	0	0.70	0.70	0.00
-0.65	0.46	8	0.70	0.70	0.00	0	0.70	0.70	0.00
-1.10	0.77	9	0.70	0.70	0.00	0	0.70	0.70	0.00
-1.60	1.12	10	0.70	0.70	0.00	0	0.70	0.70	0.00
1.60	-1.12	11	0.70	0.70	0.00	24.3	0.70	0.638	0.288
1.10	-0.77	12	0.70	0.70	0.00	24.3	0.70	0.638	0.288
0.65	-0.46	13	0.70	0.70	0.00	24.3	0.70	0.638	0.288
0.30	-0.21	14	0.70	0.70	0.00	24.3	0.70	0.638	0.288
0.05	-0.04	15	0.70	0.70	0.00	24.3	0.70	0.638	0.288
-0.05	0.04	16	0.70	0.70	0.00	24.3	0.70	0.638	0.288
-0.30	0.21	17	0.70	0.70	0.00	24.3	0.70	0.638	0.288
-0.65	0.46	18	0.70	0.70	0.00	24.3	0.70	0.638	0.288
-1.10	0.77	19	0.70	0.70	0.00	24.3	0.70	0.638	0.288
-1.60	1.12	20	0.70	0.70	0.00	24.3	0.70	0.638	0.288
1.60	-1.12	21	0.70	0.70	0.00	35.1	0.70	0.573	0.403
1.10	-0.77	22	0.70	0.70	0.00	35.1	0.70	0.573	0.403
0.65	-0.46	23	0.70	0.70	0.00	35.1	0.70	0.573	0.403
0.30	-0.21	24	0.70	0.70	0.00	35.1	0.70	0.573	0.403
0.05	-0.04	25	0.70	0.70	0.00	35.1	0.70	0.573	0.403
-0.05	0.04	26	0.70	0.70	0.00	35.1	0.70	0.573	0.403
-0.30	0.21	27	0.70	0.70	0.00	35.1	0.70	0.573	0.403
-0.65	0.46	28	0.70	0.70	0.00	35.1	0.70	0.573	0.403
-1.10	0.77	29	0.70	0.70	0.00	35.1	0.70	0.573	0.403
-1.60	1.12	30	0.70	0.70	0.00	35.1	0.70	0.573	0.403
1.60	-1.12	31	0.70	0.00	0.70	45	0.70	0.495	0.495
1.10	-0.77	32	0.70	0.00	0.70	45	0.70	0.495	0.495
0.65	-0.46	33	0.70	0.00	0.70	45	0.70	0.495	0.495
0.30	-0.42	34	0.70	0.00	0.70	45	0.70	0.495	0.495
0.05	-0.04	35	0.70	0.00	0.70	45	0.70	0.495	0.495
-0.05	0.04	36	0.70	0.00	0.70	45	0.70	0.495	0.495
-0.30	0.21	37	0.70	0.00	0.70	45	0.70	0.495	0.495
-0.65	0.46	38	0.70	0.00	0.70	45	0.70	0.495	0.495
-1.10	0.77	39	0.70	0.00	0.70	45	0.70	0.495	0.495
-1.60	1.12	40	0.70	0.00	0.70	45	0.70	0.495	0.495

Table 5

Initial Item Parameters for Two Dimensional Simple and Complex Structures (MDISC=1.0)

D_i	d_i	Item	MDISC a_{i1}	a_{i2}		α_{i1}	MDISC a_{i1}	a_{i2}	
			(Simple Structure)				(Complex Structure)		
1.60	-1.60	1	1.00	1.00	0.00	0	1.00	1.00	0.00
1.10	-1.10	2	1.00	1.00	0.00	0	1.00	1.00	0.00
0.65	-0.65	3	1.00	1.00	0.00	0	1.00	1.00	0.00
0.30	-0.30	4	1.00	1.00	0.00	0	1.00	1.00	0.00
0.05	-0.05	5	1.00	1.00	0.00	0	1.00	1.00	0.00
-0.05	0.05	6	1.00	1.00	0.00	0	1.00	1.00	0.00
-0.30	0.30	7	1.00	1.00	0.00	0	1.00	1.00	0.00
-0.65	0.65	8	1.00	1.00	0.00	0	1.00	1.00	0.00
-1.10	1.10	9	1.00	1.00	0.00	0	1.00	1.00	0.00
-1.60	1.60	10	1.00	1.00	0.00	0	1.00	1.00	0.00
1.60	-1.60	11	1.00	1.00	0.00	24.3	1.00	0.911	0.412
1.10	-1.10	12	1.00	1.00	0.00	24.3	1.00	0.911	0.412
0.65	-0.65	13	1.00	1.00	0.00	24.3	1.00	0.911	0.412
0.30	-0.30	14	1.00	1.00	0.00	24.3	1.00	0.911	0.412
0.05	-0.05	15	1.00	1.00	0.00	24.3	1.00	0.911	0.412
-0.05	0.05	16	1.00	1.00	0.00	24.3	1.00	0.911	0.412
-0.30	0.30	17	1.00	1.00	0.00	24.3	1.00	0.911	0.412
-0.65	0.65	18	1.00	1.00	0.00	24.3	1.00	0.911	0.412
-1.10	1.10	19	1.00	1.00	0.00	24.3	1.00	0.911	0.412
-1.60	1.60	20	1.00	1.00	0.00	24.3	1.00	0.911	0.412
1.60	-1.60	21	1.00	1.00	0.00	35.1	1.00	0.818	0.575
1.10	-1.10	22	1.00	1.00	0.00	35.1	1.00	0.818	0.575
0.65	-0.65	23	1.00	1.00	0.00	35.1	1.00	0.818	0.575
0.30	-0.30	24	1.00	1.00	0.00	35.1	1.00	0.818	0.575
0.05	-0.05	25	1.00	1.00	0.00	35.1	1.00	0.818	0.575
-0.05	0.05	26	1.00	1.00	0.00	35.1	1.00	0.818	0.575
-0.30	0.30	27	1.00	1.00	0.00	35.1	1.00	0.818	0.575
-0.65	0.65	28	1.00	1.00	0.00	35.1	1.00	0.818	0.575
-1.10	1.10	29	1.00	1.00	0.00	35.1	1.00	0.818	0.575
-1.60	1.60	30	1.00	1.00	0.00	35.1	1.00	0.818	0.575
1.60	-1.60	31	1.00	0.00	1.00	45	1.00	0.707	0.707
1.10	-1.10	32	1.00	0.00	1.00	45	1.00	0.707	0.707
0.65	-0.65	33	1.00	0.00	1.00	45	1.00	0.707	0.707
0.30	-0.30	34	1.00	0.00	1.00	45	1.00	0.707	0.707
0.05	-0.05	35	1.00	0.00	1.00	45	1.00	0.707	0.707
-0.05	0.05	36	1.00	0.00	1.00	45	1.00	0.707	0.707
-0.30	0.30	37	1.00	0.00	1.00	45	1.00	0.707	0.707
-0.65	0.65	38	1.00	0.00	1.00	45	1.00	0.707	0.707
-1.10	1.10	39	1.00	0.00	1.00	45	1.00	0.707	0.707
-1.60	1.60	40	1.00	0.00	1.00	45	1.00	0.707	0.707

Table 6

Initial Item Parameters for Two Dimensional Simple and Complex Structures (MDISC=1.4)

D_i	d_i	Item	MDISC	a_{i1}	a_{i2}	α_{i1}	MDISC	a_{i1}	a_{i2}
			(Simple Structure)				(Complex Structure)		
1.60	-2.24	1	1.40	1.40	0.00	0	1.40	1.40	0.00
1.10	-1.54	2	1.40	1.40	0.00	0	1.40	1.40	0.00
0.65	-0.91	3	1.40	1.40	0.00	0	1.40	1.40	0.00
0.30	-0.42	4	1.40	1.40	0.00	0	1.40	1.40	0.00
0.05	-0.07	5	1.40	1.40	0.00	0	1.40	1.40	0.00
-0.05	0.07	6	1.40	1.40	0.00	0	1.40	1.40	0.00
-0.30	0.42	7	1.40	1.40	0.00	0	1.40	1.40	0.00
-0.65	0.91	8	1.40	1.40	0.00	0	1.40	1.40	0.00
-1.10	1.54	9	1.40	1.40	0.00	0	1.40	1.40	0.00
-1.60	2.24	10	1.40	1.40	0.00	0	1.40	1.40	0.00
1.60	-2.24	11	1.40	1.40	0.00	24.3	1.40	1.276	0.576
1.10	-1.54	12	1.40	1.40	0.00	24.3	1.40	1.276	0.576
0.65	-0.91	13	1.40	1.40	0.00	24.3	1.40	1.276	0.576
0.30	-0.42	14	1.40	1.40	0.00	24.3	1.40	1.276	0.576
0.05	-0.07	15	1.40	1.40	0.00	24.3	1.40	1.276	0.576
-0.05	0.07	16	1.40	1.40	0.00	24.3	1.40	1.276	0.576
-0.30	0.42	17	1.40	1.40	0.00	24.3	1.40	1.276	0.576
-0.65	0.91	18	1.40	1.40	0.00	24.3	1.40	1.276	0.576
-1.10	1.54	19	1.40	1.40	0.00	24.3	1.40	1.276	0.576
-1.60	2.24	20	1.40	1.40	0.00	24.3	1.40	1.276	0.576
1.60	-2.24	21	1.40	1.40	0.00	35.1	1.40	1.145	0.805
1.10	-1.54	22	1.40	1.40	0.00	35.1	1.40	1.145	0.805
0.65	-0.91	23	1.40	1.40	0.00	35.1	1.40	1.145	0.805
0.30	-0.42	24	1.40	1.40	0.00	35.1	1.40	1.145	0.805
0.05	-0.07	25	1.40	1.40	0.00	35.1	1.40	1.145	0.805
-0.05	0.07	26	1.40	1.40	0.00	35.1	1.40	1.145	0.805
-0.30	0.42	27	1.40	1.40	0.00	35.1	1.40	1.145	0.805
-0.65	0.91	28	1.40	1.40	0.00	35.1	1.40	1.145	0.805
-1.10	1.54	29	1.40	1.40	0.00	35.1	1.40	1.145	0.805
-1.60	2.24	30	1.40	1.40	0.00	35.1	1.40	1.145	0.805
1.60	-2.24	31	1.40	0.00	1.40	45	1.40	0.99	0.99
1.10	-1.54	32	1.40	0.00	1.40	45	1.40	0.99	0.99
0.65	-0.91	33	1.40	0.00	1.40	45	1.40	0.99	0.99
0.30	-0.42	34	1.40	0.00	1.40	45	1.40	0.99	0.99
0.05	-0.07	35	1.40	0.00	1.40	45	1.40	0.99	0.99
-0.05	0.07	36	1.40	0.00	1.40	45	1.40	0.99	0.99
-0.30	0.42	37	1.40	0.00	1.40	45	1.40	0.99	0.99
-0.65	0.91	38	1.40	0.00	1.40	45	1.40	0.99	0.99
-1.10	1.54	39	1.40	0.00	1.40	45	1.40	0.99	0.99
-1.60	2.24	40	1.40	0.00	1.40	45	1.40	0.99	0.99

Data Generation and Analysis

Data were simulated using MD4F (Roussos, 1995). This program allows the user to specify unidimensional and two-dimensional simple and complex test structures based on the two dimensional compensatory model (Reckase, 1985). The item responses are generated by comparing $P_i(\theta_j)$, given a θ vector for an examinee, with a uniform random number (K) generated between 0 and 1 using a random number generator. If $P_i(\theta_j)$ is greater than K, a score of 1 is assigned to an examinee for item i, otherwise, a 0 is assigned.

For the unidimensional study, each of the 1,800 unidimensional data sets were analyzed using the computer program NOHARM II (Fraser , 1988) modified by Gessaroli (1995). The Approximate χ^2 was computed based on the results of fitting McDonald's one factor polynomial approximation to a normal ogive model to the data sets. Type I error rates for the Approximate χ^2 were recorded.

Stout's T_1 and Stout's T_2 were obtained using the computer program DIMTEST (Stout, Douglas, Junker, & Roussos, 1993). Type I error rates were also recorded for the two indices. It is necessary to describe the procedure used to run DIMTEST. As mentioned previously, DIMTEST splits the total test into three subtests: AT1, AT2 (assessment subtests), and PT. The items of the subtests can be assigned by experts or by using an automatic selection mode. In this analysis, the automatic selection mode was used. According to Nandakumar and Stout (1993), assigning 1/4 of the total test items to AT1 and AT2 yielded reliable results for both Type I error rate and power. Therefore, one / fourth of the total test length was assigned to each assessment subtest; for L=40, 10 items were assigned to each of the two assessment subtests; for L=80, 20 out of 80 items were assigned to each of the two assessment subtests.

A certain number of examinees should be partialled out from the total sample for factor analysis for the purpose of selection of assessment AT1 and AT2 items while the remainder are used to compute Stout's T_1 and Stout's T_2. With N=500 and N=1,000, 1/3 was used for this purpose. With N=2,000, 500 examinees were used for factor analysis. Nandakumar and Stout (1993) used the same proportion of examinees for their analysis.

For the multidimensional study, similarly, the Approximate χ^2 was computed for each of the 14,400 data sets after fitting McDonald's one factor polynomial approximation to a normal ogive model to the data sets using the computer program NOHARM II (Fraser, 1988) modified by Gessaroli (1995). Stout's T_1 and Stout's T_2 statistics were obtained using DIMTEST. The power for each index under each combination of experimental conditions was computed. The same procedure for running DIMTEST was used as for the unidimensional study.

It seems to be a common practice for many researchers to use only simple descriptive statistics and tabular summaries to analyze the results from Monte Carlo studies. Harwell, Stone, Hsu, and Kirisci (1996) argued that this increases the chance that important effects will go undetected and that the magnitude of effects may be misestimated. Thus, the best solution is to use both descriptive and inferential analyses. A variety of inferential analyses can be performed. According to Harwell, Hsu, and Kirisci (1996), for IRT Monte Carlo studies, regression methods are more attractive and preferable, because most independent variables in IRT Monte Carlo studies are metric. Thus, in order to investigate the effects of the independent variables on rejection rates, logit-linear analyses were performed for the Approximate χ^2, Stout's T_1, and Stout's T_2 separately for each of the unidimensional and multidimensional conditions. This method has been used by researchers (Breithaupt,1995; Gessaroli & De Champlain, 1996).

Specifically, the simple and interactive effects of independent variables were assessed after the most parsimonious model was fit to the response frequencies. The dependent variable was the number of rejections of the null hypothesis. The logit-linear analysis was done in a forward hierarchical manner, i.e. starting with the simplest model and then fitting incrementally more complex models to the data until the best model was found. The principle is that higher-order effects are included in the model only if the corresponding lower order effects are also included in the model. The likelihood-ratio χ^2 was employed as the fit statistic. A model was considered to be acceptable if the corresponding p-value was equal to or greater than 0.15. Any effect of the individual variable was considered to be significant if the size of its absolute Z value was greater than or equal to 2.0. (This criterion was used by Gessaroli & De Champlain (1996)). Results are reported in Chapter IV.

Chapter IV

RESULTS

In this chapter, results are presented. First, empirical Type I error rates made by the Approximate χ^2, Stout's T_1, and Stout's T_2 under the combinations of test length (L), sample size (N), and discrimination (a) are tabulated and the results of logit-linear analysis of the Type I error rates are reported. Next, the power of each statistic under the combinations of test length, sample size, test structure, discrimination, and correlation (r) are displayed and the results of logit-linear analysis of the power of the three indices are presented.

Unidimensional Data Sets

The Approximate χ^2

The Type I error rates of the Approximate χ^2 based on 100 replications under each combination of the conditions are presented in Table 7. As shown in Table 7, the Approximate χ^2 had excellent Type I error control with no incorrect rejections across all conditions simulated. Since there was no variability in the Type I error rate for the Approximate χ^2 across all conditions, logit-linear analysis was not performed with this statistic.

Stout's T_1 and Stout T_2

In Table 7, the Type I error rates of Stout's T_1 and Stout's T_2 are also displayed. The Type I error rates, although higher than the Approximate χ^2, were below the nominal level with a few exceptions. At the cell of L=80, a=1.0, and N=500, Stout's T_1 had Type I error of .07 and Stout's T_2 .12. Stout's T_2 also had Type I error of .07 under the condition of L=80, a=1.4, and N=500 (see the highlighted numbers in Table 7). Overall, greater Type I error rates were found for

Stout's T_2 than Stout's T_1. On average, Stout's T_2 had .01 higher Type I error rate than Stout's T_1.

Table 7
Type I Error Rates for the Approximate χ^2, Stout's T_1, and Stout's T_2 Based on 100 Unidimensional Data Sets

		Approximate χ^2		Stout's T_1		Stout's T_2	
		L=40	L=80	L=40	L=80	L=40	L=80
a=0.7	N=500	0.00	0.00	0.01	0.04	0.01	0.05
	N=1,000	0.00	0.00	0.01	0.02	0.01	0.03
	N=2,000	0.00	0.00	0.01	0.01	0.01	0.02
a=1.0	N=500	0.00	0.00	0.02	**0.07**	0.04	**0.12**
	N=1,000	0.00	0.00	0.00	0.02	0.00	0.04
	N=2,000	0.00	0.00	0.00	0.00	0.00	0.00
a=1.4	N=500	0.00	0.00	0.00	0.03	0.02	**0.07**
	N=1,000	0.00	0.00	0.00	0.01	0.00	0.01
	N=2,000	0.00	0.00	0.00	0.00	0.00	0.00

Logit-linear analysis was carried out with Stout's T_1 and Stout's T_2 to assess the effects of discrimination, sample size, and test length on the Type I error rates. Based on the analysis results, discrimination did not significantly affect the performance for Stout's T_1. A model with only the main effects of sample size and test length fit the data well. The model fit statistics had a $\chi^2(14)$ value of 10.99 and P value of .69. For sample size, only the parameter estimate between N=500 and N=1,000 was significant, $|Z| \geq 2.0$. The parameter estimate for test length was also shown to be significant, $|Z| \geq 2.0$. For Stout's T_2, a similar model was obtained. The model fit statistics had a $\chi^2(14)$ value of 15.5 and P value of .35. Similarly, the parameter estimate between N=500 and N=1,000 was significant, $|Z| \geq 2.0$. For test length, the parameter estimate was also significant with $|Z| \geq 2.0$. The effects of sample size and test length on the Type I error rates of the

two indices are shown in Table 8.

As shown in Table 8, for both Stout's T_1 and Stout's T_2, higher Type I error rates were related to smaller sample sizes and longer test length. A stronger sample size effect was observed when sample size dropped from 1,000 to 500 with Stout's T_2 being more pronounced. In the logit-linear analysis, the parameter estimate for this effect was significant. However, the parameter estimate for the effect between N=1,000 and N=2,000 was not significant. Test length had a slightly stronger effect for Stout's T_2 than for Stout's T_1.

Table 8

The Effects of Sample Size and Test Length on the Type I Error Rates of Stout's T_1 and Stout's T_2 Based on 100 Unidimensional Data Sets

	Stout's T_1	**Stout's T_2**		**Stout's T_1**	**Stout's T_2**
N=500	.028	.052	L=40	.006	.010
N=1,000	.010	.015	L=80	.022	.038
N=2,000	.003	.005			

Multidimensional Data Sets

In the multidimensional case, the use of two different dimensional test structures (simple and complex) resulted in very different performance for the Approximate χ^2, Stout's T_1, and Stout's T_2. Thus, logit-linear analysis was performed separately with each test structure for each index. In the following discussion, power ≥ .80 is considered as good power (see italics in Tables below) and power ≥ .95 is classified as excellent power (highlighted in the following tables).

Two Dimensional Simple Test Structure

As shown in Table 9, it seems that the Approximate χ^2, Stout's T_1, and Stout's T_2 had equally strong power in most cases except that the Approximate χ^2 was more strongly affected at

Table 9

The Power of the Approximate χ^2, Stout's T_1, and Stout's T_2 with Two Dimensional Simple Test Structure Over 100 Replications

MDISC	Sample Size	Correlation	The Approximate χ^2 Test Length		Stout's T_1 Test Length		Stout's T_2 Test Length	
			40	80	40	80	40	80
0.7	500	.0	1.00	1.00	1.00	1.00	1.00	1.00
		.4	1.00	1.00	1.00	1.00	1.00	1.00
		.57	.83	.94	.89	1.00	.91	1.00
		.7	.05	.03	.58	.99	.62	.99
	1,000	.0	1.00	1.00	1.00	1.00	1.00	1.00
		.4	1.00	1.00	1.00	1.00	1.00	1.00
		.57	1.00	1.00	1.00	1.00	1.00	1.00
		.7	.87	.99	.98	1.00	.98	1.00
	2,000	.0	1.00	1.00	1.00	1.00	1.00	1.00
		.4	1.00	1.00	1.00	1.00	1.00	1.00
		.57	1.00	1.00	1.00	1.00	1.00	1.00
		.7	1.00	1.00	1.00	1.00	1.00	1.00
1.0	500	.0	1.00	1.00	1.00	1.00	1.00	1.00
		.4	1.00	1.00	1.00	1.00	1.00	1.00
		.57	1.00	1.00	.99	1.00	.99	1.00
		.7	.51	.62	.92	1.00	1.00	1.00
	1,000	.0	1.00	1.00	1.00	1.00	1.00	1.00
		.4	1.00	1.00	1.00	1.00	1.00	1.00
		.57	1.00	1.00	1.00	1.00	1.00	1.00
		.7	1.00	1.00	1.00	1.00	1.00	1.00
	2,000	.0	1.00	1.00	1.00	1.00	1.00	1.00
		.4	1.00	1.00	1.00	1.00	1.00	1.00
		.57	1.00	1.00	1.00	1.00	1.00	1.00
		.7	1.00	1.00	1.00	1.00	1.00	1.00
1.4	500	.0	1.00	1.00	1.00	1.00	1.00	1.00
		.4	1.00	1.00	1.00	1.00	1.00	1.00
		.57	1.00	1.00	.99	1.00	1.00	1.00
		.7	.99	1.00	1.00	1.00	1.00	1.00
	1,000	.0	1.00	1.00	1.00	1.00	1.00	1.00
		.4	1.00	1.00	1.00	1.00	1.00	1.00
		.57	1.00	1.00	1.00	1.00	1.00	1.00
		.7	1.00	1.00	1.00	1.00	1.00	1.00
	2,000	.0	1.00	1.00	1.00	100	1.00	1.00
		.4	1.00	1.00	1.00	100	1.00	1.00
		.57	1.00	1.00	1.00	100	1.00	1.00
		.7	1.00	1.00	1.00	100	1.00	1.00

N=500 and r=.7 for a=.7. Under these conditions, power was much lower (.05 for L=40 and .03 for L=80) than Stout's T_1 and Stout's T_2. Also, for N=500, r=.7 and a=1.0, the power for the Approximate χ^2 was reduced (.51 for L=40 and .62 for L=80). Overall mean power was .97 for the Approximate χ^2 and .99 for Stout's T_1 and Stout's T_2. In the following section, logit-linear analyses of the effects of the independent variables on the power of the Approximate χ^2, Stout's T_1, and Stout's T_2 are reported.

The Approximate χ^2 Logit-linear analysis was carried out to test the effects of test length, sample size, item discrimination, and correlation on the power of the Approximate χ^2. Based on the analysis, test length did not significantly affect the power. Thus, a model with a 2-way interaction of sample size by correlation and a 2-way interaction of sample size by discrimination fit the frequencies adequately. The fit statistic $\chi^2(18)$ was 10.87 and P=.90. However, not all the parameter estimates were significant. For the sample size by correlation interaction, only the parameter estimate between N=500 and N=1,000 for r=.57 and r=.7 was shown significant, $|Z| \geq 2.0$. This interaction appears to result from a ceiling effect at N=1,000 for r=.57. For the sample size by discrimination interaction, the parameter estimate between N=500 and N=1,000 for a=.7 and a=1.0 and the parameter estimate between N=500 and N=1,000 for a=1.0 and a=1.4 were significant. Again the interactions appear to result from ceiling effects at N=1,000 for a=1.4.

Table 10

The Effects of Sample Size by Correlation and Sample Size by Discrimination on the Power of the Approximate χ^2 Based on Two Dimensional Simple Test Structure Data Sets

	Power		
	N=500	N=1,000	N=2,000
r=.0	**1.00**	**1.00**	**1.00**
r=.4	**1.00**	**1.00**	**1.00**
r=.57	**.96**	**1.00**	**1.00**
r=.7	.53	**.97**	**.97**
a=0.7	.73	**.98**	**.98**
a=1.0	*.89*	**1.00**	**1.00**
a=1.4	**.99**	**1.00**	**1.00**

The main effects of sample size, test length, discrimination, and correlation on the power of the Approximate χ^2 are shown in Table 11. Excellent power was observed for N≥1,000, a ≥ 1.0, r ≤.57, and both test length. This contributed to the ceiling effects in the 2-way interactions for the Approximate χ^2.

Table 11

The Main Effects of Independent Variables on the Power of the Approximate χ^2 Based on Two Dimensional Simple Test Structure Data Sets

			Power				
a=0.7	*.910*	L=40	**.951**	N=500	*.874*	r=.0	**1.000**
a=1.0	**.964**	L=80	**.961**	N=1,000	**.994**	r=.4	**1.000**
a=1.4	**1.000**			N=2,000	**1.000**	r=.57	**.987**
						r=.7	*.840*

Stout's T_1 and Stout's T_2 Since similar performance patterns were observed for Stout's T_1 and Stout's T_2, logit-linear analysis results for the two indices are reported at the same time. The

analysis indicated that for both Stout's T_1 and Stout's T_2, a model with only main effects of independent variables was sufficient to describe the rejection frequencies. The model fit resulted in a $\chi^2(63)=23.57$ and P=1.0 for Stout's T_1 and a $\chi^2(63)=26.41$ and P=1.0 for Stout's T_2. Based on the logit-linear analysis of Stout's T_1, the parameter estimates for the following effects were significant: the effect between a=.7 and a=1.0; the effect between N=500 and N=1,000; the effect of test length; and the effect between r=.57 and r=.7. Similar parameter estimation results were shown for Stout's T_2.

The power of Stout's T_1 and Stout's T_2 for each independent variable is presented in Table 12. Both indices yielded very strong power in each case. Although there was a slight decrease in power for a=.7, and r=.7 for N=500, overall the power was still greater than .97 for each index. As shown in Table 12, similar power was observed for both indices and it appears that there was no meaningful difference between Stout's T_1 and Stout's T_2.

Table 12
The Main Effects of Independent Variables on the Power of Stout's T_1 and Stout's T_2 Based on Two Dimensional Simple Test Structure Data Sets

	Power			Power	
	Stout's T_1	**Stout's T_2**		**Stout's T_1**	**Stout's T_2**
a=0.7	**.977**	**.979**	N=500	**.986**	**.990**
a=1.0	**.996**	**1.000**	N=1,000	**1.000**	**1.000**
a=1.4	**1.000**	**1.000**	N=2,000	**1.000**	**1.000**
L=40	**.982**	**.986**	r=.0	**1.000**	**1.000**
L=80	**1.000**	**1.000**	r=.4	**1.000**	**1.000**
			r=.57	**.993**	**.994**
			r=.7	**.971**	**.977**

Two Dimensional Complex Test Structure

Based on the results of the analyses with two dimensional complex test structure, all three indices were sensitive to test structure. The results for the Approximate χ^2 followed by the results for Stout's T_1 and Stout's T_2 are presented.

The Approximate χ^2 The power of the Approximate χ^2 with two dimensional complex test structure is shown in Table 13. For a=.7 with both test lengths, excellent power of .99 and 1.0 was observed only for r=.0 and N=2,000. For the remaining cells, the maximum power was only .13. For a=1.0, dramatic improvement occurred for the cells of r=.0 and N=1,000 with a minimum power of .77 for L=40 and .88 for L=80. For N=2,000, power was 1.0 for both test lengths. With these exceptions, power was extremely poor and no more than .30. For a=1.4,

Table 13

The Power of the Approximate χ^2 with Two Dimensional Complex Test Structure Over 100 Replications

		N=500		N=1,000		N=2,000	
MDISC	Correlation	L=40	L=80	L=40	L=80	L=40	L=80
0.7	.0	.00	.00	.13	.06	**.99**	**1.00**
	.4	.00	.00	.00	.00	.00	.00
	.57	.00	.00	.00	.00	.00	.00
	.7	.00	.00	.00	.00	.00	.00
1.0	.0	.00	.00	.77	*.88*	**1.00**	**1.00**
	.4	.00	.00	.00	.00	.21	.30
	.57	.00	.00	.00	.00	.00	.00
	.7	.00	.00	.00	.00	.00	.00
1.4	.0	.14	.14	**1.00**	**1.00**	**1.00**	**1.00**
	.4	.00	.00	.06	.02	**1.00**	**1.00**
	.57	.00	.00	.00	.00	**.99**	**1.00**
	.7	.00	.00	.00	.00	*.87*	**.98**

dramatic improvement occurred for N=1,000 and r=.0. Perfect power was observed for r=.0 and N=1,000 over the two test lengths. For N=2,000, the Approximate χ^2 performed extremely well with minimum power of .87 for r=.7 and L=40 and minimum power of .98 for all other conditions. It was interesting to note that other than for N=2,000 and a=1.4 power was not satisfactory for r ≥.4. In fact, The Approximate χ^2 performed extremely poorly and completely failed to identify the correct number of dimensions in most cases. There were only 15 cases out of 72 conditions with power over .80. The average power was .23.

Logit-linear analysis was performed with the power of the Approximate χ^2. A model with only main effects fit the data well. The statistics for the model fit were $\chi^2(63)=74.72$ and P=.15. The parameter estimates for all the main effects were significant with $|Z|\geq 2.0$ except for test length. The model with test length effect excluded was tried, but the model fit statistics did not meet the model fit criterion (P≥.15). Therefore, this effect was kept in the model. The main effects of the independent variables are displayed in Table 14.

Table 14

The Main Effects of Independent Variables on the Power of the Approximate χ^2 Based on Two Dimensional Complex Test Structure Data Sets

			Power				
a=0.7	.091	N=500	.012	L=40	.227	r=.0	.562
a=1.0	.173	N=1,000	.163	L=80	.233	r=.4	.144
a=1.4	.425	N=2,000	.514			r=.57	.111
						r=.7	.103

As shown in Table 14, increasing discrimination and sample size increased power.

Greater increase was observed for a=1.4 and N=2,000 than the other levels of discrimination and sample size. Yet, the increase was far from satisfactory. The maximum power was only .514. Increasing correlation was related to decrease in power. A drastic drop in power for r=.4 was observed. Beyond this level, similarly poor power was shown. Similar power was shown for both test lengths. Overall, the Approximate χ^2 performed extremely poorly with the complex test structure except for a=1.4 and N=2,000, where power ranged from .87 to 1.0 (see Table 14).

Stout's T_1 and Stout's T_2 Much better power was shown for Stout's T_1 and Stout's T_2 than for the Approximate χ^2 (see Table 15). However, for a majority of the cases the power of these indices was not satisfactory. For each index, 51 of 72 times the power was below .80. Each of the independent variables had strong effects on the power of the two indices. Generally speaking, power increased as test length, discrimination, and sample size increased and dimension correlation decreased. For example, for both indices, with L=40, 6 of 36 times the power of each index was above .80 while for L=80, 15 of 36 times, the power was above .80. With discrimination, out of 24 times, 4 times for a=.7, 8 times for a=1.0, and 9 times for a=1.4, the power was above .80. With sample size, out of 24 times, power was above .80 4 times for N=500, 7 times for N=1,000, and 10 times for N=2,000. When correlation was examined, of 18 cases, power was satisfactory (above .80) for 14 cases for r=.0, 6 cases for r=.4, one case for r=.57, and none for r=.7. In the following section, the logit-linear analyses of the power of Stout's T_1 and Stout's T_2 are reported in detail.

Table 15

The Power of Stout's T_1 and Stout's T_2 with Two Dimensional Complex Test Structure Over 100 Replications

Stout's T_1

		N=500		N=1,000		N=2,000	
MDISC	Correlation	L=40	L=80	L=40	L=80	L=40	L=80
	.0	.09	.62	.64	**.99**	**.95**	**1.00**
	.4	.01	.13	.10	.72	.33	**1.00**
0.7	.57	.00	.05	.03	.19	.02	.60
	.7	.01	.05	.00	.01	.00	.04
	.0	.46	*.92*	**.96**	**1.00**	**1.00**	**1.00**
	.4	.06	.46	.28	**.95**	.53	**1.00**
1.0	.57	.00	.13	.08	.42	.13	*.88*
	.7	.00	.06	.00	.09	.00	.28
	.0	*.88*	**.99**	**1.00**	**1.00**	**1.00**	**1.00**
	.4	.38	*.88*	.55	**.99**	.72	**.99**
1.4	.57	.09	.33	.20	.62	.21	.50
	.7	.01	.06	.02	.03	.01	.06

Stout's T_2

		N=500		N=1,000		N=2,000	
MDISC	Correlation	L=40	L=80	L=40	L=80	L=40	L=80
	.0	.16	.67	.66	**.99**	**.96**	**1.00**
	.4	.04	.19	.13	.76	.38	**1.00**
0.7	.57	.04	.08	.04	.26	.04	.64
	.7	.02	.11	.01	.01	.01	.05
	.0	.56	*.93*	**.96**	**1.00**	**1.00**	**1.00**
	.4	.11	.53	.33	**.95**	.55	**1.00**
1.0	.57	.00	.15	.11	.53	.17	*.89*
	.7	.02	.09	.01	.12	.02	.29
	.0	*.91*	**.99**	**1.00**	**1.00**	**1.00**	**1.00**
	.4	.45	*.88*	.59	**1.00**	.77	**.99**
1.4	.57	.15	.43	.27	.64	.29	.53
	.7	.03	.15	.06	.06	.02	.06

Logit-linear analysis indicated that the model fit included the highest order interaction (4-

way interaction). To simplify the analysis, logit-linear analysis was performed for each sample size.

For N=500, for Stout's T_1 and Stout's T_2 the simplest model that adequately explained the frequencies included the following effects of the independent variables: test length, discrimination, correlation, and discrimination by correlation interaction. The model fit statistics had a $\chi^2(11)$ value of 9.86 and P=.54 for Stout's T_1 and a $\chi^2(11)$ value of 13.27 and P=.28 for Stout's T_2. All the main effects had $|Z| \geq 2.0$. Power increased as test length and discrimination increased and correlation decreased.

Discrimination by correlation interaction effects for Stout's T_1 and Stout's T_2 are displayed in Table 16. For Stout's T_1, based on the logit-linear analysis results, out of the 6 parameter estimates, 4 estimates were significant. All the interaction effects were ordinal. The

Table 16

The Effects of Discrimination by Correlation on the Power of Stout's T_1 and Stout's T_2 for N=500 with Two Dimensional Complex Test Structure Data Sets

	Power					
	Stout's T_1			**Stout's T_2**		
	a=.7	a=1.0	a=1.4	a=.7	a=1.0	a=1.4
r=.0	.355	.690	*.935*	.415	.745	**.950**
r=.4	.070	.260	.630	.115	.320	.665
r=.57	.025	.065	.210	.060	.075	.290
r=.7	.030	.030	.035	.065	.055	.090

first significant interaction effect occurred between r=.0 and r=.4 for a=.7 and a=.10. That is, when discrimination moved from .7 to 1.0, power increased by .335 for r=.0 and .19 for r=.4. The

second significant discrimination by correlation effect was observed between r=.4 and r=.57 for a=.7 and a=1.0. As discrimination went from .7 to 1.0, there was a substantially greater increase in power for r=.4 than for r=.57 (.19 compared to .04). The third significant discrimination by correlation effect existed between r=.0 and r=.4 for a=1.0 and a=1.4. A greater increase in power for r=.4 than for r=.0 was observed as discrimination increased from 1.0 to 1.4. The increase for r=.4 was .37 while for r=.0 was .245. The fourth significant discrimination by correlation interaction effect was due to the fact that there was .225 greater increase in power for r=.4 than r=.57 as discrimination increased from 1.0 to 1.4 (see Table 16).

For Stout's T_2, similarly, all interaction effects were ordinal. Based on the parameter estimates results, except for the effect between r=.57 and r=.7 for a=.7 and a=1.0, all the effects were significant. The 5 significant effects contributed to overall discrimination by correlation interaction. For example, the first significant effect was due to the fact that there was a much greater increase in power for r=.0 than r=.4 (.330 compared to .205) when discrimination increased from .7 to 1.0. The second significant discrimination effect existed between r=.4 and r=.57 for a=.7 and a=1.0. The increase for r=.4 was .205 compared to the increase of .015 for r=.57 as discrimination moved from .7 to 1.0. The third significant discrimination by correlation interaction effect was attributable to the fact that a greater increase in power (.345) occurred for r=.4 than for r=.0 (.205) as discrimination increased from 1.0 to 1.4. The effect between r=.4 and r=.57 for a=1.0 and a=1.4 was also substantial. As discrimination went form 1.0 to 1.4, a greater increase in power was observed for r=.4 than for r=.57 (.345 compared to .215). The last significant discrimination by correlation interaction effect appears to be attributable to the fact

that there was a substantially greater increase in power at r=.57 than at r=.7 as discrimination increased from 1.0 to 1.4. The increase was .18 greater for r=.57 than for r=.7 (.215 compared to .035) (see Table 16). Very similar interaction patterns were shown for both indices. There was not much difference in power between the two indices. On average, .03 greater power was observed for Stout's T_2 than Stout's T_1.

For N=1,000, based on the logit-linear analysis results, for Stout's T_1, the best model that fit the data included the following effects: the main effects of discrimination, test length, correlation, and 2-way interactions of test length by correlation and discrimination by correlation. The test of model fit had a $\chi^2(8) = 7.29$ and P=.51. All the main effects were significant. Power increased as test length and discrimination increased and correlation decreased.

For Stout's T_2 with N=1,000, a model with all the main effects of test length, discrimination, correlation, and three 2-way interactions of discrimination by test length, discrimination by correlation, and test length by correlation fit the data well. The model fit statistics were $\chi^2(6) = 9.55$ and P=.15. Although the model fit statistics indicated that there were three 2-way interactions, none of the individual parameter estimates for discrimination by test length interaction was significant. However, when this interaction was excluded from the model, the fit statistics did not meet the criterion, P≥.15. Therefore, this interaction was kept within the model. For the main effects, all the effects for discrimination, test length, and correlation were shown significant. The increase in power was related to an increase in test length and discrimination and a decrease in correlation.

The test length by correlation interaction effects for Stout's T_1 and Stout's T_2 are presented in Table 17. For Stout's T_1, all parameter estimates for test length by correlation interaction were significant, $|Z| \geq 2.0$. All interaction effects were ordinal (see Table 17). The significant effects were attributable to a substantially different

Table 17

The Effects of Test Length by Correlation on the Power of Stout's T_1 and Stout's T_2 for N=1,000 with Two Dimensional Complex Test Structure Data Sets

	Power			
	Stout's T_1		**Stout's T_2**	
	L=40	L=80	L=40	L=80
r=.0	*.867*	**.997**	*.873*	**.997**
r=.4	.310	*.887*	.350	*.903*
r=.57	.103	.410	.140	.477
r=.7	.007	.043	.027	.063

increase for each level of correlation as test length increased from 40 items to 80 items. For example, when test length went from 40 items to 80 items, power increased .13 for r=.0 and .557 for r=.4. This was caused by the fact that a ceiling effect occurred for r=.0 for L=80. Another significant effect was observed between r=.4 and r=.57 for L=40 and L=80. Greater increase in power was found for r=.4 than for r=.57 (.577 increase for r=.4 compared to .307 increase for r=.57). The last significant test length by correlation effect was explained by the fact that there was a much greater increase in power for r=.57 (.307) compared to the increase for r=.7 (.036) when L=80 was used. This appears to be a floor effect.

Similarly, for Stout's T_2, parameter estimates for the test length by correlation were

significant, $|Z| \geq 2.0$. For r=.0, power increased from .873 to .997 while for r=.4 power increased from .35 to .903. Similar to Stout's T_1, this appears to be caused by a ceiling effect for r=.0 at L=80. For r=.4, power increased from .35 to .903 compared to the power increase from .14 to .447 for r=.57. When the effect between r=.57 and r=.7 was examined, power increased by .337 for r=.57 compared to a negligible increase (.036) for r=.7, a floor effect. Other than for r=.0 and L=80, a slightly higher power was found for Stout's T_2 than Stout's T_1. The mean power for Stout's T_1 was .453 while for Stout's T_2 was .479 (see Table 17).

The effects for the discrimination by correlation interaction for Stout's T_1 and Stout's T_2 are displayed in Table 18. For Stout's T_1, out of 6 interaction effects, two parameter estimates were significant, $|Z| \geq 2.0$. For Stout's T_1, all the interaction effects were ordinal. It appears the two significant effects were between r=.0 and r=.4 for a=1.0 and a=1.4 and r=.57 and r=.7 for a=1.0 and a=1.4. For example, when discrimination moved from 1.0 to 1.4, power increased by .02 for r=.0 and .155 for r=.4. Again, this was caused by a ceiling effect at r=.0 for a=1.4 (see Table 18). The other significant effect occurred between r=.57 and r=.7 for a=1.0 and a=1.4.

Table 18

The Effects of Discrimination by Correlation on the Power of Stout's T_1 and Stout's T_2 for N=1,000 with Two Dimensional Complex Test Structure Data Sets

	Power					
	Stout's T_1			**Stout's T_2**		
	a=.7	a=1.0	a=1.4	a=.7	a=1.0	a=1.4
r=.0	*.815*	**.980**	**1.000**	*.825*	**.980**	**1.000**
r=.4	.410	.615	.770	.445	.640	.795
r=.57	.110	.250	.410	.150	.320	.455
r=.7	.005	.045	.025	.010	.055	.060

where power increased by .16 for r=.57 compared to a decrease of .02 for r=.7. This also appears to be a floor effect.

For Stout's T_2, based on the logit-linear analysis results, only the parameter estimate between r=.0 and r=.4 for a=1.0 and a=1.4 was significant. As for Stout's T_1, a ceiling effect was found. At r=.0, power increased by .02, .135 less than the increase (.155) for r=.4. In addition to this significant effect, there were two non-negligible effects. The effect between r=.57 and r=.7 for a=.7 and a=1.0 was fairly substantial. There was a power increase of .17 for r=.57 compared to the increase of .045 for r=.7 as discrimination increased from .7 to 1.0. Another substantial effect was observed between r=.57 and r=.7 for a=1.0 and a=1.4. Power increased by .135 for r=.57 compared to an increase of .005 for r=.7. These two effects may also contribute to the overall discrimination by correlation interaction (see Table 18). In fact, the interaction effects had very similar patterns for both indices. Fewer significant interaction effects were observed for Stout's T_2 than for Stout's T_1 but still contributed to overall effect. This was simply due to the slightly greater power for Stout's T_2.

The effects of the test length by discrimination for Stout's T_2 (This interaction for Stout's T_1 was not included in the model) are shown in Table 19 and Figure 1. As reported above, according to the logit-linear analysis results, none of the interaction effects for Stout's T_2 were significant. However, excluding this interaction from the model resulted in a worse model fit, $P<.15$. As shown in Table 19, there was .102 less increase in power for a=1.4 than other levels of discrimination as test length increased from 40 items to 80 items. When a=.7 and 1.0 were examined, approximately the same amount of increase in power was shown (.295 for a=.7

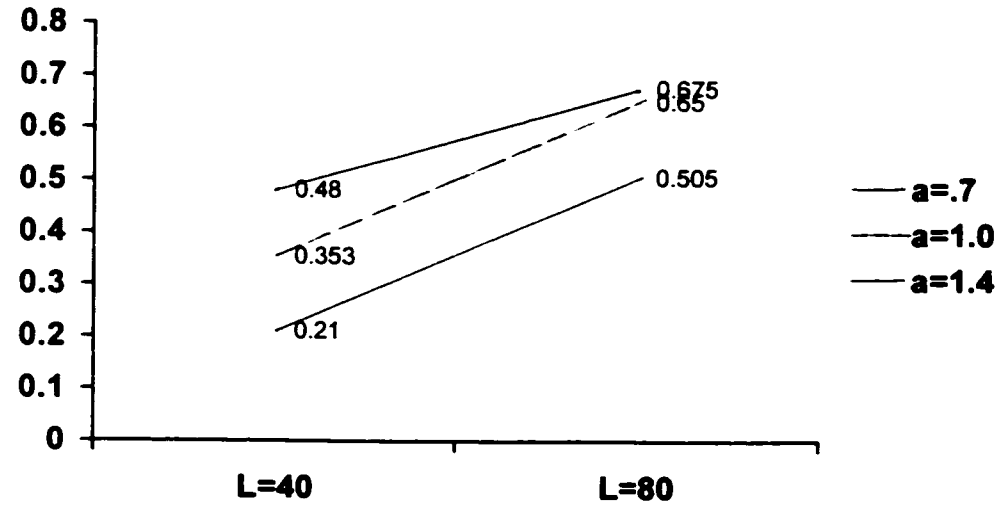
and .297 for a=1.0) as test length increased (see Table 19). The difference in power increase for a=1.4 from the increase for the other levels of discrimination across test lengths likely accounts for the test length by discrimination interaction (see Figure 1).

Table 19

The Effects of Discrimination by Test Length on the Power of Stout's T_2 for N=1,000 with Two Dimensional Complex Test Structure Data Sets

	Power			
	Stout's T_1		**Stout's T_2**	
	L=40	L=80	L=40	L=80
a=0.7	N/A	N/A	.210	.505
a=1.0	N/A	N/A	.353	.650
a=1.4	N/A	N/A	.480	.675

Figure 1 The Effects for Test Length by Discrimination for Stout's T_2 for N=1,000



For N=2,000, the best model for Stout's T_1 included the main effects of test length, discrimination, correlation, and all 2-way interactions. The model fit statistics were $\chi^2(6)=1.06$

and P=.28. For discrimination, only the parameter estimates for the first two levels of discrimination were significant. For the first two levels of discrimination, power increased as discrimination increased. As discrimination increased from 1.0 to 1.4, power decreased. For test length and correlation, all the effects were significant. Power increased as test length increased and correlation decreased.

For Stout's T_2, the best model included the main effects of test length, discrimination, correlation, and all 2-way interactions. The model fit statistics were $\chi^2(6)=5.02$ and p=.54. For the main effects, all the parameter estimates for test length, discrimination, and correlation were significant. Except for a=1.4, increasing test length and discrimination increased power and increasing correlation decreased power. When discrimination increased to 1.4, power decreased.

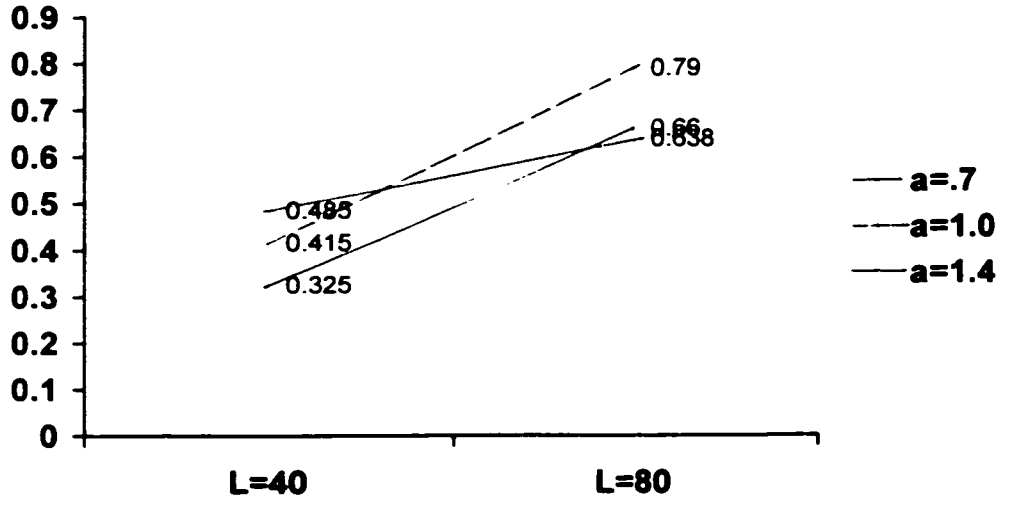
The effects of discrimination by test length interaction for Stout's T_1 and Stout's T_2 are presented in Table 20. For Stout's T_1, based on the parameter estimates results, both interaction effects were significant. It appears that significant interaction effects occurred between a=.7 and a=1.4 for L=40 and L=80 and between a=1.0 and a=1.4 for L=40 and L=80 (see Table 20 and Figure 2). As shown in Figure 2, the interaction effects were disordinal for a=.7 and a=1.4 and a=1.0 and a=1.4 across the two test lengths. This would account for each of the two significant contrasts. No significant test length by discrimination effects were found between a=.7 and a=1.0 for L=40 and L=80. The same amount of increase in power was observed for the two levels of discrimination across the two test lengths (see Figure 2).

Table 20

The Effects of Discrimination by Test Length on the Power of Stout's T_1 and Stout's T_2 for N=2,000 with Two Dimensional Complex Test Structure Data Sets

	Power			
	Stout's T_1		**Stout's T_2**	
	L=40	L=80	L=40	L=80
a=0.7	.325	.660	.348	.673
a=1.0	.415	.790	.435	.795
a=1.4	.485	.638	.520	.645

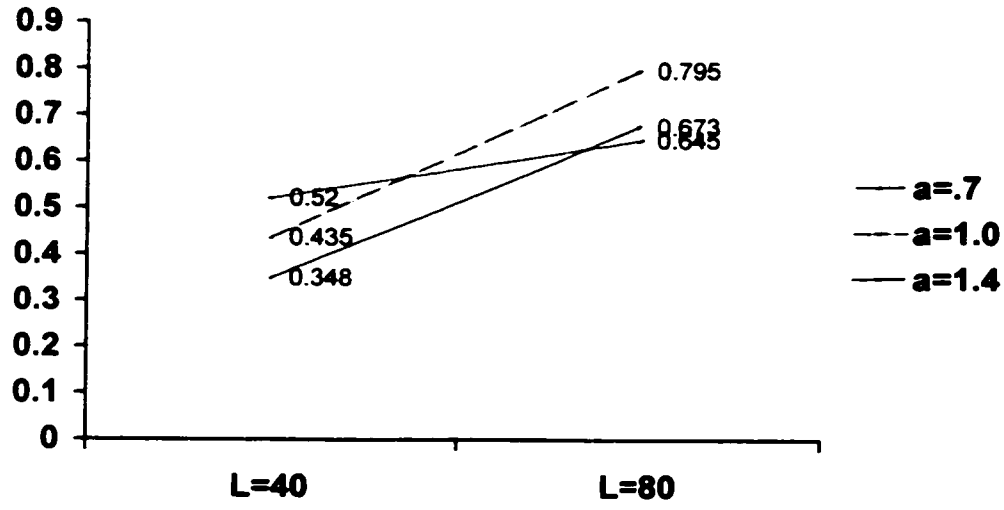
Figure 2 The Effects for Test Length by Discrimination for Stout's T_1 for N=2,000



For Stout's T_2, in logit-linear analysis, all the parameter estimates for discrimination by test length interaction were significant. The significant effects existed between a=.7 and a=1.4 and between a=1.0 and a=1.4 for L=40 and L=80. As shown in Figure 3, these two effects were disordinal with the line for a=1.4 crossing the lines for a=.7 and a=1.0. For a=.7, power increased by .325 compared to a much smaller increase of .125 for a=1.4. For a=1.0, power increased by

.36 compared to an increase of .125 for a=1.4. A much greater increase for a=.7 and a=1.0 than the increase for a=1.4 caused the disordinal interaction effects. Both Stout's T_1 and Stout's T_2 performed similarly.

Figure 3 The Effects for Test Length by Discrimination for Stout's T_2 for N=2,000



The effects of test length by correlation interaction for Stout's T_1 and Stout's T_2 are shown in Table 21. Parameter estimates in logit-linear analysis indicated that for Stout's T_1, there were two significant effects. It appears that one of the significant effects was attributable to the fact that there was a substantially greater increase in power (.47) for r=.4 compared to a much less increase in power for r=.0 when test length increased from 40 items to 80 items (see Table 21). The power increase differed by .453 between r=.0 and r=.4. For r=.0, excellent power of .983 and perfect power were achieved for L=40 and L=80 respectively, resulting in a ceiling effect. Another significant test length by correlation effect was due to a much greater increase in

Table 21

The Effects of Test Length by Correlation on the Power of Stout's T_1 and Stout's T_2 for N=2,000 with Two Dimensional Complex Test Structure Data Sets

	Power			
	Stout's T_1		**Stout's T_2**	
	L=40	L=80	L=40	L=80
r=.0	**.983**	**1.000**	**.987**	**1.000**
r=.4	.527	**.997**	.567	**.997**
r=.57	.120	.660	.167	.678
r=.7	.030	.127	.017	.133

power for r=.57 than for r=.7 as test length increased. The increase for r=.57 was .54 while for r=.7 the increase was only .097. This seems to be a floor effect.

For Stout's T_2, similarly, based on the parameter estimates in logit-analysis, there were two significant test length by correlation interaction effects. By examining Table 21, it was found that the two significant effects occurred between r=.0 and r=.4 for L=40 and L=80 and between r=.57 and r=.7 for L=40 and L=80. When test length increased, for r=.0, excellent and perfect power (.987 for L=40 and 1.0 for L=80) were achieved resulting in a ceiling effect while for r=.4, power was only .567 for L=40 but increased to .997 for L=80. For r=.57, power increased by .511 compared to a much less increase of .116 for r=.7 (a floor effect) as test length increased. Again, the interaction effects for both indices followed a similar pattern. However, Stout's T_2 was slightly more powerful than Stout's T_1 with an average power of .568 compared to an average power of .556 for Stout's T_1.

The effects of discrimination by correlation interaction for Stout's T_1 and Stout's T_2 are

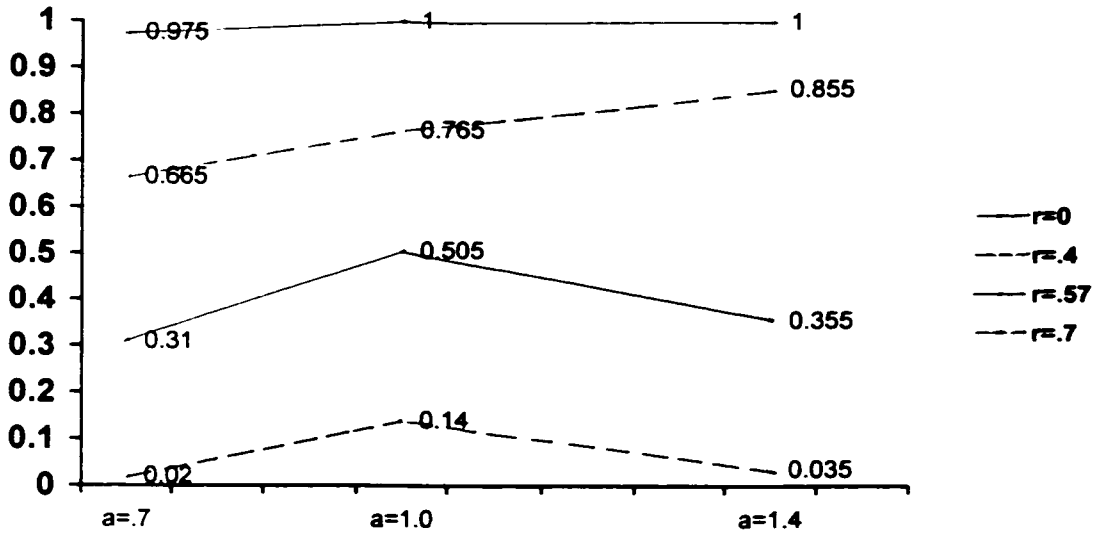
displayed in Table 22. For both indices, in the logit-linear analysis, none of the parameter estimates was significant. However, eliminating this interaction violated the model fit criterion, P ≥.15. The effects of discrimination by correlation for Stout's T_1 were shown in Figure 4.

Table 22

The Effects of Discrimination by Correlation on the Power of Stout's T_1 and Stout's T_2 for N=2,000 with Two Dimensional Complex Test Structure Data Sets

	Power					
	Stout's T_1			**Stout's T_2**		
	a=.7	a=1.0	a=1.4	a=.7	a=1.0	a=1.4
r=.0	**.975**	**1.000**	**1.000**	**.980**	**1.000**	**1.000**
r=.4	.665	.765	*.855*	.690	.775	*.880*
r=.57	.310	.505	.355	.340	.530	.410
r=.7	.020	.140	.035	.030	.155	.040

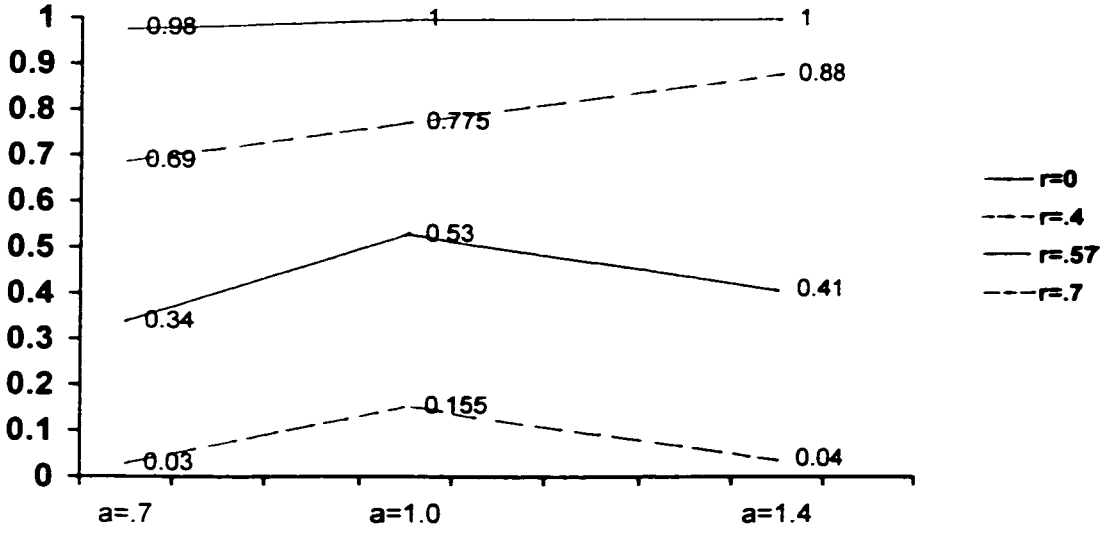
Figure 4 The Effects for Discrimination by Correlation for Stout's T_1 for N=2,000



Although all the effects were not significant, it appears that the effects between r=.4 and r=.57 for a=1.0 and a=1.4 was quite substantial. Similar effect was also observed between r=.4 and r=.7 for a=1.0 and a=1.4 For the former effect, there was an increase of .09 for r=.4 but a decrease of .15 for r=.57. The difference in power change between the two levels of correlation was .24. For the latter, power change was .195. These two effects contributed to the overall model fit.

With discrimination by correlation interaction for Stout's T_2, similarly, parameter estimates in logit-analysis indicated that none of the parameter estimates was significant. However, as shown in Table 22 and Figure 5, the effects between r=.4 and r=.57 and r=.4 and r=.7 for a=1.0 and a=1.4 were substantial. For r=.4, there was an increase of .105 compared to

Figure 5 The Effects for Discrimination by Correlation for Stout's T_2 for N=2,000



a decrease of .12 for r=.57 and a decrease of .115 for r=.7. The difference in power change between r=.4 and r=.57 was .225 and between r=.4 and r=.7 was .22. The direction of these effects were quite different. Again, these two effects contributed to the overall model fit.

In the following section, sample size effects for Stout's T_1 and Stout's T_2 are examined and discussed. It appears as sample size increased, more interactions were present. For example, when a sample size of 500 was used, the logit-linear model included only discrimination by correlation interaction. As sample size increased to 1,000, for Stout's T_1, the logit-linear model included two 2-way interactions, test length by correlation and discrimination by correlation while for Stout's T_2, the logit-model included three 2-way interactions: test length by discrimination, test length by correlation, and discrimination by correlation. When sample size moved from 1,000 to 2,000, the logit-linear model included all three 2-way interactions for both indices.

Sample size effects for discrimination by correlation interactions are shown in Table 23. For both indices, for N=500 and N=1,000 at r ≥.57, as sample size and discrimination increased, power increased except for r=.7 for Stout's T_1 (see Table 23). For N=2,000 and r≤.4, increasing discrimination increased power. For r≥.57, as discrimination increased from .7 to 1.0, power increased. As discrimination increased from 1.0 to 1.4, power decreased. At r=.7, both indices did not work well for all three sample sizes and discriminations. Under these conditions, power ranged from .01 to .155. For a=1.4 and r ≥.57, both indices lacked power with N≥1,000. Under these conditions, increasing sample size either decreased power or had no effects on power (see Table 23).

Table 23

Sample Size Effects for Discrimination by Correlation Interactions for Stout's T_1 and Stout's T_2

	Stout's T_1				Stout's T_2		
	a=.7	a=1.0	a=1.4		a=.7	a=1.0	a=1.4
				N=500			
r=.0	.355	.690	*.935*		.415	.745	**.950**
r=.4	.070	.260	.630		.115	.320	.665
r=.57	.025	.065	.210		.060	.075	.290
r=.7	.030	.030	.035		.065	.055	.090
				N=1,000			
r=.0	*.815*	**.980**	**1.000**		*.825*	**.980**	**1.000**
r=.4	.410	.615	.770		.445	.640	.795
r=.57	.110	.250	.410		.150	.320	.455
r=.7	.005	.045	.025		.010	.055	.060
				N=2,000			
r=.0	**.975**	**1.000**	**1.000**		**.980**	**1.000**	**1.000**
r=.4	.665	.765	*.855*		.690	.775	*.880*
r=.57	.310	.505	.355		.340	.530	.410
r=.7	.020	.140	.035		.030	.155	.040

Since in the logit-linear analysis, test length by discrimination interaction was not present with N=500 for each index and with N=1,000 for Stout's T_1, sample size effects could only be compared between N=1,000 and N=2,000 for Stout's T_2 for this interaction. Sample size effects between N=1,000 and N=2,000 for test length by discrimination interactions for Stout's T_2 are shown in Table 24.

In most cases, for each discrimination, increasing sample size and test length increased power. The increase was more pronounced for a=1.0, L=80, and N=2,000. However, for a=1.4 , L=80, and N=2,000, power decreased.

Table 24

Sample Size Effects Between N=1,000 and N=2,000 for Test Length by Discrimination Interactions for Stout's T_2

	N=1,000		N=2,000	
	L=40	L=80	L=40	L=80
a=0.7	.210	.505	.348	.673
a=1.0	.353	.565	.435	.795
a=1.4	.480	.675	.520	.645

Sample size effects on test length by correlation interaction are presented in Table 25. Since this interaction was not significant for N=500 base on the logit-linear analysis, only N=1,000 and N=2,000 were used in the comparison.

For both indices, with L=40, good or excellent power was shown for r=.0 for N=1,000 and N=2,000. For the other correlation levels, stronger sample size effects were shown for r=.4 (.217 higher for N=2,000 for both indices) than other levels of correlation. While for r ≥.57,

Table 25

Sample Size Effects Between N=1,000 and N=2,000 for Test Length by Correlation Interactions for Stout's T_1 and Stout's T_2

	N=1,000			
	Stout's T_1		Stout's T_2	
	L=40	L=80	L=40	L=80
r=.0	*.867*	**.997**	*.873*	**.997**
r=.4	.310	*.887*	.350	*.903*
r=.57	.103	.410	.140	.477
r=.7	.007	.043	.027	.063
	N=2,000			
r=.0	**.983**	**1.000**	**.987**	**1.000**
r=.4	.527	**.970**	.567	**.997**
r=.57	.120	.660	.167	.678
r=.7	.030	.127	.017	.133

an increase in sample size did not make much improvement in power. With L=80, good or excellent power was observed for r≤.4. A greater sample size effect was shown for r=.57 with N=2,000 with an increase of .25 and .201 for r=.57 for Stout's T_1 and Stout's T_2 respectively.

The sample size effects on the independent variables are presented in Table 26. For both indices, for a=.7 and a=1.0, power increased as sample size increased. For a=1.4, power increased as sample size increased from 500 to 1,000. From 1,000 to 2,000, power was fairly.

Table 26

Sample Size Effects on the Independent Variables for Stout's T_1 and Stout's T_2

Independent Variables	Stout's T_1			Stout's T_2		
	N=500	N=1,000	N=2,000	N=500	N=1,000	N=2,000
a=0.7	.126	.335	.493	.164	.358	.510
a=1.0	.261	.473	.603	.299	.501	.615
a=1.4	.453	.551	.561	.499	.578	.583
r=.00	.660	**.932**	**.992**	.703	**.935**	**.993**
r=.40	.320	.598	.762	.367	.627	.782
r=.57	.100	.257	.390	.142	.308	.427
r=.70	**.032**	**.025**	**.065**	**.070**	**.045**	**.075**
L=40	.166	.322	.408	.208	.348	.434
L=80	.390	.584	.696	.433	.610	.704

stable. For correlation, in quite a few cases, ceiling or floor effects resulted in some of the interactions (see the underlined highlighted numbers in Table 26). For each test length, power increased as sample size increased. Greater increase was observed between N=500 and N=1,000 than between N=1,000 and N=2,000.

The summary and discussion of the results are presented in the following chapter.

Chapter V

SUMMARY AND DISCUSSION OF THE RESULTS

In this section, the summary followed by the discussion of the results from the unidimensional and multidimensional studies for each index is presented.

Summary of the Results

The Approximate χ^2

Based on the results of the unidimensional study, the Approximate χ^2 had excellent Type I error control with zero rejections of the unidimensional assumption across all conditions simulated in this study.

In the multidimensional case, the power of Approximate χ^2 was shown to be very sensitive to test structure (simple test structure and complex test structure) simulated in this study. With two dimensional simple test structure the Approximate χ^2 demonstrated excellent power under all conditions except for the cases of r=.7 for a=.7 and r=.7 for a=1.0 with N=500. Sample size by correlation and sample size by discrimination interactions were present. The sample size by correlation effect was observed at the cell of N=500 and r=.7 where power dropped dramatically from .96 for r=.57 to .53 for r=.7. For N≥1,000, increasing correlation did not significantly affect power. A sample size by discrimination interaction existed at N=500 wherein increasing discrimination from .7 to 1.0 was associated with an increase in power while for the other sample sizes, discrimination had no significant effects on the power of this index. Excellent power was achieved for N≥1,000 across the three discrimination levels. Ceiling effects seem to be the cause of both sample size by correlation and sample size by discrimination

interactions. Test length did not appear to have much effect on the power of the Approximate χ^2. The mean power of the Approximate χ^2 was .956.

The picture for complex structure was very different. The Approximate χ^2 performed extremely poorly under most conditions. There were a few exceptions where good or excellent power was shown. For example, for a≤1.0, excellent power was shown for N=2,000 for r=.0 with both test lengths. For a=1.0 and L=80, when N=1,000 was used, good power of .88 was achieved. For a=1.4 and N=2,000, the Approximate χ^2 performed extremely well. Under this condition, test length and correlation had little or no effect on power. The minimum power was .87 for r=.7 and L=40. The dramatic improvement in power for the Approximate χ^2 is likely due to the high discrimination value (1.4) and large sample size. This finding should be investigated in future studies. Overall, for complex structure, good or excellent power was obtained for 15 of 72 occasions.

The Approximate χ^2 was found to have extremely good Type I error control, poor power with two dimensional complex test structure under most conditions, but good power with two dimensional simple test structure.

Stout's T_1 and Stout's T_2

In the unidimensional case, both Stout's T_1 and Stout's T_2 was shown to have Type I error rates lower than nominal level under most conditions. Inflated Type I error rates were observed in a few cases. For example, Type I error of .07 was observed for Stout's T_1 and .12 for Stout's T_2 with N=500 and L=80 at a=1.0. Type I error of .07 was also observed with N=500 and L=80 at a=1.4 for Stout's T_2. Based on the Logit-linear analysis results, only the main effects of sample

size and test length affected the Type I error rates of the two indices. Smaller sample size and longer test length resulted in higher Type I error rates with N=500 and L=80 being more pronounced.

Based on the results of the multidimensional study, the use of different test structures made a major difference in the power of Stout's T_1 and Stout's T_2. Excellent power was observed for Stout's T_1 and Stout's T_2 with simple test structure in almost all conditions. Although the power of the two indices was shown to be affected by the independent variables, the effect was weak. Except for the cell of a=.7, r=.7 and N=500 for L=40 where low power, .58 for Stout's T_1 and .62 for Stout's T_2, was shown, good (.80 or above) or excellent power (.95 or above) was achieved.

The power of Stout's T_1 and Stout's T_2 decreased remarkably for complex test structure. It was inadequate under many conditions. Slightly higher power was observed for Stout's T_2 than Stout's T_1 with a mean power of .45 as compared to a mean power of .43 for Stout's T_1. For each index, good or excellent power was obtained for only 21 of 72 combinations. In logit-linear analysis for Stout's T_1 and Stout's T_2, the simplest model included the highest order interactions (4-way interactions). To simplify the analysis, Logit-linear analysis was carried out for each sample size. Based on the results, discrimination, test length, and correlation each affected the power of the two indices. As sample size increased, more interactions were present. For example, for N=500, the logit-linear model involved only a discrimination by correlation interaction as well as main effects. For the main effects, power increased as test length and discrimination increased and correlation decreased. For Stout's T_1, significant discrimination by correlation

interaction effects were found between r=.0 and r=.4 and between r=.4 and r=.57 for a=.7 and a=1.0 as well as for a=1.0 and a=1.4. As discrimination increased from .7 to 1.0, there was a much greater increase in power for r=.0 than for r=.4 and a much greater increase in power for r=.4 than for r=.57. When discrimination increased from 1.0 to 1.4, a much greater increase in power occurred for r=.4 than r=.0 and r=.57. For Stout's T_2, except for the effect between r=.57 and r=.7 for a=.7 and a=1.0, all the discrimination by correlation interaction effects were significant. As discrimination increased from .7 to 1.0, there was a greater increase in power for r=.0 compared to less increase for r=.4 as well as a greater increase for r=.4 compared to less increase for r=.57. When discrimination went from 1.0 to 1.4, a much greater increase in power was observed for r=.4 than the other levels of correlation. Also, a greater increase in power occurred for r=.57 compared to much less increase for r=.7. Ceiling or floor effects account for some of the interactions (see Table 15).

For N=1,000, the simplest model included main effects and two 2-way interactions (test length by correlation and discrimination by correlation) for Stout's T_1. For Stout's T_2, the best model involved main effects and all three 2-way interactions (test length by discrimination, test length by correlation, and discrimination by correlation). For both indices, with the main effects power increased as test length and discrimination increased and correlation decreased. For Stout's T_1, with discrimination by correlation interaction, only two effects were shown to be significant. These effects were those between r=.0 and r=.4 for a=1.0 and a=1.4 as well as the effects between r=.57 and r=.7 for a=1.0 and a=1.4. A greater increase was observed for r=.4 than for r=.0, and for r=.57 than r=.7 as discrimination increased from 1.0 to 1.4. Again, ceiling effect

for r=.0 and floor effect for r=.7 resulted in these interactions. With test length by discrimination interaction, that none of the effects were shown to be significant by the logit-linear analysis. For test length by correlation interaction, all the effects were significant. A much greater increase in power was shown for r=.4 than other levels of correlation as test length increased. For r=.57, there was also a much greater power increase compared to r=.7 as test length increased. Good or excellent power for r=.0 were obtained across the three discrimination levels while extremely poor power was observed for r=.7. These ceiling and floor effects contributed to most of the interactions (see Table 16 and Table 17).

For Stout's T_2, with discrimination by correlation interaction, only one effect was shown to be significant by the logit-linear analysis. This effect was due to a much greater increase in power for r=.57 compared to a negligible increase for r=.7 as discrimination increased from 1.0 to 1.4. However, other non significant effects may be combined to contribute to overall model fit. As for test length by correlation interaction, all the effects were shown significant by logit-analysis. These effects were attributable to the fact that as test length increased, there was a much greater power increase for r=.4 than other levels of correlation. Also, there was a much greater increase in power for r=.57 compared to r=.7. Similar to Stout's T_1, ceiling effect for r=.0 and floor effect for r=.7 caused most of the test by correlation and discrimination by correlation interactions (see Table 16 and Table 17). None of the effects for test length by discrimination were significant.

For N=2,000, based on the results of the logit-linear analysis, the simplest model included three 2-way interactions (test length by discrimination, test length by correlation, and

discrimination by correlation) as well as the main effects. For both indices, with the main effects, power increased as discrimination and test length increased and correlation decreased. For both indices, all test length by discrimination interaction effects were significant. However, these interaction effects were disordinal. These disordinal interaction effects were attributable to the fact that power was highest for a=1.4 for the 40 item test yet lowest for the 80 item test (see Table 19, Figure 2 and Figure 3). None of the discrimination by correlation effects were significant. However, some of the non-negligible effects might cumulatively contribute to the overall model fit. For example, for r=.57 and r=.7 there was a substantial decrease in power for a=1.4 compared to an increase in power for r=.4 (see Table 21, Figure 4, and Figure 5). The test length by correlation interaction was due to the fact that there was a substantially greater power increase for r=.4 and r=.57 compared to a ceiling effect for r=.0 and a floor effect for r=.7 as test length increased.

In conclusion, the Approximate χ^2 demonstrated perfect Type I error control under all conditions and strong power with two dimensional simple test structure in most cases. However, the Approximate χ^2 failed with two dimensional complex test structure under most conditions except for the case of a=1.4 and N=2,000, wherein good or excellent power was observed. It also performed well with r=.0 and N=2,000 for a=.7 and a=1.0. Stout's T_1 and Stout's T_2 had lower Type I error rates than nominal level in most cases except for the conditions of N=500, a=1.0 for L=80 and N=500, a=1.4, and L=80 where Type I error rates were inflated. The maximum Type I error rate of .12 was observed for N=500, a=1.4 and L=80. Slightly better power was observed for both indices with two dimensional simple test structure and better power with two

dimensional complex test structure than the Approximate χ^2. However, under quite a number of conditions, unsatisfactory power with two dimensional complex test structure was found for Stout's T_1 and Stout's T_2. All the interaction effects were ordinal except for test length by discrimination for N=2,000. Quite unexpectedly, Stout's T_1 and Stout's T_2 did not perform well at a=1.4 with r≥.57 as increasing sample size from 1,000 to 2,000 did not seem to enhance power; instead, a negative effect was observed in some cases. Overall, for only 21 of 72 combinations of discrimination, sample size, correlation, and test length did they reach good or excellent power (≥.80). In contrast, with for the Approximate χ^2, good or excellent power was found for 15 of 72 combinations.

Discussion

In this section, results for the three indices are compared and discussed. The Approximate χ^2 had excellent Type I error control, excellent power with two dimensional simple test structure, and poor power with two dimensional complex test structure. Stout's T_1 and Stout's T_2, on the other hand, had slightly higher Type I error rates than the Approximate χ^2 though in most cases the Type I error rates were lower than the nominal level with a few exceptions. Both indices demonstrated excellent power with two dimensional simple test structure and better power with two dimensional complex test structure than the Approximate χ^2. The slightly higher Type I error rates and better power of Stout's T_1 and Stout's T_2 than the Approximate χ^2 seem to suggest that there is a direct relationship between the Type I error rates and the power demonstrated by the three indices. The extremely poor power with two dimensional complex test structure and zero Type I error rates shown by the Approximate χ^2 seem to suggest some cause for concern for the

conservative performance of this index. Although, this situation did not appear to be a problem with two dimensional simple test structure and two dimensional complex test structure under the condition of N=2,000 and a=1.4, it does not exclude the more conservative nature of this index. The extremely low Type I error rates of the Approximate χ^2 may be also enhanced by the fact that a uniform discrimination value was used to simulate item responses. It is likely that no variability in item discrimination played a part in the Type I error rate for the Approximate χ^2. Gessaroli, De Champlain, and Folske (1997) had similar findings when they simulated unidimensional data sets using a uniform discrimination value of .80. Other researches (Breithpaut, 1995; Gessaroli & De Champlain, 1996) also showed the Approximate χ^2 had Type I error well below the nominal level. The extremely good performance of the Approximate χ^2 with high discrimination (a=1.4) and large sample size (N=2,000) and the extremely poor performance with complex structure under most of the other conditions suggests the dependency of this index on discrimination level and sample size. Gessaroli, De Champlain, and Folske (1997) also showed that the Approximate χ^2 was very sensitive to sample size. These authors found that for a 45 item test with simple test structure and a dimensional correlation of .7, the power of this index was only .02 for N=1,000. However, when N=5,000 was used, power increased dramatically to 1.0. The findings in this study raised an interesting question: How might the results of the Approximate χ^2 have differed with two dimensional complex test structure with sample sizes larger than 2,000?

The slightly higher Type I error rates and better power with two dimensional complex test structure shown by Stout's T_1 and Stout's T_2 appear to indicate a slightly more liberal nature of

these indices than the Approximate χ^2. However, the power of Stout's T_1 and Stout's T_2 was far from optimal in many cases. It is also interesting to note that as discrimination and sample size increased, the Type I error rates of the two indices decreased while longer test length produced higher Type I error rates. Correspondingly, test length had a strong effect on the power of the two indices with complex test structure.

The result that longer test length produced higher Type I error rates for Stout's T_1 and Stout's T_2 was different from earlier researchers (De Champlain, 1992; Breithaupt, 1995; Gessaroli & De Champlain, 1996) who stated that test length had no effect on the Type I error rates of the two indices. However, most of these researchers used a test length shorter than 80 items. For example, De Champlain (1992) and Gessaroli & De Champlain (1996) compared 15, 30, and 45 item tests. Breithaupt (1995) compared 30 and 45 item tests.

The fact that longer test length produced higher Type I error rates for these two indices was somewhat contradictory to Stout's asymptotic theory that the mean absolute covariance between items j and k at all levels of ability approaches zero as the test length approaches infinity. Accordingly, increasing test length should produce a lower Type I error rate, since longer test has smaller mean absolute conditional item covariance. The finding does not seem to support this theory. One alternative to assess Stout's asymptotic theory would be to examine the change in the amount of the conditional item covariance as test length increases.

When the performance of Stout's T_1 and Stout's T_2 was compared, Stout's T_2 was found to have slightly higher Type I error rates and power than Stout's T_1. The result was corresponding to the statement of Stout, Douglas, Junker, and Roussos (1993) that Stout's T_2 is more powerful in

simulation studies at the expense of a higher Type I error rate. Yet, both the higher Type I error and gain in power were small. This finding was similar to that of Breithaupt (1995) who showed that the power of Stout's T_2 was only slightly greater than that of Stout's T_1.

The results concerning the power of the three indices seems to suggest that they are very sensitive to different test structures simulated. Comparing the results of the present study with those of previous researchers appears to indicate that the three indices are also sensitive to dimension dominance. For example, Gessaroli, De Champlain, and Folske (1997) showed that both Stout's T_2 and the Approximate χ^2 did not work well with highly correlated (r=.7) two dimensional simple test structure. They found that under the condition of r=.7 and a=.8 for N=1,000, the Approximate χ^2 had power of .02 for L=45 and .03 for L=75. Stout's T_2 had power of .44 for L=45 (data for L=75 were not reported). Breithaupt (1995) showed similar findings. According to Gessaroli, De Champlain, and Folske (1997), the conservative nature of the Approximate χ^2 could be attributable to the loss of information incurred when transforming the residual joint proportions to Fisher z values and finally correlation coefficients. However, this situation was not obvious with two dimensional simple test structure based on the results of the present study. Under similar conditions (a=.7, N=1,000, and r=.7) used by Gessaroli, De Champlain, and Folske (1997), strong power was observed for both the Approximate χ^2 and Stout's T_2. The Approximate χ^2 had power of .87 with L=40 and .99 with L=80 and Stout's T_2 had power of .98 with L=40 and 1.0 with L=80. One possible explanation of the different findings was that the test structure simulated with different dimensional importance in these studies may have resulted in different power. Breithaupt (1995) and Gessaroli, De Champlain,

and Folske (1997) used a simple structure with 80% of the items loaded on the first dimension and 20% loaded on the second dimension. In the present study, more importance was put on the second dimension. That is, 75% of the items were related to the first dimension and 25% to the second dimension. This may also explain the similar findings between the present study and the findings of De Champlain and Gessaroli (1998). They found that the Approximate χ^2 demonstrated extremely good power (1.0) with a similar two dimensional simple test structure to the one used in the present study.

The power of the Approximate χ^2, Stout's T_1, and Stout's T_2 dramatically decreased with two dimensional complex test structure, though Stout's T_1 and Stout's T_2 were much less affected. Specifically, the Approximate χ^2 completely failed in many cases. For example, For the Approximate χ^2, for 15 of 72 combinations, the power was greater than .80 and most cases occurred for a=1.4 and N=2,000. For both Stout's T_1 and Stout's T_2, for 21 of 72 combinations, the power was greater than .80. The findings were inconsistent with the findings of previous researchers. Nandakumar (1994) and Nandakumar and Stout (1993) showed that the power of Stout's T_2 was .97 with highly correlated (r=.7) two dimensional complex test structure. De Champlain and Gessaroli (1998) found that the Approximate χ^2 had extremely good power (1.0) with highly correlated (r=.7) two dimensional complex test structure. One possible reason for the different findings was that the different dimension dominance simulated may have played a part. All the previous authors used a structure where two dimensions were set as equally important. In the present study, the complex structure was simulated with the first dimension set as dominant and the importance of the second dimension gradually increased. This resulted in a dimension

ratio of 3:1 i.e. a structure with a less important second dimension than the one used by previous researchers. The different findings seem to suggest that the performance of the three indices is dependent upon how two dimensional complex test structure is simulated.

An interesting finding of the present study was that both Stout's T_1 and Stout's T_2 did not work well with high discrimination (1.4) and large sample size (2,000) as correlation increased. Power intended to decrease as high discrimination (1.4) and larger sample size (2,000) were used at $r \geq .57$ (see Table 14). Stout (1987) pointed out that the poor performance of Stout's T_1 with high discrimination was likely to be attributable to the combination of guessing with high discrimination value. However, in the present study, both Stout's T_1 and Stout's T_2 did not seem to work well with a=1.4, $r \geq .57$, and N=2,000 even under a no guessing condition.

It is not surprising that the power of the Approximate χ^2, Stout's T_1, and Stout's T_2 decreased as correlation increased because increasing correlation will bring the dimensionality closer to one, i.e. noncentrality will be smaller. In this study, as the other variables increased, the increase in power for Stout's T_1 and Stout's T_2 resulted in ceiling effects at r=.0 and floor effects for r=.7. These ceiling and floor effects were likely due to large noncentrality differences for these two correlation levels.

Based on the present study, the differences between the performance of the Approximate χ^2, Stout's T_1, and Stout's T_2 are obvious in terms of Type I error control and power. The Approximate χ^2 had excellent Type I error control and performed well in identifying two dimensional simple test structure, yet, had an extremely conservative nature in a majority of cases with two dimensional complex test structure. Stout's T_1 and Stout's T_2, on the other hand,

had slightly higher Type I error rates under most conditions, worked well in identifying two dimensional simple test structure, and yielded better power with two dimensional complex test structure than the Approximate χ^2.

According to Mood, Graybill, and Boes (1974), a good index is one for which the Type I error rate is small (ideally 0) and the probability of rejecting a false null hypothesis is large (ideally unity). Based on the conditions simulated in this study, none of the three indices satisfies these criterion. Strictly speaking, an ideal test such as the one Mood, Graybill, and Boes (1974) mentioned is never found. The Approximate χ^2 had good Type I error rate control and yielded poor power with two dimensional complex test structure, while Stout's T_1 and Stout's T_2 had better power with two dimensional complex test structure but slightly higher Type I error rates. The poor power of the Approximate χ^2 with two dimensional complex test structure is likely attributable to its extremely low Type I error rate. Stout's T_1 and Stout's T_2 had slightly higher Type I error rates and correspondingly better power than the Approximate χ^2 with two dimensional complex test structures.

The extremely good performance of the Approximate χ^2 with large sample size and high discrimination with two dimensional complex test structure may suggest heavy dependence on sample size and discrimination. It seems that to increase its power with two dimensional complex test structure, a larger sample size than 2,000 may be required especially with low discrimination values. This discussion given hereby can not be taken as absolute conclusions as the test items were not typical of the real world. However, it does suggest that further research is warranted.

Conclusions, limitations of the study, and suggestions for future research are presented in

Chapter VI.

Chapter VI

CONCLUSIONS

In this chapter, implications of the results of the study and recommendations to practitioners are first presented. Then, limitations of the study are described and suggestions for research are offered.

Implications and Recommendations

It should be pointed out that the results obtained are heavily dependent upon the data manipulated in this study. Readers and practitioners should be very careful in generalizing them to other conditions. The results are influenced by the methods used to generate data in this study. For example, the same discrimination value used for each item in a test is not realistic. Guessing was not used. The comparisons of power did not take into account the noncentrality differences.

However, useful information is provided in this study, especially the information concerning the impact of test structure and discrimination values. Based on this study, the three indices have been shown to be sensitive to different test structures. Therefore, in choosing appropriate dimensional assessment techniques, test structure is an important factor to be considered. For example, when a two dimensional simple structure is examined, each of the three indices may be used. For two dimensional complex structure, such factors as dimension dominance, discrimination, test length, dimension correlation, and sample size should be carefully considered before choosing any of the three indices.

It is clear that each index examined in this study possesses certain weaknesses. It is

difficult to achieve good power and keep a nominal Type I error rate at the same time. For example, the Approximate χ^2 had excellent Type I error control but lacked power with two dimensional complex structure. Stout's T_1 and Stout's T_2 yielded better power with two dimensional complex structure and also slightly higher Type I error rates. In dimensional assessment, it may be important to obtain additional information using auxiliary techniques through examining the amount of item residual covariance or conditional item covariance. One advantage of the Approximate χ^2 is that it is based on nonlinear factor analysis, which allows the researcher to fit two, three, or more dimensional models to the data sets. Thus, as an auxiliary method, assessing more than one dimension model fit is another way of confirming multidimensionality of the test. In this case, more accurate estimates can be made through comparing two pieces of information, the values of the Approximate χ^2 and item residual covariance obtained after model fit. Gessaroli, De Champlain, and Folske (1997) suggested that another index, the Approximate Likelihood Ratio $\chi^{2.}$ based on McDonald's nonlinear factor analysis be used. They found that the Approximate Likelihood Ratio χ^2 yielded better power than the Approximate χ^2 with multidimensional simple test structure involving high dimensional correlation.

Recently, descriptive statistics have been developed based on Stout's nonparametric framework. These techniques are Agglomerative Hierarchical Cluster Analysis (HCA) (Stout, Habing, Douglas, Kim, Roussos, & Zhang, 1996), Dimensionality Evaluation To Enumerate Contributing Traits (DETECT) (Kim, Zhang, & Stout, 1995), and Δ, an index of multidimensionality (Gao, 1997). These techniques are used to determine a test's dimensional

structure or the amount of multidimensionality involved by assessing the covariance of item-pair responses conditioning on examinees' trait levels. The information obtained through these techniques can be used as supporting information to Stout's T_1, Stout's T_2, and other dimensional assessment methods.

Limitations of the Study and Suggestions for Future Research

The major limitations of the study are those often encountered in Monte Carlo studies. One of the limitations is that the data in this study were simulated using a two parameter compensatory model. A guessing parameter was not specified. Another aspect of lacking reality of the data was that a uniform discrimination level was used for all the items in the test. This may be too much a sacrifice of reality for assessing impact of specific discrimination levels on the performance of the three indices. Also, noncentrality differences were not considered in the comparisons of the power. This has restricted the generalizability of the results.

Second, in the multidimensional study, only one simple test structure and one complex test structure were examined. Based on the analysis results, the Approximate χ^2, Stout's T_1, and Stout's T_2 are sensitive to different test structures. It would be more informative if different examples of test structure other than those used in the study had been examined. For example, a structure often found in a reading test can be studied where there is one general dimension and one or more subdimensions where no item is related to more than one of these subdimensions.

Finally, in this study, the Approximate χ^2 was found to be conservative with multidimensional complex test structure. A well-known disadvantage of any χ^2 is that it is dependent on sample size. In statistical analysis, it is a common practice to use relatively large

sample sizes to increase the power of conservative statistics. As mentioned before, Gessaroli, De Champlain, and Folske (1997) found the Approximate χ^2 yielded perfect power with two dimensional simple test structure involving high correlation (.7) when a sample size of 5,000 was used. De Champlain and Gessaroli (1998) showed that the Approximate χ^2 demonstrated perfect power with two dimensional simple test structure and two dimensional complex test structure involving high correlation (.7). The test structures used by these authors were very different from those used in this study. More information may have been obtained if a sample size larger than 2,000 had been used and additional types of simple and complex test structure had been tested.

Based on the results obtained and limitations of this study, the following suggestions are offered regarding future research concerning dimension assessment.

First, it would be important to assess the performance of the Approximate χ^2, Stout's T_1, Stout's T_2, and the Approximate Likelihood Ratio χ^2 with a larger number of simulated test structures. For example, these structures may involve those where the item discrimination varies more realistically. The ratio of the first dimension dominance to the second dimension dominance should also vary. Instead of 3:1 used in this study, the ratio may be set at 2:1, 1:1, 3:2, and 4:1 etc. In addition, one might simulate science reading tests including several paragraphs where in each paragraph heavy emphasis was put on a distinct subject matter (simple test structure), or one might simulate a reading test with a major dimension for reading and two or more minor dimensions relating to passage content (complex test structure). However, in power study, noncentrality differences must be taken into account in any design involving dimension correlation.

Second, in this study, the power of the Approximate χ^2 with two dimensional complex structure was far from optimal. However, its excellent performance with high discrimination (1.4) and large sample size (2,000) presented a sharp contrast to other conditions. For future studies, it would be interesting to know what sample size may be appropriate for the Approximate χ^2 to achieve better power with complex test structure. Due to the fact that the discrimination value was set the same for all items in this study, the generalization of the results is limited. Thus, another focus of future studies should be assessing the impact of different discrimination levels in a more practical way. One way of doing this is to assess indices with data simulated based on the mean and SD of estimates of low, medium, and high discrimination and difficulty from real data. This would allow for more generalization of the results.

Third, according to Gessaroli, De Champlain, and Folske (1997), the Approximate Likelihood Ratio χ^2 demonstrated great improvement over the power of the Approximate χ^2. However, the Approximate Likelihood Ratio χ^2 was examined under limited conditions. It would be interesting for researchers to assess the power of the Approximate Likelihood Ratio χ^2 with a variety of conditions.

Fourth, while evaluating current inferential statistics such as the indices examined in this study and the Approximate Likelihood Ratio χ^2, relevant descriptive methods (e.g. the methods developed recently with Stout's nonparametric framework) may also be used to provide supporting information on the effectiveness of these indices.

Finally, as mentioned above, different data generation methods may affect the assessment results in Monte Carlo studies. The method used in this study generated a data set

that was not realistic. In future research, multiple levels of discrimination should be nested within a test so that more realistic item responses could be simulated for both Type I error and power studies.

This study provided some evidence concerning the performance of the Approximate χ^2, Stout's T_1, and Stout's T_2 with conditions that were different from those used by previous researchers. However, due to the method used to generate the data sets, the generalization of findings was limited. Clearly, much more research should be carried out using other designs and additional indices to assess test dimensionality.

References

Ackerman, T. A. (1987). A comparison study of the unidimensional IRT estimation of compensatory and noncompensatory multidimensional item response data (Report No. 87-12). Iowa City, IA: The American College Testing Program.

Ackerman, T. A. (1994). Creating a test information profile for a two-dimensional latent space. Applied Psychological Measurement, 18, 257-275.

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

Ansley, T.N., & Forsyth, R.A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. Applied Psychological Measurement, 9, 37-48.

Armor, D. J. (1974). Theta reliability and factor scaling. In H. L. Constner (Ed.), Sociological Methodology (pp. 17-50). San Francisco CA: Jossey-Bass.

Batley, R., & Boss, M. W. (1993). The effects on parameter estimation of correlated dimensions and a distribution-related trait in a multidimensional item response model. Applied Psychological Measurement, 17, 131-141.

Ben-Simon, A., & Cohen, Y. (1990). Rosenbaum's test of unidimensionality: Sensitivity analysis. Paper presented at the Annual Meeting of the American Educational Research Association, Boston.

Berger, M. P. F., & Knol, D. L. (1990). On the assessment of dimensionality in multidimensional item response theory models. Paper presented at the meeting of the American Educational Research Association, Boston, MA.

Bentler, P. (1972). A lower-bound method for the dimension-free measurement of internal consistency. Social Science Research, 1, 343-357.

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

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical Theories of Mental Test Scores. Reading, MA: Addison Wesley.

Birnbaum, M., & Tatsuoka, K. (1982). On the dimensionality of achievement test data. Journal of Educational Statistics, 19, 259-266.

Blais, J. G., & Laurier M. (1995). Methodological consideration in using DIMTEST to assess unidimensionality. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Bock, R. D., & Lieberman. M. (1970). Fitting a response model for n dichotomously scored items. Psychometrika, 35, 179-196.

Bock, R.D., & Aitkin, M.A.(1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika, 46, 443-456.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. Applied Psychological Measurement, 12, 261-280.

Breithaupt, K. (1995). Dimensionality of binary response data: a test of two statistical indices. Unpublished Master thesis. University of Ottawa, ON.

Browne, M. W. (1977). The analysis of patterned correlation matrices by generalized least-squares. British Journal of Mathematical and Statistical Psychology, 30, 113-124.

Camilli, G. (1992). A conceptual analysis of differential item functioning in terms of a multidimensional item response model. Applied Psychological Measurement, 16, 129-147.

Carmines, E. G., & Zeller, R. A. (1979). Reliability and Validity Assessment. Beverly Hills CA: Sage.

Carroll, J. B. (1945). The effect of difficulty and chance success on the correlation between items or between tests. Psychometrika, 10, 1-19.

Cattell, R. B., & Tsujioka, B. (1964). The importance of factor-trueness and validity, versus homogeneity and orthogonality, in test scales. Educational and Psychological Measurement, 37, 827-838.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. Psychometrika, 40, 5-32.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16, 297-334.

De Champlain, A. (1992). Assessing test dimensionality using the approximate chi-square statistics. Unpublished doctoral dissertation, University of Ottawa, Ottawa.

De Champlain, A., & Gessaroli, M. E. (1992). Assessing test dimensionality using an approximate chi-square statistic. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, California.

De Champlain, A., & Gessaroli, M. E. (1998). Assessing the dimensionality of item response matrices with small sample sizes and short test lengths. Applied Measurement in Education, 11(3), 231-253.

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

Dubois, P. H. (1970). Varieties of psychological test homogeneity. American Psychologist, 25, 532-536.

Fraser, C. (1988). NOHARM II: A fortran program for fitting unidimensional normal ogive models of latent trait theory. Armidale, N.S.W.: University of New England, Centre for Behavioral Studies.

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

Gao, F., & Stout, W. (1997). A dimtest-based index of the amount of multidimensionality. Paper presented at the NCME annual meeting, Chicago, IL.

Gessaroli, M. E., & De Champlain, A. (1992). The assessment of dimensionality: a review of procedures and methods. Unpublished manuscript.

Gessaroli, M.E. (1994). The assessment of dimensionality via local and essential independence: A comparison in theory and practice. In D. Laveault, B.D. Zumbo, M.E. Gessaroli, & M.W. Boss (Eds.). Modern theories in measurement: Problems and issues. Ottawa, Ontario: University of Ottawa.

Gessaroli, M. E. (1995). Assessing dimensionality using non-linear factor analysis. Paper presented at the annual meeting of the American Educational Research Association, Los Angels.

Gessaroli, M. E., & De Champlain, A. (1996). Using an approximate chi-square statistic to test the number of dimensions underlying the responses to a set of items. Journal of Educational Measurement, 33, 157-179.

Gessaroli, M. E., De Champlain, A. F., & Folske, Jane. C. (1997). Assessing dimensionality using a likelihood-ratio chi-square test based on a non-linear factor analysis of item response data. Paper presented at the annual meeting of the National Council on Measurement in Education. Chicago, IL.

Green, B. F. (1956). A method of scalogram analysis using summary statistics. Psychometrika, 21, 79-88.

Green, S. B., Lissitz, R. W., & Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. Educational and Psychological Measurement, 37, 827-838.

Greene, V. C., & Carmines, E. G. (1980). Assessing the reliability of linear composites. In K. F. Schuessler (Ed.), Sociological Methodology (pp. 160-175). San Francisco CA: Jossey-Bass.

Hambleton, R. K., & Traub, R. E. (1973). Analysis of empirical data using two logistic latent trait models. British Journal of Mathematical and Statistical Psychology, 26, 195-211.

Hambleton, R. K. (1980). Latent ability scales: Interpretations and uses. New Directions for Testing and Measurement, 6, 73-97.

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

Harwell, M., Stone, C. A., Hsu, T., & Kirisci, L. (1996). Monte Carlo studies in item response theory. Applied Psychological Measurement, 20(2), 101-125.

Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. Multivariate Behavioral Research, 19, 49-78.

Hattie, J. (1985). Methodology review: assessing unidimensionality of tests and items. Applied Psychological Measurement, 9, 139-164.

Hattie, J. A., & Hansford, B. F. (1982). Communication apprehension: An assessment of Australian and United States data. Applied Psychological Measurement, 6, 225-233.

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

Heise, D. R., & Bohrnstedt, G. W. (1970). Validity, invalidity, and reliability. In E. F. Borgatta, & G. W. Bohrnstedt (Eds.), Sociological Methodology (pp. 104-129). San Francisco CA: Jossey-Bass.

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. The Annals of Statistics, 14, 1523-1543.

Hulin, C.L., Drasgow, F., & Parsons, L.K. (1983). Item Response Theory. Homewood, IL: Dow-Jones Irwin.

Humphreys, L. G. (1985). General intelligence: An integration of factor, test, and simplex theory. In B. B. Wolman (Ed.), Handbook of Intelligence (pp. 201-224). New York: Wiley.

Humphreys, L. G. (1986). An analysis and evaluation of test and item bias in the prediction context. Journal of Applied Psychology, 71, 327-333.

Hutten, L. (1979). An empirical comparison of the goodness of fit of three latent trait models to real test data. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Hutten, L. (1980). Some empirical evidence for latent trait model selection. Paper presented at the annual meeting of the American Educational Research Association, Boston MA.

Junker, B., & Stout, W. F. (1994): Robustness of ability estimation when multiple traits are present with one trait dominant: A comparison in theory and practice. In D. Laveault, B.D. Zumbo, M.E. Gessaroli, & M.W. Boss (Eds.). Modern theories in measurement: Problems and issues. Ottawa, Ontario: University of Ottawa.

Kim, R. K. Zhang, J., & Stout, W. (1995). A new index of dimensionality--DETECT. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Kingston, N. (1986). Assessing the dimensionality of the GMAT verbal and qualitative measures using full-information factor analysis (Report No. TM 860 575). Princeton, NJ: Educational Testing Service. (ERIC Document Reproduction Service No. Ed 275 698).

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

Laforge, R. (1965). Components of reliability. Psychometrika, 30, 187-195.

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

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from a retrospective study of disease. Journal of the National Cancer Institute, 22, 719-748.

McDonald, R. P. (1967). Nonlinear factor analysis. Psychometrika Monograph, 32(4) Pt. 2.

McDonald, R. P.(1981). The dimensionality of tests and items. British Journal of Mathematical and Statistical Psychology, 33, 161-183.

McDonald, R. P. (1982). Linear versus nonlinear models in item response theory. Applied Psychological Measurement, 6, 379 -396.

McDonald, R. P. (1985). Factor analysis and related methods. Hillsdale, NJ: Lawrence Erlbaum Associates.

McDonald, R. P. (1989). Future directions in item response theory. International Journal of Education Research, 13, 205-220.

McDonald, R. P.(1994). Testing for approximate dimensionality. In D. Laveault, B. D. Zumbo, M. E. Gessaroli, & M. W. Boss (Eds.). Modern theories in measurement: Problems and issues.

McDonald, R. P., & Marsh, H. W. (1990). Choosing a multivariate model: Noncentrality and goodness or fit. Psychological Bulletin, 107, 247-255.

McDonald, R.P., & Mok, M.M.-C. (1995). Goodness of fit item response models. Multivariate Behavioral Research, 30(1), 23-40.

Mckinley, R. (1989). Confirmatory analysis of test structure using multidimensional item response theory (Research Rep. RR - 89 - 31). Princeton, NJ: Educational Testing Service.

Muraki, E., & Engelhard, G. (1985). Full-information factor analysis: Applications of EAP scores. Applied Psychological Measurement, 9, 417-430.

Muthen, B. (1978). Contributions to factor analysis of dichotomous variables. Psychometrika, 43, 551-560.

Nandakumar, R. (1987). Refinement of Stout's procedure for assessing latent trait dimensionality. Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.

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

Nandakumar, R. (1994). Assessing dimensionality of a set of item responses-comparison of different approaches. Journal of Educational Measurement, 31, 17-35.

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

Novic, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. Psychometrika, 32, 1-13.

Oshima, T. C., & Miller, M. D. (1992). Multidimensionality and item bias in item response theory. Applied Psychological Measurement, 16, 237-249.

Reckase, M.D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. Journal of Educational Statistics, 4, 207-230.

Reckase, M.D. (1985). The difficulty of test items that measure more than one ability. Applied Psychological Measurement, 6, 401-412.

Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building unidimensional tests using multidimensional items. Journal of Educational Measurement, 25, 193-203.

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

Rosenbaum, P. (1984). Testing the local independence assumption in item response theory. (Tech. Rep. No. 84-85). Princeton, NJ: Educational Testing Service.

Roussos, L. (1995). Hierarchical agglomerative clustering computer program user's manual. UrbanaChampaign: Statistical Laboratory for Educational and Psychological Measurement, Department of Statistics, University of Illinois.

Roussos, L. (1995). MD4F: A computer program for generating thetas and response strings corresponding to the M2PL model. UrbanaChampaign: Statistical Laboratory for Educational and Psychological Measurement, Department of Statistics, University of Illinois.

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

Smith, K. W. (1974). Forming composite scales and estimating their validity through factor analysis. Social Forces, 53, 169-180.

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

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

Stout, W., Douglas, J., Junker, B., & Roussos, L. (1993). DIMTEST Manual. Department of Statistics, University of Illinois at Urbana-Champaign, IL.

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

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

Thissen, D, Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple categorical-response modes. Journal of Educational Measurement, 26, 247-260.

Traub, R.E. (1983). A prior considerations in choosing an item response model. In R.K. Hambleton (Ed.), Applications of item response theory (pp.57-70). British Columbia, Canada: Educational Research Institute of British Columbia.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. Psychometrika, 38, 1-10.

Watkins, D., & Hattie, J. A. (1980). An investigation of the internal structure of the Bigg's Study process questionnaire. Educational and Psychological Measurement, 40, 1125-1130.

Wilson, D., Wood, R., & Gibbons, R. D. (1987). TESTFACT: Test scoring, item statistics, and item factor analysis. Mooresville, IN: Scientific, Software, Inc.

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, 125-145.

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

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

Zhang, J., & Stout, W. (1996). A new theoretical DETECT index of dimensionality and its estimation. Submitted for publication.

Zwick. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. Psychological Bulletin, 99, 432-442.

Zwick. R. (1987). Assessing the dimensionality of NAEP reading data. Journal of Educational Measurement, 24, 293-308.