# The Effect of Type of Multidimensional Model on the Assessment of Approximate Unidimensionality in Item Response Spaces

M. Padraig M. M. McLoughlin[1]

Morehouse College

Paper presented at the annual meeting of the Eastern Educational Research Association,

Hilton Head, SC, 16 February 2001.

[1] Part of this paper is based on the author's doctoral dissertation, written at Georgia State University and supervised by John H. Neel, Associate Professor of Research, Measurement, and Statistics.

## ABSTRACT

### THE EFFECT OF TYPE OF MULTIDIMENSIONAL MODEL ON THE ASSESSMENT OF APPROXIMATE UNIDIMENSIONALITY IN ITEM RESPONSE SPACES

M. Padraig M. M. McLoughlin, Ph.D.
Department of Mathematics, Morehouse College, Atlanta, GA 30314

The purpose of this study was to investigate the effect type of bivariate multidimensional model has on the power of the DeChamplain - Gessaroli approximate chi-square statistic (DGACS) and the Stout- Nandakumar T statistic (SNTS). Specifically, two multidimensional models were considered: the Sympson (1978) partially compensatory multidimensional IRT (SPCMIRT) model and the McKinley - Reckase (1983) compensatory multidimensional IRT (MRCOMIRT) model. The intent was realized by 1) an empirical investigation of the power of the DGACS as operationalised by the McDonald - Fraser NOHARM procedure followed by the DeChamplain - Gessaroli CHIDIM procedure for testing if a set of item responses is approximately unidimensional with respect to the SPCMIRT and MRCOMIRT models ; 2) an empirical investigation of the power of the SNTS as operationalised by the Stout - Nandakumar DIMTEST procedure for testing if a set of item responses is approximately unidimensional with respect to the SPCMIRT and MRCOMIRT models; 3) a comparison of power between the DGACS and SNTS for SPCMIRT data; 4) a comparison of power between the DGACS and SNTS for MRCOMIRT data; 5) a comparison of power for the DGACS between SPCMIRT and MRCOMIRT data; and, 6) a comparison of power for the SNTS between SPCMIRT and MRCOMIRT data.

For SPCMIRT modeled data the DGACS had type II error rates across all conditions that were lower that the SNTS and demonstrated adequate power in 21 of 27 conditions; whereas, the SNTS demonstrated adequate power in only 1 of 27 conditions.

For MRCOMIRT modeled data the DGACS had type II error rates across most conditions that were lower that the SNTS and demonstrated adequate power in 21 of 27 conditions; whereas, the SNTS demonstrated adequate power in 5 of 27 conditions.

For SPCMIRT modeled data, given that the SNTS suffers with test length less than or equal to 60 items and small or moderate sample sizes, caution might dictate that it be used with confidence only with large samples ($n > 2{,}000$) and long tests ($i > 60$).

When assessing approximate unidimensionality of a short to moderate length test where the number of items is less than or equal to 60, it seems warranted to question the utility of the SNTS for data modeled with the Sympson PCMIRT model and to recommend use of the DGACS as a more appropriate statistic for assessing approximate unidimensionality. Further, it seems warranted to advise caution in the use of the SNTS for data modeled with the McKinley - Reckase COMIRT model; whereas, the DGACS seems a more preferred statistic for assessing approximate unidimensionality when correlation between dimensions is not high.

However, more research is needed and it is advised that a more diverse sets of parametres, conditions, and dimensions should be studied before any conclusions regarding the utility of the SNTS or DGACS for use with PCMIRT or COMIRT models can be stated with *high* confidence.

This research extends previous results by varying test length, sample size, and considering a more diverse set of parameter combinations than was previously considered. In addition, this research investigated the comparative utility of DGACS and SNTS for testing if a set of item responses is approximately unidimensional with respect to the Sympson PCMIRT model which heretofore had not been executed.

## INTRODUCTION

Item Response Theory (IRT) is a general statistical theory which relates item and test performance to examinee ability measured by the items. It is a theory which links observable data (item and test performance) to unobservable data (examinee ability). IRT item statistics are not sample or test dependent; thus, "examinees [can be] compared even when the examinees have not taken the same items" (Crocker & Algina, 1986, p. 346) and the resulting estimates of IRT `parameters will not vary greatly. IRT statistics are expressed at the item level such that item statistics are not contingent on the test from whence they came; ergo, the statistics are locally computed then globally excogitated. Furthermore, the IRT "provides a measure of precision for each ability level" (Hambleton, Swaminathan, & Rogers, 1991, p. 5). Denote the mathematical models employed in IRT as item response models (IRMs), IRMs specify "an examinee's probability of answering a given item correctly depends on the examinee's ability or abilities and the characteristics of the item" (Hambleton, et al., 1991, p. 9). Most IRMs employed have associated with them the underlying assumptions of invariance, monotonicity, local independence, and a particular dimensionality (most often unidimensionality).

The DeChamplain - Gessaroli approximate $\chi^2$ statistic (DGACS) and Stout-Nandakumar T statistic (SNTS) are claimed to be non-parametric statistics for assessing approximate unidimensionality of binary item responses and are operationalised in the computer programs CHIDIM (DeChamplain & Tang, 1997a) and DIMTEST (Stout, Junker, Nandakumar, Chang, & Steidinger, 1992a), respectively. Therefore, it is claimed that no particular parametric distribution is assumed for the underlying ability distribution or for the item response functions (IRFs) that generate the item responses in the mathematical derivation of the probability distributions of $\chi^2$ or T (see Gessaroli & DeChamplain, 1996 and Stout, 1987 & 1990 for details).

"The assumption common to the IRT models most widely used is that one ability is measured by the items that make up the test. This is called the assumption of unidimensionality" (Hambleton, et al., 1991, pg 9). Since most IRT procedures (BILOG, LOGIST, etc.) used in practice today presume unidimensionality, it seems unreasonable to expect any IRT model where the number of items is greater than or equal to one ($n \geq 1$) to be unidimensional because in practical testing situations the response to an item is dependent on several proficiencies. So, use of said procedures needs to be justified by a statistical process which confirms *approximate*

unidimensionality so the use of IRT in situations where there is more than one dimension can be justified.

For example, consider the following mathematics item: Solve for x: $x^2 - 3x + 5 = 0$. A response to this item not only depends on the examinee's understanding of algebra, but is possibly influenced to some degree by the amount of study time, conditions of the testing situation, the examinee's health, etc. Hence, various authors (see Humphreys, 1982; Traub, 1983; or Harrison, 1986 for more detailed arguments) opine that the unidimensional item response theory (UIRT) assumption of unidimensionality will always be violated with real data. Indeed, regardless of the test theory employed, "such tests are found to measure one [or more] major or dominant dimension[s], but also numerous small factors representing non-error noise intrinsic to behavioral measurement of psychologically [or educationally] important traits…" (Humphreys, 1982, p. 2). Thus, "the key question becomes the assessment of the amount of deviation from the [unidimensional] assumption that can occur in the multidimensional model generating test response data before the claimed performance properties of a particular statistical procedure used for a particular application . . . becomes unacceptably degraded" (Junker & Stout, 1994, p. 32).

DeChamplain and Gessaroli have proposed an approximate chi-square procedure to test for approximate unidimensionality of an incomplete latent space based on McDonald's weak principle of local independence (see McDonald 1967, 1979, 1981, 1994 for theoretical details). DeChamplain and Tang's computer program, CHIDIM, tests the null hypothesis that the off-diagonal elements of a residual correlation matrix are equal to zero after fitting a one dimensional NLFA model to a data matrix using the computer program NOHARM87 (Fraser, 1998). DeChamplain and Gessaroli have demonstrated its utility for testing if a set of item responses is approximately unidimensional with respect to a normal ability distribution. DeChamplain and Tang (1993) demonstrated the utility of two approximate chi-squared statistics', precursors of the DeChamplain - Gessaroli approximate $\chi^2$ statistic, for testing if a set of item responses is approximately unidimensional with respect to normal, positively skewed leptokurtic, and negatively skewed leptokurtic ability distributions. McLoughlin (2000a, 2000b) demonstrated the utility of the DeChamplain - Gessaroli approximate $\chi^2$ statistic for testing if a set of item responses is approximately unidimensional with respect to normal, uniform, bimodal, "positive"

chi-square, "negative" chi-square, positively skewed, and negatively skewed ability distributions for short to moderate length tests.

Stout and Nandakumar have proposed an asymptotically normal procedure to test for approximate unidimensionality of an incomplete latent space based on Stout's theory of essential independence (see Stout, 1987, 1990; Nandakumar, 1991a for theoretical details). The Stout, et al. computer program, DIMTEST, tests the null hypothesis that the off-diagonal elements of a residual covariance matrix over all item pairs average zero. Stout and Nandakumar have demonstrated its utility for testing if a set of item responses is approximately unidimensional with respect to a normal ability distribution. Nandakumar and Yu (1994) demonstrated the utility of the Stout T, precursor to the Stout - Nandakumar T statistic, for testing if a set of item responses is approximately unidimensional with respect to normal, bimodal, "positive" chi-square, "negative" chi-square, positively skewed, and negatively skewed ability distributions. McLoughlin (2000a, 2000b) showed the Stout - Nandakumar T does not perform as well as the DeChamplain - Gessaroli approximate $\chi^2$ statistic for testing if a set of item responses is approximately unidimensional with respect to normal, uniform, bimodal, "positive" chi-square, "negative" chi-square, positively skewed, and negatively skewed ability distributions for short to moderate length tests.

Hambleton and Swaminathan (1985) suggest that IRT is based on two axioms:
1) performance of an examinee on a test can be explained or predicted by a set of factors called abilities or latent traits; and, 2) the relationship between item performance and ability can be described by a monotonically increasing function between ability parameters and item parameters (or characteristics) called an item response function (IRF). An IRF is approximated by an item response model (IRM). Suppes and Zanotti (1981) showed that the monotonicity assumption cannot be ignored, but can be relaxed such that other IRT models can use monotonically non-decreasing functions.

The common assumptions made for most IRMs are invariance, local independence, and unidimensionality (Hambleton & Swaminathan, 1985). These assumptions are considered so important they are treated in many cases as postulates. Therefore, henceforth, if a reference to unidimensionality is made such that it is referencing the assumption of unidimensionality and

approximations to unidimensionality the term axiomatic unidimensionality will be used synonymously with the IRT assumption of unidimensionality.

There is not a test for invariance, but it has been shown that the properties of local independence and unidimensionality are testable. Invariance implies that characteristics of an examinee do not depend on the set of items used nor do the item parameters depend on the ability distribution of examinees sampled. Local independence implies that when the ability (abilities are) is held constant, an examinee's performance on any pair of items is statistically independent.

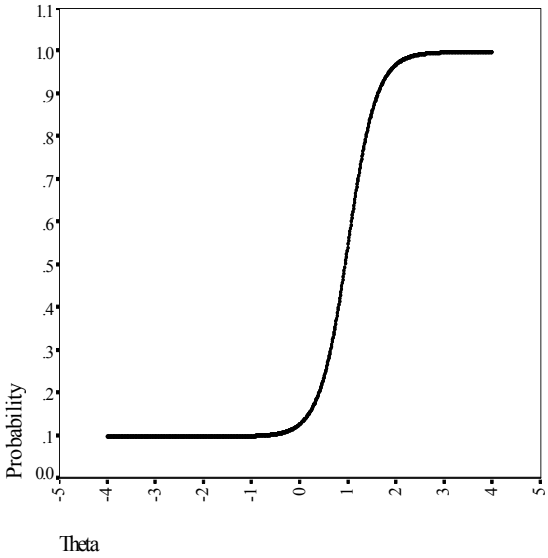
The normal ogive model for unidimensional (d = 1) item response theory (UIRT) assumes local independence and a normal continuous density function for a response function for the item $U_j$ such that

$$P(U_j = 1 | a_i, \quad b_i, \quad c_i, \quad \theta_j) = c_i + (1 - c_i) \int_{-\infty}^{a_i(\theta_j - b_i)} \frac{e^{-z^2/2}}{\sqrt{2\pi}} dz \qquad (1)$$

is the probability that a randomly chosen $j^{th}$ examinee with ability $\theta_j$ answers item i correctly, $a_i$ is the item discrimination parameter, $b_i$ is the item difficulty parameter, and $c_i$ is the item pseudo-guessing parameter [see figure 1].

*Figure 1*

*IRF with a = 2, b = 1, and c = .1*




So, a test $\mathbf{U}_N$ composed of items $U_1, U_2, \ldots, U_N$ with monotone item response functions $P_i(\theta)$ where $U_i = 1$ denotes a correct response and $U_i = 0$ denotes an incorrect response has a test

response distribution defined by $P_N(\mathbf{u}_N) = P(\mathbf{U}_N = \mathbf{u}_N)$, $\mathbf{u} \in \mathbf{U}$ such that $\mathbf{U}$ is the space of all possible test response patterns. Therefore, the fundamental general form equation if an IRT model for $\mathbf{U}_N$ is

$$P(\mathbf{U}_N = \mathbf{u}_N) = \int_{\theta} [P(U_N = u_N \mid \Theta = \theta] f(\theta) d\theta \qquad (2)$$

$\forall$ u э $\Theta$ is a continuous random variable with density $f(\Theta = \theta)$ which is shortened to $f(\theta)$ (Junker, 1991).

The unidimensional normal ogive model does not lend itself to mathematically simple computation; hence, a more tractable set of IRT models are the unidimensional item response theory logistic models (Birnbaum, 1968), specified as:

$$P(U_j = 1 \mid a_i, \quad b_i, \quad c_i, \quad \theta_j) = c_i + (1 - c_i) \frac{e^{Da_i(\theta_j - b_i)}}{1 + e^{Da_i(\theta_j - b_i)}}, \; i = 1, 2, \ldots n \wedge n \in \mathbb{N} \qquad (3)$$

where n is the number of items on the test, $P(U_j = 1 \mid a_i, \quad b_i, \quad c_i, \quad \theta_j)$ is the probability that a randomly chosen $j^{th}$ examinee with ability $\theta_j$ answers item i correctly, $a_i$ is the item discrimination parameter, $b_i$ is the item difficulty parameter, and $c_i$ is the item pseudo-guessing parameter, and, D is a scaling constant. $a_i \in (0, \infty)$. $b_i \in (-\infty, \infty)$. $c_i \in [0,1]$.

Suppose $L(1.702x)$ denotes a logistic continuous density function with D = 1.702 and $N(x)$ denotes a normal ogive continuous density function, Haley (1952) proved that $| N(x) - L(1.702x) | < .01$ for $x \in (-\infty, \infty)$.

When d > 1, many logistic models have been proposed. In the case of d > 1, denote such IRT models as multidimensional item response theory (MIRT) models. It is noteworthy that MIRT models reduce to the Birnbaum UIRT model when d = 1.

When the complete latent space has been specified, local independence holds. When unidimensionality is true, then local independence is true. However, the converse is not true; local independence does not imply unidimensionality. Note:

Theorem 1: Unidimensionality implies local independence (Lord & Novick, 1968, theorem 16.8.1).

Therefore, by contraposition, it is logically equivalent that:

Corollary 1: Not locally independent implies not unidimensional.

If the data were truly unidimensional, then local independence would hold (recall theorem 1). Therefore, lack of unidimensionality should be investigated by considering lack of local independence. Thus, an investigation of the patterns of local dependence in the conditional covariances (covariances of item pairs conditioned on a unidimensional model and estimates of ability) would yield sufficient evidence of a lack of unidimensionality (an application of corollary 1).

The partially compensatory multidimensional item response theory (PCMIRT) (Sympson, 1978) m-dimensional models are specified as:

$$P(U_{ij}=1|\mathbf{a}_i,\quad \mathbf{b}_i,\quad c_i,\quad \boldsymbol{\theta}_j)=c_i+\frac{1-c_i}{\prod_{k=1}^{m}(1+e^{(-1.702a_{ik}(\theta_{jk}-b_{ik}))})},\ i=1,2,\ldots,n\ ;\ n\in\mathbb{N} \quad (4)$$

where n is the number of items on the test, $P(U_{ij}=1|\mathbf{a}_i,\quad \mathbf{b}_i,\quad c_i,\quad \boldsymbol{\theta}_j)$ is the probability that a randomly chosen $j^{th}$ examinee with ability $\boldsymbol{\theta}_j$ answers item i correctly, $\boldsymbol{\theta}_{jk}$ is the ability parameter for person j on dimension k, $a_{ik}$ is the discrimination parameter for item i on dimension k, $b_{ik}$ is the item difficulty parameter for item i on dimension k , $c_i$ is the pseudo-guessing parameter for item i. For the case where there are two dimensions the model reduces to:

$$P(U_{ij}=1|\mathbf{a}_i,\quad \mathbf{b}_i,\quad c_i,\quad \theta_1,\quad \theta_2)=c_i+\frac{1-c_i}{(1+e^{(-1.702a_{1i}(\theta_1-b_{1i}))})(1+e^{(-1.702a_{1i}(\theta_1-b_{1i}))})},$$

$$i=1,2,\ldots,n\ ;\ n\in\mathbb{N} \quad (5)$$

where n is the number of items on the test, $\theta_{j1}$ is the $j^{th}$ examinee's ability on dimension one, $\theta_{j2}$ is the $j^{th}$ examinee's ability on dimension two, P is the probability that a randomly chosen $j^{th}$ examinee with ability $(\theta_1, \theta_2)$ answers item i correctly, $a_{i1}$ is the discrimination parameter for

dimension one for item i, $a_{i2}$ is the discrimination parameter for dimension two for item i, $b_{i1}$ is the difficulty parameter for dimension one for item i, $b_{i2}$ is the difficulty parameter for dimension two for item i, and $c_i$ is the pseudo-guessing parameter for item i.

The compensatory multidimensional item response theory (COMIRT) m-dimensional (McKinley & Reckase, 1983) models are specified as:

$$P(U_{ij}=1\,|\,\mathbf{a}_i,\quad c_i,\quad \mathbf{b}_i,\quad \theta_j) = c_i + \frac{1-c_i}{1+e^{(-1.702\sum_{k=1}^{m} a_{ik}(\theta_{jk}-b_{ik}))}},\ i=1,2,\ldots,n\ ;\ n\in\mathbb{N} \quad (6)$$

$\theta_{j1}$ is the $j^{th}$ examinee's ability on dimension one, $\theta_{j2}$ is the $j^{th}$ examinee's ability on dimension two, . . . , $\theta_{jm}$ is the $j^{th}$ examinee's ability on dimension m, P is the probability that a randomly chosen $j^{th}$ examinee with ability ($\theta_1$, $\theta_2$, …, $\theta_m$) answers item i correctly, $a_{i1}$ is the discrimination parameter for dimension one for item i, $a_{i2}$ is the discrimination parameter for dimension two for item i, . . . , $a_{im}$ is the discrimination parameter for dimension m for item i, $b_{i1}$ is the difficulty parameter for dimension one for item i, $b_{i2}$ is the difficulty parameter for dimension two for item i, . . . , $b_{im}$ is the difficulty parameter for dimension m for item i, and $c_i$ is the pseudo-guessing parameter for item i.

For the case where there are two dimensions the model reduces to:

$$P(U_{ij}=1\,|\,\mathbf{a}_i,\quad c_i,\quad \mathbf{b}_i,\quad \theta_1,\quad \theta_2) = c_i + \frac{1-c_i}{(1+e^{((-1.702a_{1i}(\theta_1-b_{1i}))+(-1.702a_{2i}(\theta_2-b_{2i})))})}$$

$$i=1,2,\ldots,n\ ;\ n\in\mathbb{N} \quad (7)$$

where n is the number of items on the test, $\theta_{j1}$ is the $j^{th}$ examinee's ability on dimension one, $\theta_{j2}$ is the $j^{th}$ examinee's ability on dimension two, P is the probability that a randomly chosen $j^{th}$

examinee with ability ($\theta_1$, $\theta_2$) answers item i correctly, $a_{i1}$ is the discrimination parameter for dimension one for item i, $a_{i2}$ is the discrimination parameter for dimension two for item i, $b_{i1}$ is the difficulty parameter for dimension one for item i, $b_{i2}$ is the difficulty parameter for dimension two for item i, and $c_i$ is the pseudo-guessing parameter for item i.

Much attention has focused on the usage of non-linear factor analysis (NLFA) to model IRT data based on the results of McDonald. McDonald (1967) and Takane & de Leeuw (1987) have shown that IRT models can be considered to be special cases of NLFA. Therefore, dimensionality assessment of IRMs proceeding from NLFA is logically justifiable. Work in this area attempted to assess full dimensionality (Muraki & Engelhart, 1985; Dorans & Lawrence, 1988) or to assess lack of unidimensionality (Yen, 1984). However, *full* information non-linear factor analysis methods "yield a discrepancy function based on the ratio of the likelihood under the fitted model to the likelihood of the saturated model in which we fit the multidimensional distribution to the empirical frequencies. The discrepancy function so obtained is asymptotically distributed as chi-square with $df = 2^p - t$ where t is the number of parameters in the fitted item response model" (McDonald, 1994, p. 73) and p is the number of items. McDonald notes that in applications difficulty will arise that $2^p$ is large relative to sample size and Gessaroli notes that "require[s] very large samples, even with fairly small test lengths because this method uses all of the information in all of the $2^p$ response patterns" (Gessaroli, 1994, p. 102). McDonald condensed the problem with full information factor analysis by stating, "that the full information contained in the higher joint moments of the binary responses suffer the usual sampling instability of such moments. In general, if additional information is not contributed by chance, we would expect it to increase misfit, not reduce it" (McDonald, 1994, p. 75).

So, attention expanded such that focus was on *limited*-information NLFA procedures using non-parametric procedures to assess full dimensionality or lack of unidimensionality (Hambleton & Rovinelli, 1986; DeChamplain & Tang, 1993; Berger & Knol, 1990; DeChamplain & Gessaroli, 1991; McDonald, 1994; McDonald & Mok, 1995; Gessaroli & DeChamplain, 1996). Limited information non-linear factor analysis methods proposed by Christofferson (1975) and Muthén (1978) yield a discrepancy function based on the ratio of the likelihood under the fitted model to the likelihood of the model with the first and second moments. The discrepancy function

so obtained is asymptotically distributed as chi-square with $df = \frac{p(p+1)}{2} - t$. "Large samples are not needed, do not require pooling of answer-pattern frequencies, and arguably should be computationally more efficient" (McDonald, 1994, p. 76).

A method advanced using NLFA to assess lack of unidimensionality is the approximate chi-square test of DeChamplain & Gessaroli (Gessaroli & DeChamplain, 1996) which seems to be a promising method which uses NLFA. The statistic tests the null hypothesis that the off diagonal elements of the matrix of residual correlations of a NLFA are equal to zero after fitting the data with a one-dimensional NLFA model. The DeChamplain - Gessaroli method uses the approximate $\chi^2$ statistic originally proposed by Bartlett (1950) and outlined by Steiger (1980).

First, using NOHARM87 (Fraser, 1998), a one dimensional model is fitted using NLFA. For each pair of items the proportion of examinees who correctly answered only item i, only item j, and both items is calculated. Let these proportions be respectively designated $p_i^{(0)}$, $p_j^{(0)}$, and $p_{ji}^{(0)}$. Second, for each pair of items the expected as well as residual joint proportions of examinees who correctly answered items i and j are estimated using the one dimensional model. Let the residual joint proportions be designated $p_{ji}^{(r)}$. Third, the estimated residual correlations for each pair of items is calculated. Let the estimated residual correlations be designated $r_{ji}^{(r)}$ such that

$$r_{ji}^{(r)} = \frac{p_{ji}^{(r)}}{\sqrt{p_i^{(0)}(1-p_i^{(0)})p_j^{(0)}(1-p_j^{(0)})}}. \qquad (8)$$

Fourth, each of the estimated residual correlations is transformed to a Fisher z. Let the transformed estimated residual correlations be designated $z_{ji}^{(r)}$ such that

$$z_{ji}^{(r)} = .5\ln(1+r_{ji}^{(r)}) - .5\ln(1-r_{ji}^{(r)}). \qquad (9)$$

Fifth, the Bartlett $\chi^2$ statistic is calculated such that

$$\chi^2 = (N-3)\sum_{i=2}^{k}\sum_{j=1}^{i-1}(z_{ji}^{(r)})^2 \qquad (10)$$

where N is the number of examinees and k is the number of items. The DeChamplain - Gessaroli $\chi^2$ is approximately distributed as a central $\chi^2$ with $df = .5k(k-1) - t$ where t is the total number

of independent factor analytic parameters estimated (Gessaroli & DeChamplain, 1996). The decision rule is reject $H_0$ if $\chi^2_{comp} \geq \chi^2_{((.5k(k-1)-t),\alpha)}$ where $\chi^2_{((.5k(k-1)-t),\alpha)}$ is the upper 100(1 -$\alpha$) percentile of the $\chi^2_{(.5k(k-1)-t)}$ distribution and $\alpha$ the desired level of significance.

Monte Carlo simulations executed and reported have varied test length and sample size; and to a lesser extent IRF type, correlations between abilities, and the effect of test reliability (DeChamplain, 1995; DeChamplain, 1996; DeChamplain & Gessaroli, 1996; Gessaroli & DeChamplain, 1996; Gessaroli, DeChamplain, & Folske, 1997; McLoughlin, 2000a, 2000b). Item-ability loadings have generally been simple structure with a ratio of 80:20, 75:25, or 50:50.

One empirical study has been conducted on DGACS using real data (DeChamplain, 1995) and it was shown that the DGACS confirmed established dimensionality structure for a battery of the LSAT. Two chi-squared statistics which were precursors to the DGACS were shown to be robust to mild departure from the ability distribution assumption of $N(\mu, \sigma)$ (DeChamplain & Tang, 1993). The DGACS was shown to be robust to departure from the ability distribution assumption of $N(\mu, \sigma)$ for short or moderate length tests (McLoughlin, 2000a, 2000b). One empirical study has been conducted on DGACS that showed it can discriminate between one and more than one dimension underlying the test and maintains a good type II error rate even when correlations between two abilities are as high as 0.7 (DeChamplain & Gessaroli, 1996). No studies have been conducted for different item-ability loadings other than simple structure for the compensatory multidimensional item response theory (COMIRT) d = 2 case. It has been found to be as robust to the presence of guessing as the Stout - Nandakumar T (DeChamplain & Gessaroli, 1996).

All the studies modeled multidimensional item response theory (MIRT) data using compensatory multidimensional item response theory two parametre logistic (COMIRT2PL) model or compensatory multidimensional item response theory three parametre logistic (COMIRT3PL) model with but one fixed c. No empirical evidence has been reported on data where $d \geq 3$. It is claimed that the DGACS avoids strong parametric modeling assumptions, uses a model to which IRT is related (NLFA) in assessing dimensionality, does not suffer the problem of the Stout - Nandakumar T method of sensitivity to a small number of examinees or items.

Advantages claimed to this methodology are: 1) dimensionality of a test can be investigated rather than just unidimensionality (the logical extension of the previous discussion substituting the NOHARM87 step with an m-dimensional model rather than a 1-dimensional model); 2) it avoids

strong parametric modeling assumptions; 3) uses a model to which IRT is related (NLFA) in assessing dimensionality; 4) the approximate $\chi^2$ has the advantage of being able to test the dimensionality of long tests using small samples; 5) dimensionality of short tests can be investigated; 6) the statistic is based on a discrepancy function (the discrepancy between the observed and fitted item-covariance matrices); and, 7) is contained in two available programs, NOHARM87 and CHIDIM.

Weaknesses noted for this methodology are: 1) the results used to calculate the $\chi^2$ statistic are based on unweighted ordinary least square estimates (from NOHARM87) which are less preferred theoretically than unweighted generalized least square estimates; 2) it is conservative; and, 3) the statistic possibly suffers as does any other $\chi^2$ statistic in that it is sensitive to large sample size. If the number of examinees is sufficiently large rejection of the null hypothesis that the off diagonal elements of the matrix of residual correlations of a NLFA are equal to zero might almost be guaranteed.

The Stout - Nandakumar T statistic was developed to assess lack of axiomatic unidimensionality under a null hypothesis for essential unidimensionality; that is, the null hypothesis that the off-diagonal elements of a residual covariance matrix over all item pairs average zero. The Stout - Nandakumar T statistic (SNTS) is designed to test the hypothesis on a test of N items such that the test is divided into three subtests: the assessment test 1 (AT1), the assessment test 2 (AT2), and the partitioning test (PT).

Let $N_1$ be the number of items in AT1, $N_2$ be the number of items in AT2; hence, $N - N_1 - N_2$ is the number of items in PT. Restrict N such that $N \geq 20$ (Stout, Douglas, Junker, & Roussos, 1993). Further, restrict $\max(N_1, N_2) \leq \frac{N}{4}$ and $\min(N_1, N_2) \geq 4$ (Gao, 1997). Let M be the number of examinees. Restrict M such that $M \geq 250$ (Nandakumar & Stout, 1993).

AT1 is selected such that it is dimensionally homogeneous. It has been suggested by Stout and others that the selection of AT1 should be based on 1) hierarchical cluster analysis (HCA), 2) dimensionality evaluation to enumerate contributing traits (DETECT), 3) expert opinion based on content, or 4) principle component factor analysis (PCFA). There has been only one investigation of the HCA method for choosing AT1 (Roussos, Stout, & Marsden, 1993), one investigation of the content expert opinion method (Nandakumar, 1993), and no investigation of the DETECT method.

Using PCFA to select AT1 the items selected should be the same signed most heavily loading items on the second factor (Stout, 1990; Stout & Nandakumar, 1993; and Nandakumar, 1994). It should be noted (Nandakumar, 1994) that PCFA is a data selection procedure and is not part of the distribution theory of the Stout T.

AT2 is selected to correct for a "slight" bias in the Stout T statistic such that AT2 items are matched in difficulty with AT1 items. PT items are all the items left not assigned to AT1 or AT2.

Scores on the PT subtest are used to group the M examinees into K mutually exclusive and exhaustive subgroups and within each $k^{th}$ subgroup (k = 1, 2, 3, . . . , K) two variance estimates ($\hat{\sigma}_k^2$ and $\hat{\sigma}_{U,k}^2$) are computed using item responses of AT1. Let $U_{ijk}$ be the response of the $j^{th}$ examinee to the $i^{th}$ item from the $k^{th}$ group. Let $J_k$ denote the number of examinees in the $k^{th}$ group. Let $Y_j^{(k)} = \sum_{i=1}^{N_1} \frac{U_{ijk}}{N_1}$. Let $\overline{Y}^{(k)} = \sum_{j=1}^{J_k} \frac{Y_j^{(k)}}{J_k}$. Thus,

$$\hat{\sigma}_k^2 = \sum_{j=1}^{J_k} \frac{(Y_j^{(k)} - \overline{Y}^{(k)})^2}{J_k}. \tag{11}$$

Let $\hat{p}_i^{(k)} = \sum_{i=1}^{J_k} \frac{U_{ijk}}{J_k}$. Thus, $\hat{\sigma}_{U,k}^2 = \sum_{i=1}^{N_1} \frac{\hat{p}_i^{(k)}(1-\hat{p}_i^{(k)})}{N_1^2}$. (12)

$\hat{\sigma}_k^2$ is the variance estimate of the AT1 subtest amongst examinees in subgroup k and $\hat{\sigma}_{U,k}^2$ is the estimate of the 'unidimensional' variance by summing the item variances of AT1.

Let $\hat{\delta}_{4,k} = \sum_{i=1}^{N_1} [\hat{p}_i^{(k)}(1-\hat{p}_i^{(k)})(1-2\hat{p}_i^{(k)})^2]$ and $\hat{\mu}_{4,k} = \sum_{j=1}^{J_k} \frac{(Y_j^{(k)} - \overline{Y}_j^{(k)})^4}{J_k}$. Thus, the appropriate standard error for the $k^{th}$ subgroup is

$$S_k = \sqrt{\frac{(\hat{\mu}_{4,k} - \hat{\sigma}_k^4) + \frac{\hat{\delta}_{4,k}}{N_1^4} + 2\sqrt{\frac{(\hat{\mu}_{4,k} - \hat{\sigma}_k^4)\hat{\delta}_{4,k}}{N_1^4}}}{J_k}}. \tag{13}$$

Hence, $T_1 = \frac{1}{\sqrt{K}} \sum_{k=1}^{K} \frac{\hat{\sigma}_k^2 - \hat{\sigma}_{U,k}^2}{S_k}$. (14)

Two variance estimates ( $\hat{\sigma}^2_{2,k}$ and $\hat{\sigma}^2_{2,U,k}$ ) are computed using item responses of AT2. Let

$Y^{(k)}_{2,j} = \sum_{i=1}^{N_2} \frac{U_{ijk}}{N_2}$. Let $\overline{Y}^{(k)}_2 = \sum_{j=1}^{J_k} \frac{Y^{(k)}_{2,j}}{J_k}$. Thus,

$$\hat{\sigma}^2_k = \sum_{j=1}^{J_k} \frac{(Y^{(k)}_{2,j} - \overline{Y}^{(k)}_2)^2}{J_k}. \quad (15)$$

Let $\hat{p}^{(k)}_{2,i} = \sum_{i=1}^{J_k} \frac{U_{ijk}}{J_k}$. Thus, $\hat{\sigma}^2_{2,U,k} = \sum_{i=1}^{N_2} \frac{\hat{p}^{(k)}_{2,i}(1-\hat{p}^{(k)}_{2,i})}{N_2^2}$. (16)

$\hat{\sigma}^2_{2,k}$ is the variance estimate of the AT2 subtest amongst examinees in subgroup k and $\hat{\sigma}^2_{2,U,k}$ is the estimate of the 'unidimensional' variance by summing the item variances of AT2. Let $\hat{\delta}_{2,4,k}$ $= \sum_{i=1}^{N_2} [\hat{p}^{(k)}_{2,i}(1-\hat{p}^{(k)}_{2,i})(1-2\hat{p}^{(k)}_{2,i})^2]$ and $\hat{\mu}_{2,4,k} = \sum_{j=1}^{J_k} \frac{(Y^{(k)}_{2,j} - \overline{Y}^{(k)}_2)^4}{J_k}$. Thus, the appropriate standard error for the $k^{th}$ subgroup is

$$S_{2,k} = \sqrt{\frac{(\hat{\mu}_{2,4,k} - \hat{\sigma}^4_{2,k}) + \frac{\hat{\delta}_{2,4,k}}{N_2^4} + 2\sqrt{\frac{(\hat{\mu}_{2,4,k} - \hat{\sigma}^4_{2.k})\hat{\delta}_{2,4,k}}{N_2^4}}}{J_k}}. \quad (17)$$

$$\text{Hence, } T_2 = \frac{1}{\sqrt{K}} \sum_{k=1}^{K} \frac{\hat{\sigma}^2_{2,k} - \hat{\sigma}^2_{2,U,k}}{S_{2,k}}. \quad (18)$$

Finally, the Stout - Nandakumar T for testing essential unidimensionality is

$$T = \frac{T_1 - T_2}{\sqrt{2}}. \quad (19)$$

T is asymptotically distributed $N(0,1)$ under $H_0$. The decision rule is reject $H_0$ if $T \geq |Z_{\alpha/2}|$ where $Z_{\alpha/2}$ is the upper $100(1 - \frac{\alpha}{2})$ percentile of the normal distribution and $\alpha$ the desired level of significance.

Originally, Stout (Stout, 1987) proposed only choosing an AT1 and PT (thus, the test for unidimensionality was the $T_1$ statistic), but it was shown that for short tests ($N \leq 30$) type I error

rates were inflated (Stout, 1990; Gessaroli & DeChamplain, 1991) and that if there is much guessing and high discrimination (a > 1.1) the statistic is biased toward rejection of the null hypothesis (Stout & Nandakumar, 1993). The introduction of AT2 and the $T_2$ statistic was added to account for the bias in $T_1$ by Nandakumar in her doctoral dissertation (1987). Nandakumar & Stout (1993) noted that $T_1$ is sensitive to dimensionality and bias; whereas, $T_2$ is sensitive to bias. Consequently, T is sensitive to dimensionality.

Stout's T statistic (STS) and the Stout - Nandakumar T statistic (SNTS) have been studied extensively. Monte Carlo simulations executed and reported have varied test length, sample size, IRF type, correlations between abilities, ability distributions, and item-ability loadings (Stout, 1987, 1990; Gessaroli & DeChamplain, 1991; Junker & Stout, 1993; Nandakumar, 1991a, 1991b, 1993; Nandakumar & Yu, 1994, 1996; DeChamplain & Gessaroli, 1996; Roussos, Stout, & Marsden, 1993, 1998; Wang, 1994; Hattie, Krakowski, Rogers, & Swaminathan, 1996).

Empirical studies have been conducted on STS using real data (Nandakumar, 1993; Wang & Hocevar, 1994) and it was shown that STS established dimensionality structure for several standardised tests whilst confirming multidimensional structure of tests composed of half items from one standardized test and half items from another dissimilar test based on content. STS has been shown to be robust to departure from the ability distribution assumption of $N(\mu, \sigma)$ (Nandakumar & Yu, 1994, 1996), can discriminate between one and more than one dimension underlying the test, maintains the desired type I error rate even when correlations between two abilities are as high as 0.7 (Stout, 1987; Ang & Miller, 1993), and maintains adequate power when correlations between two abilities are as high as 0.7 (Nandakumar & Yu, 1994, 1996). It has been shown that the SNTS is not as robust as the STS to departure from the ability distribution assumption of $N(\mu, \sigma)$ (McLoughlin 2000a). For both the STS and SNTS, it has been shown that different item-ability loadings for the d = 2 case (simple structure, some simple and mixed structure) does not diminish power or inflate Type I error rates (Stout, 1987; Roussos, Stout, & Marsden, 1998). The SNTS has been found to be robust to the presence of guessing or common discrimination (Hattie, Krakowski, Rogers, & Swaminathan, 1996).

All the studies but one (Hattie, et al., 1996) modeled the MIRT data using COMIRT2PL. Hattie, et al. found it did not perform well with partially compensatory multidimensional two parametre logistic (PCMIRT2PL) data. Little empirical evidence has been reported on data where

d = 3 (only two: Hattie, et al., 1996 and Roussos, et al., 1998) and in those reports there were severe restrictions placed on the choices of the item parameters. No empirical evidence has been reported on data where d > 3. It is claimed that the SNTS avoids strong parametric modeling assumptions, but it is sensitive to sample size of examinees or items. Research indicates if the number of examinees is small ($M \leq 750$) or the number of item is small ($N \leq 25$) partitioning of the test into AT1, AT2, and PT yield partitions too small to compute T reliably (Nandakumar, 1991b; DeChamplain & Gessaroli, 1996).

Advantages claimed to this methodology are: 1) it avoids strong parametric modeling assumptions; 2) the SNT has the advantage of being able to test the dimensionality of long tests or large samples; 3) has been very extensively studied; and, 4) is contained in a commercially available program, DIMTEST.

Weaknesses noted for this methodology are: 1) the results are dependent on the choice of AT1; 2) it is conservative; and, 3) the statistic suffers that it is sensitive to small sample size of items or examinees.

It has been claimed that when item responses are modeled properly by an essentially unidimensional model, items of all the subtests AT1, AT2, and PT will be measuring the same dominant dimension; hence, T will be "small." However, when item responses are modeled improperly by an essentially unidimensional model, items of AT1 will be dimensionally different from items in the other subtests (AT2 and PT); hence, T will be "large."

The two most promising methods are the Gessaroli - DeChamplain approximate $\chi^2$ test based on limited-information non-linear factor analytic procedures (NLFA) which was first proposed by McDonald in 1967 and the Stout - Nandakumar T test based on Stout's theory of essential unidimensionality which was first proposed by Stout in 1987.

Stout (1990) notes that a psychometric interpretation of essential independence is that $\Theta$ measures individual examinee differences that are essential or dominant in influencing item pool performance. Nandakumar (1994) notes that essential independence requires the average of $|\text{cov}(U_i, U_j \mid \Theta = \theta)|$ over all item pairs to be 'small' when conditioned on a dominant ability; whereas, local independence requires $\text{cov}(U_i, U_j \mid \Theta = \theta) = 0$ for all $\theta$. Therefore, Nandakumar claims it is the case that essential independence conditions on dominant abilities; but, local independence (strong and weak forms) requires conditioning on all abilities. A key detail is for

$\{U_i, i \geq 1, \ i \in \mathbb{N}\}$ , it may be $d > 1$, but with $d_E = 1$, the claim is that UIRT modeling will be adequate. Therefore, the axiomatic dimensionality may be greater than one, but the essential dimensionality is equal to one.

Several researchers have argued that test data will almost always be multidimensional (Lumsden, 1961; Reckase, Ackerman, & Carlson, 1988; Wang, 1986, 1988; Yen, 1985). Traub (1983), Humphreys (1984), and Harrison (1986) made compelling arguments that the UIRT assumption of unidimensionality will *always* be violated with real data, then why search for an acceptable UIRT model. Why not model all IRT data with MIRT models?

First, one can argue that a goal of modeling is parsimony. Second, since UIRT is more easily interpretable than MIRT, it follows for many testing situations an approximately unidimensional test would be preferable to a multidimensional test. Third, several studies have shown that UIRT parameters are recovered well with MIRT data when the inter-dimensional correlations are moderate to high (Drasgow & Parsons, 1983; Harrison, 1986). Fourth, it has been shown that the unidimensionality assumption is met when a set of items from MIRT measures the same linear composite of multidimensional abilities (Reckase, Ackerman, & Carlson, 1988). Fifth, "the practice of reporting a single score, often number correct, is a widespread practice for all sorts of test administrations. Implicit in this is the conclusion that there is one dominant dimension driving the test response mechanism, at least up to a good approximation" (Junker & Stout, 1994, p. 51).

Indeed, suppose a test practitioner has created a mathematics test with subtests defined by the contents algebra, geometry, and pre-calculus. Suppose the test is an achievement battery designed to test mastery of the three areas. One justifiably uses expert opinion to conclude that three scores should be used. However, suppose the test is a placement instrument designed for placement in remedial mathematics or a first course in calculus for entering freshmen at the college level. One may use expert opinion to conclude that one score should be used. Nonetheless, this is an example of Junker and Stout's position that in reporting a single score the conclusion that there is one dominant dimension driving the test response mechanism is implicit. Ergo, a statistical procedure such as DIMTEST or NOHARM-CHIDIM assists in justifying the use of UIRT analysis and report of a single score.

Therefore, the purpose of this study was to investigate the effect type of bivariate multidimensional model has on the power of the DGACS and SNTS. Specifically, two multidimensional models are considered: the Sympson (1978) partially compensatory multidimensional IRT (PCMIRT) model and the McKinley - Reckase (1983) compensatory multidimensional IRT (COMIRT) model. The intent was realized by 1) an empirical investigation of the power of the McDonald - Fraser NOHARM procedure followed by the DeChamplain - Gessaroli CHIDIM procedure for testing if a set of item responses is approximately unidimensional with respect to the Sympson PCMIRT and McKinley - Reckase COMIRT models ; 2) an empirical investigation of the power of the Stout - Nandakumar DIMTEST procedure for testing if a set of item responses is approximately unidimensional with respect to the Sympson PCMIRT and McKinley - Reckase COMIRT models; 3) an empirical investigation of the power of the DGACS and SNTS for PCMIRT data; and, 4) an empirical investigation of the power of the DGACS and SNTS for COMIRT data.

This research extends previous results by varying test length, sample size, and considering a more diverse set of parameter combinations. In addition, this research investigates the comparative utility of DGACS and SNTS for testing if a set of item responses is approximately unidimensional with respect to the Sympson PCMIRT model which heretofore had not been executed.

**METHODS**

The research uses Monte Carlo methods as outlined by Harwell, Stone, Hsu, & Kirisci (1996) and Spence (1983).

The DGACS and SNTS were tested with four factors manipulated: model, test length, sample size, and correlation between dimensions. Two models were used, the bivariate Sympson PCMIRT and bivariate Reckase - McKinley COMIRT models.

Three test lengths (J) were used to model a range from short to moderate lengths of tests: 20, 40, and 60. Three sample sizes (N) were used to model a range from small to moderately large samples: 250, 750, and 2000. Three correlations between dimensions were used to model a range from no to strong correlations between dimensions, $\rho = 0, .4, .8$. The $2 \times 3 \times 3 \times 3$ design data sets in each cell were replicated 100 times.

For each sample size, test length, and correlation examinee responses were generated using SPSS 8.0 (SPSS Institute, 1998). The number of dimensions was set to two. PCMIRT model responses were generated such that $a_{2i}$ and $a_{1i} \in \{.5, 1, 1.5, 2\}$; $b_{2i}$ and $b_{1i} \in \{-2, -1, 0, 1, 2\}$; $c_i \in \{0, .05, .1, .15, .2\}$ where 80% of the items had $c_i = 0$, 5% of the items had $c_i = .05$, 5% of the items had $c_i = .1$, 5% of the items had $c_i = .15$, and 5% of the items had $c_i = .2$. COMIRT model responses were generated likewise.

For a given $(\theta_1, \theta_2)$ level and a given $i^{th}$ item, the probability of a correct response was obtained, $P_i(\theta_1, \theta_2)$. Then a random number, x, from $U[0,1]$ was generated. If $P_i(\theta_1, \theta_2) \geq x$, the examinee was said to have answered the item correctly; whereas, if $P_i(\theta_1, \theta_2) < x$, the examinee was said to have answered the item incorrectly. Correct responses were coded as 1 and incorrect as 0.

The computer program NOHARM87 (Fraser, 1998) was run on each data set with the following specifications: an exploratory solution was requested, an axiomatic unidimensional model was fitted, criterion for convergence was set at the default value of 0.0001; and, the pseudo-guessing parameters were set at the default value of zero. Upon completion of the NOHARM87 execution, CHIDIM (DeChamplain & Tang, 1997a) was run with $\alpha = .05$. Rejection rates were recorded.

Then, the computer program DIMTEST (Stout, et al. 1992a) was run on each data set with the following specifications: an exploratory solution was requested, the default principal components factor analysis with half the number of subjects and one-quarter of the items for computation of the tetrachoric correlation matrix for selection of assessment test one (AT1), default selection of assessment test two (AT2) with half the number of subjects and one-quarter of the items for computation, default selection of the partitioning set (PT) with the same half of subjects and the remaining one-half of the items for computation. Default $\alpha = .05$. Rejection rates were recorded.

Finally, rejection rates were compared between the DGACS and SNTS for PCMIRT data; between the DGACS and SNTS for COMIRT data; between PCMIRT and COMIRT data for the SNTS; and between PCMIRT and COMIRT data for the DGACS.

Justification for the use of exploratory solutions in each case is based on practical considerations. If a researcher is attempting to determine if a data set is approximately unidimensional that attempt is exploratory in nature. There may well be prior information about

the pseudo-guessing parameters, but the researcher wished to approach the question such that such prior knowledge might not exist; hence, setting the pseudo-guessing parameters to zero in the NOHARM87 stage of the analysis. Allowing for expert opinion to select the items for use in AT1 may well be practical in some situations, but in many it is not - one is trying to determine *if* a unidimensional model may be used or not; hence, setting DIMTEST to perform the principle components factor analysis was thought to be most common.

**RESULTS**

Rejection rates were recorded and compared for the DGACS and SNTS. In addition, the agreement of cell case number between run rejections was examined per cell between the DGACS and SNTS. The agreement of cell case across run rejections was examined separately for the DGACS and SNTS. Agreement above 10% was considered to be possibly an artifact of the data construction procedure. No agreement index above 10% was found.

For all the tables in the results section, let D - G abbreviate DeChamplain - Gessaroli, S - N abbreviate Stout - Nandakumar, j represent the number of items, n represent the number of subjects, $d_E$ represents essential unidimensionality, and $d_F$ represents fundamental unidimensionality (see McLoughlin, 2000a for details).

The number of rejections per 100 runs of approximate unidimensionality for data modeled with the Sympson PCMIRT model are summarized in table 1. The number of rejections per 100 runs of approximate unidimensionality for data modeled with the McKinley - Reckase COMIRT model are summarized in table 2. The number of rejections per 100 runs of approximate unidimensionality for the DGACS are summarized in table 3. The number of rejections per 100 runs of approximate unidimensionality for the SNTS are summarized in table 4.

The DGACS demonstrated excellent power for data derived from a Sympson PCMIRT model in 21 of 27 conditions. The exceptions were: the 20 item, 250 subject, 0 correlation condition; the 40 item, 750 subject, .4 correlation condition; the 40 item, 250 subject, .8 correlation condition; the 60 item, 250 subject, .8 correlation condition; and, the 20 item, 750 subject, .8 correlation condition where the DGACS demonstrated fair power; and, the 20 item, 250 subject, .4 correlation condition and the 20 item, 250 subject, .8 correlation condition where the DGACS demonstrated poor power.

The DGACS demonstrated excellent power for data derived from a McKinley - Reckase COMIRT model in 21 of 27 conditions. The exceptions were the 250 subject or 750 subject, .8 correlation condition regardless of test length where the DGACS demonstrated poor power.

The SNTS demonstrated excellent power for data derived from a Sympson PCMIRT model in 2 of 27 conditions. The cases where excellent power was demonstrated were: the 60 item, 750 subject, 0 correlation condition; and, the 60 item, 2,000 subject, 0 correlation condition. The SNTS demonstrated fair power in three other conditions: the 60 item, 250 subject, 0 correlation condition; the 40 item, 750 subject, 0 correlation condition; and, the 40 item, 2,000 subject, 0 correlation condition. In the other 22 conditions the SNTS demonstrated poor power to detect the multidimensional nature of the data.

The SNTS demonstrated excellent power for data derived from a McKinley - Reckase COMIRT model in 9 of 27 conditions. All the conditions where excellent power existed were in conditions such that the number of items were 40 or 60 and the correlation between dimensions was 0 or .4 (see table 4). The SNTS demonstrated fair power in three other conditions: the 40 item, 250 subject, 0 correlation condition; the 20 item, 2,000 subject, 0 correlation condition; and, the 40 item, 250 subject, .4 correlation condition. In the other 15 conditions the SNTS demonstrated poor power to detect the multidimensional nature of the data.

The DGACS, nonetheless, demonstrated poor power in six conditions: the 250 subject and 750 subject $\rho = .8$ conditions (regardless of test length). The DGACS demonstrated excellent power in all other conditions.

The SNTS demonstrated poor power in all conditions when $\rho = .8$. In addition, the SNTS demonstrated poor power in the 20 item, 250 subject, $\rho = 0$ condition; the 20 item, 750 subject, $\rho = 0$ condition; the 20 item, 250 subject, $\rho = .4$ condition; and, the 20 item, 750 subject, $\rho = .4$ condition.

The SNTS demonstrated fair power in the 40 item, 250 subject, $\rho = 0$ condition; 40 item, 750 subject, $\rho = 0$; 20 item, 2,000 subject, $\rho = 0$ condition; and, 40 item, 250 subject, $\rho = .4$ condition; and, 20 item, 2,000 subject, $\rho = .4$ condition.

The SNTS demonstrated excellent power in the 60 item, 250 subject, $\rho = 0$ condition; the 60 item, 750 subject, $\rho = 0$ condition; the 40 item, 2,000 subject, $\rho = 0$ condition; the 60 item, 2,000 subject, $\rho = 0$ condition; the 60 item, 250 subject, $\rho = .4$ condition; the 40 item, 750 subject, $\rho = .4$

condition; the 60 item, 750 subject, $\rho = .4$ condition; the 40 item, 2,000 subject, $\rho = .4$ condition; and, the 60 item, 2,000 subject, $\rho = .4$ condition.

*Table 1*

*Number of Rejections of Unidimensionality for the Bivariate Sympson Partially Compensatory Model* (nominal $\alpha = .05$; 100 runs per cell)

| j | n | r | D - G Approx. $\chi^2$ Reject $H_0$: $d_F = 1$ | S - N Approx. T Reject $H_0$: $d_E = 1$ |
|---|---|---|---|---|
| 20 | 250 | 0 | 71 | 9 |
| 40 | 250 | 0 | 100 | 37 |
| 60 | 250 | 0 | 100 | 61 |
| 20 | 750 | 0 | 100 | 18 |
| 40 | 750 | 0 | 100 | 60 |
| 60 | 750 | 0 | 98 | 79 |
| 20 | 2,000 | 0 | 100 | 30 |
| 40 | 2,000 | 0 | 100 | 73 |
| 60 | 2,000 | 0 | 100 | 89 |
| 20 | 250 | .4 | 26 | 8 |
| 40 | 250 | .4 | 71 | 19 |
| 60 | 250 | .4 | 92 | 27 |
| 20 | 750 | .4 | 92 | 15 |
| 40 | 750 | .4 | 100 | 24 |
| 60 | 750 | .4 | 98 | 52 |
| 20 | 2,000 | .4 | 100 | 10 |
| 40 | 2,000 | .4 | 100 | 13 |
| 60 | 2,000 | .4 | 100 | 13 |
| 20 | 250 | .8 | 27 | 7 |
| 40 | 250 | .8 | 67 | 10 |
| 60 | 250 | .8 | 74 | 14 |
| 20 | 750 | .8 | 67 | 10 |
| 40 | 750 | .8 | 95 | 17 |
| 60 | 750 | .8 | 98 | 18 |
| 20 | 2,000 | .8 | 100 | 16 |
| 40 | 2,000 | .8 | 100 | 2 |
| 60 | 2,000 | .8 | 100 | 8 |

*Table 2*

*Number of Rejections of Unidimensionality for the Bivariate McKinley - Reckase Compensatory Model* (nominal $\alpha = .05$; 100 runs per cell)

| j | n | r | D - G Approx. $\chi^2$ Reject $H_0$: $d_F = 1$ | S - N Approx. T Reject $H_0$: $d_E = 1$ |
|---|---|---|---|---|
| 20 | 250 | 0 | 100 | 35 |
| 40 | 250 | 0 | 100 | 74 |
| 60 | 250 | 0 | 100 | 83 |
| 20 | 750 | 0 | 100 | 49 |
| 40 | 750 | 0 | 100 | 77 |
| 60 | 750 | 0 | 100 | 98 |
| 20 | 2,000 | 0 | 100 | 69 |
| 40 | 2,000 | 0 | 100 | 86 |
| 60 | 2,000 | 0 | 100 | 91 |
| 20 | 250 | .4 | 93 | 20 |
| 40 | 250 | .4 | 99 | 66 |
| 60 | 250 | .4 | 100 | 87 |
| 20 | 750 | .4 | 100 | 32 |
| 40 | 750 | .4 | 100 | 85 |
| 60 | 750 | .4 | 99 | 94 |
| 20 | 2,000 | .4 | 100 | 55 |
| 40 | 2,000 | .4 | 98 | 93 |
| 60 | 2,000 | .4 | 100 | 97 |
| 20 | 250 | .8 | 1 | 4 |
| 40 | 250 | .8 | 5 | 13 |
| 60 | 250 | .8 | 2 | 19 |
| 20 | 750 | .8 | 14 | 2 |
| 40 | 750 | .8 | 33 | 16 |
| 60 | 750 | .8 | 48 | 35 |
| 20 | 2,000 | .8 | 100 | 3 |
| 40 | 2,000 | .8 | 97 | 16 |
| 60 | 2,000 | .8 | 99 | 32 |

*Table 3*

*Number of Rejections of Unidimensionality for the DGACS* (nominal $\alpha = .05$; 100 runs per cell)

| j | n | r | Sympson PCMIRT Model | McKinley - Reckase COMIRT Model |
|---|---|---|---|---|
| 20 | 250 | 0 | 71 | 100 |
| 40 | 250 | 0 | 100 | 100 |
| 60 | 250 | 0 | 100 | 100 |
| 20 | 750 | 0 | 100 | 100 |
| 40 | 750 | 0 | 100 | 100 |
| 60 | 750 | 0 | 98 | 100 |
| 20 | 2,000 | 0 | 100 | 100 |
| 40 | 2,000 | 0 | 100 | 100 |
| 60 | 2,000 | 0 | 100 | 100 |
| 20 | 250 | .4 | 26 | 93 |
| 40 | 250 | .4 | 71 | 99 |
| 60 | 250 | .4 | 92 | 100 |
| 20 | 750 | .4 | 92 | 100 |
| 40 | 750 | .4 | 100 | 100 |
| 60 | 750 | .4 | 98 | 99 |
| 20 | 2,000 | .4 | 100 | 100 |
| 40 | 2,000 | .4 | 100 | 98 |
| 60 | 2,000 | .4 | 100 | 100 |
| 20 | 250 | .8 | 27 | 1 |
| 40 | 250 | .8 | 67 | 5 |
| 60 | 250 | .8 | 74 | 2 |
| 20 | 750 | .8 | 67 | 14 |
| 40 | 750 | .8 | 95 | 33 |
| 60 | 750 | .8 | 98 | 48 |
| 20 | 2,000 | .8 | 100 | 100 |
| 40 | 2,000 | .8 | 100 | 97 |
| 60 | 2,000 | .8 | 100 | 99 |

*Table 4*

*Number of Rejections of Unidimensionality for the SNTS* (nominal $\alpha = .05$; 100 runs per cell)

| j | n | r | Sympson PCMIRT Model | McKinley - Reckase COMIRT Model |
|---|---|---|---|---|
| 20 | 250 | 0 | 9 | 35 |
| 40 | 250 | 0 | 37 | 74 |
| 60 | 250 | 0 | 61 | 83 |
| 20 | 750 | 0 | 18 | 49 |
| 40 | 750 | 0 | 60 | 77 |
| 60 | 750 | 0 | 79 | 98 |
| 20 | 2,000 | 0 | 30 | 69 |
| 40 | 2,000 | 0 | 73 | 86 |
| 60 | 2,000 | 0 | 89 | 91 |
| 20 | 250 | .4 | 8 | 20 |
| 40 | 250 | .4 | 19 | 66 |
| 60 | 250 | .4 | 27 | 87 |
| 20 | 750 | .4 | 15 | 32 |
| 40 | 750 | .4 | 24 | 85 |
| 60 | 750 | .4 | 52 | 94 |
| 20 | 2,000 | .4 | 10 | 55 |
| 40 | 2,000 | .4 | 13 | 93 |
| 60 | 2,000 | .4 | 13 | 97 |
| 20 | 250 | .8 | 7 | 4 |
| 40 | 250 | .8 | 10 | 13 |
| 60 | 250 | .8 | 14 | 19 |
| 20 | 750 | .8 | 10 | 2 |
| 40 | 750 | .8 | 17 | 16 |
| 60 | 750 | .8 | 18 | 35 |
| 20 | 2,000 | .8 | 16 | 3 |
| 40 | 2,000 | .8 | 2 | 16 |
| 60 | 2,000 | .8 | 8 | 32 |

## DISCUSSION

McDonald (1981) defined local independence as the strong principle of local independence and relaxed the definition to create the weak principle of local independence. Combined with monotonicity, his theory gives rise to the concept of fundamental unidimensionality (McLoughlin, 2000a). Whereas, Stout defined essential independence and weak monotonicity so he could define his theory of essential unidimensionality (Stout, 1987, 1990). Each is a theoretical construct designed to provide the statistician, psychometrician, or edumetrician with a foundation under which a test of approximate unidimensionality of a data set is executable - - the former tested by the NOHARM - CHIDIM procedure and the latter by the DIMTEST procedure. The competing theories and methodologies exist for the practical reason that the axiomatic unidimensionality assumption associated with UIRT models presumably cannot be strictly met with real data "because several cognitive, personality, and test taking factors always affect performance, at least to some extent. These factors might include level of motivation, anxiety, ability to work quickly, tendency to guess when in doubt about answers, and cognitive skills in addition to the dominant one being measured by the test items" (Hambleton, et al., 1991, p. 9). Furthermore, a test may be constructed to measure a particular content, but might in practice measure unintended content. However, the axiomatic unidimensionality assumption might be approximated, and that approximation must be adequately justified in order that a UIRT model is used to interpret test scores.

Again, one can argue that a goal of modeling is parsimony. Since UIRT is more easily interpretable than MIRT, it follows for many testing situations an essentially or fundamentally unidimensional test would be preferable to a multidimensional test. So, "the practice of reporting a single score, often number correct, is a widespread practice for all sorts of test administrations. Implicit in this is the conclusion that there is one dominant dimension driving the test response mechanism, at least up to a good approximation" (Junker & Stout, 1994, p. 51).

Given the practical reason for the theoretical construction, it is therefore crucial that the practical application of the theory demonstrates adequate precision of detection that a hypothesis of approximate unidimensionality is not rejected when a data set has been constructed by a unidimensional model and adequate precision of detection that a hypothesis of approximate unidimensionality is rejected when a data set has been constructed by a multidimensional model.

A limitation of this study was "a difficulty shared by all methods designed to reject or not to reject a null hypothesis of k dimensions [$k \geq 1$] is that all such hypotheses are false *a priori*" (McDonald, 1994, p. 83). Indeed, it should be noted that "the issue of dimensionality refers to a data matrix that is the result of the interaction of persons and test questions. The dimensionality refers only to the matrix, not to the items (or the test) by themselves, nor to the people. Thus, a set of items (a test) cannot be described *properly* [2] as being unidimensional, only the item response data matrix can be given that descriptive term" (Reckase, 1994, p. 88). However, if certain data matrix structures can be shown to be approximately unidimensional, then it may be helpful to researchers.

Another limitation was that Monte Carlo procedures attempt to simulate real testing conditions. However, not all testing situations can be simulated because real data are not as factorially pure as are found in this study; therefore, the results of this study cannot be generalised beyond the conditions simulated herein because they are specific to the simulated conditions.

Nonetheless, given the difference between McDonald's theory of fundamental unidimensionality and Stout's theory of essential unidimensionality, one might induce that the DGACS would be more conservative than the SNTS. The induction is predicated on the proofs of theoretical relationship between axiomatic, fundamental, and essential unidimensionality. Axiomatic unidimensionality implies fundamental unidimensionality, fundamental unidimensionality implies essential unidimensionality; whilst the converse of each statement is false (McLoughlin, 2000a). Thus, one might reasonably presume, would result in the DGACS have higher type II error rates because essential unidimensionality is less restrictive.

However, for the bivariate Sympson partially compensatory model, the DGACS demonstrated higher power than the SNTS for *all* conditions. For the bivariate McKinley - Reckase compensatory model, the DGACS demonstrated higher power than the SNTS for *most* conditions; the exceptions were the 250 subject, $\rho = .8$ conditions (regardless of test length) where both statistics demonstrated poor power and the SNTS was nominally better.

When a test is short, the number of subjects is low, and the correlation between dimensions is high, the DGACS failed to adequately detect the multidimensional nature of the data regardless of model used.

---

[2] Italics added.

In conditions where correlation between dimensions was high, the DGACS demonstrated as good or better power for PCMIRT data as opposed to COMIRT data. The opposite was true when correlation between dimensions was moderate or low. This may be due to the nature of the model: for compensatory data when correlation is high between dimensions it might be the case that the second dimensions compensates to the point that the dimension becomes somewhat indistinguishable from the first because of the additive property between dimensions as opposed to the partially compensatory model which has a multiplicative property between dimensions.

When the correlation between dimensions was high, the DGACS failed to adequately detect the multidimensional nature of the data regardless of model used. The DGACS performed quite poorly for short tests when the correlation between abilities was .8 using COMIRT or PCMIRT modeled data; but, performed acceptably when the number of subjects was 2,000. However, the SNTS performed quite poorly when the correlation between abilities was .8 using COMIRT or PCMIRT modeled data - - regardless of number of items or number of subjects.

Overall the SNTS demonstrated as good or better power for COMIRT data as opposed to PCMIRT data. This result is different than was observed for the DGACS. This result tends to reinforce the result found for the DGACS: for compensatory data when correlation is high between dimensions it might be the case that the second dimension compensates to the point that the dimensions become somewhat convoluted with the first dimension or dominated to such a point that it becomes undetectable regardless of statistic used as opposed to the partially compensatory model. Further, the results derived herein support the findings of previous research by Hattie, Krakowski, Rogers, and Swaminathan (1996): The SNTS performs better with COMIRT data than with PCMIRT data (they used the Hattie COMIRT model) and for short tests ($n < 25$) the SNTS performed poorly. However, their results report the SNTS performed better with COMIRT data than was found by this researcher. However, such seems not to be the case for PCMIRT data. It seems that the SNTS is sensitive to data modeled by a PCMIRT model.

It could be that under the COMIRT model, an approximate linear composite existed for the data in the .8 correlation between dimensions condition that was not present for the PCMIRT model. Hence, if such is true, it would support the findings of previous research by Reckase, Ackerman, and Carlson (1988).

The results found in this paper is may be due to the nature of the programs employed. First, the DGACS is computed in the program CHIDIM which follows NOHARM87 which used the non-linear factor analysis robust method of normal-ogive harmonic analysis to fit a unidimensional model to the data set. Given the proof that IRT models can be considered to be special cases of NLFA (McDonald, 1967 and Takane & de Leeuw, 1987), one could infer that this might produce a better fit. Nandakumar (1993) showed the results DIMTEST (Stout, et al., 1992) are dependent on the choice of AT1. Different choices of the AT1 produce different AT2s and PTs, which hence result in different T values. It could have been the automated choice of AT1 led to the results of this study.

Nonetheless, Nandakumar (1993) notes that the automated choice of AT1 would benefit a novice user. "The procedure is automated and totally data-dependent in its selection of assessment subtest items, making it more user friendly. The automation of size of assessment subtests could especially benefit the novice user" (Nandakumar, 1993, p. 63). However, given the dependency of T on the choice of AT1, spurious results may culminate, whereas with added input from expert opinion, a more complete analysis of the dimensionality of the data set might result. More research is needed that investigates the variation of T values given different methods of selection of AT1 and different sizes of AT1.

The fact that the DGACS had type II error rates across all conditions that were lower that the SNTS for data modeled with the Sympson PCMIRT model and that the DGACS had type II error rates across most conditions that were lower that the SNTS for data modeled with the McKinley - Reckase COMIRT might be due the nature of data sets employed in this study. Since the pseudo-guessing parameter was not fixed and the discrimination parameter was not bounded above by 1.1, these conditions might have created problems for the DIMTEST procedure. Previous studies of the Stout - Nandakumar T used a fixed pseudo-guessing parameter (0, .1, or .2) (Gessaroli & DeChamplain, 1996; Hattie, et al., 1994; Nandakumar & Stout, 1993; Stout, 1987) or did not mention what values were used for the pseudo-guessing parameter though reference to a three parameter model was made (Nandakumar, 1994). The effect of a varying pseudo-guessing parameter and not bounding the discrimination parameter were not the focus of this study, but should be in subsequent studies.

Stout, et al. (1993) in the DIMTEST manual recommend the number of items be greater than or equal to 20 (p. 1); Gessaroli & DeChamplain (1996) found the SNTS did not perform well when the number of items was 15; Nandakumar (1991a) recommends the number of examinees be greater than or equal to 750; DeChamplain & Gessaroli (1991) found the Stout T problematic with sample sizes of 500 or less; whilst this research indicates problems with the SNTS for sample sizes of 250, 750, and 2,000 but especially for sizes of 250 and 750 and problems with the SNTS when the number of items is 20, 40, or 60 but especially so for 20 or 40 item tests. Given that Stout's T and the Stout - Nandakumar T were developed using large sample distribution theory, the results contained in this study for the 250 subject case are not surprising, but the results for the 750 and 2,000 subject cases were surprising given previous studies.

With regard to number of items, twenty item tests may not be considered typical, but 20 item subtests are not rare. The range represented by forty and sixty items tests is common. Nonetheless, many tests are longer (e.g.: the SAT-V). Tests length should be varied and lengthened in subsequent studies. A focus on the number of items for a given number of subjects should be part of a subsequent study. If it is found, for instance, that there needs to be more than 300 items for the SNTS Type II error rates to stabilise for a given test length, then its practical usefulness might be questionable. Further, by lengthening the tests in subsequent studies, it might be found that at a certain point the DGACS begins to suffer sensitivity to test length. All the studies of the DGACS used short to moderate test lengths. It seems appropriate at this time given the results of this and previous research to subject the DGACS to longer tests.

With regard to sample size, it is known that $\chi^2$ statistics suffer as sample size increases. The point at which the DGACS deteriorates such that its utility is questionable was not found in this study nor was that question the focus of this study, but that question should be addressed in subsequent research. Given that the SNTS seems to suffer with small or moderate sample sizes, caution might dictate that it be used with confidence only with quite large samples ($N > 2,000$)[3].

"In the event that more than 2,000 examinees are being used, DIMTEST automatically splits them into roughly equal sized subsets that are made as large as possible without exceeding 2,000 in any subset. For each of these subsets [the DIMTEST procedure selection of PT and calculation of T] are performed separately resulting in a separate calculation of T for each subset. In this case,

the DIMTEST statistic is a normalized sum of the separate T statistics" (Stout, et al., 1993, p. 3). Study of DIMTEST for exceptionally large samples should be conducted (insofar as it seems only one published study has been conducted on the SNTS to date with *exceptionally* large samples, Nandakumar & Junker, 1993 [and one on Stout's T, Stout, 1987]).

Furthermore, once an upper limit to the size of the sample can be identified for the DGACS, it may be advisable to devise a method of sample splitting similar to the DIMTEST procedure to allow for large samples for CHIDIM.

Essential unidimensionality refers to an infinite pool of items from which a test is created. A problem for this study (indeed any study of procedures to assess approximate unidimensionality) is that essential unidimensionality has not been defined in such a way that any test of finite length (which are the only tests actually in existence) can be constructed to be essentially unidimensional or not. Essential unidimensionality is designed to identify whether there exists a sufficiently dominant dimension so that UIRT can be used to report and interpret scores. An overriding question is, therefore, "what is an essentially unidimensional test and how do we know that it is?" The answer to this question has not been produced yet. Research into this question is of great import in order to more succinctly assess the tractability of UIRT modeling. Indeed, the same is true of fundamental unidimensionality! Thus, for both the DGACS and SNTS, problems with data derived from either a COMIRT or PCMIRT model with correlation between abilities of .8 might not necessarily be a problem. Rejection of approximate unidimensionality may not necessarily imply rejection of the use of a unidimensional model because the model may be robust. This is because an educationally or psychologically meaningful definition of the approximate dimensionality of a test is not necessarily congruent with a statistically meaningful definition. More research into the structure of multidimensional models is advised.

Ease of use is a problem that should be addressed. DIMTEST is *far* more user friendly than NOHARM-CHIDIM. DIMTEST comes in one commercially available package; whereas NOHARM and CHIDIM are separate packages. CHIDIM requires the upper triangular matrix (lower off-diagonal entries) of raw Pearson product-moment covariance matrix and the upper triangular matrix (lower off-diagonal entries) of residual correlations matrix which are not saved as separate files by NOHARM, but are included in the NOHARM report. CHIDIM requires

---

[3] The term large sample size is a subjective one. Stout, et al. (1993) define samples of less that 2,000 "small,"

different formatting of the raw Pearson product-moment covariance matrix and the residual correlations matrix than is reported by NOHARM. These obstacles make the running of NOHARM-CHIDIM extremely unfriendly. It would be highly desirable if Fraser, McDonald, DeChamplain, and Tang pooled the programs into a single package. It would facilitate more research into NOHARM-CHIDIM without diminishing the importance of NOHARM as a stand alone program.

Further research is needed as to type II error rates (and type I) error rates for both the DGACS and SNTS: varying parameters, test length, number of subjects, number of dimensions, and structures of tests.

It seems warranted in light of the results of this study to question the utility of the SNTS for data modeled with the Sympson PCMIRT model. It seems warranted in light of the results of this study to advise caution when assessing approximate unidimensionality for short or moderate length tests using the SNTS for data modeled with the McKinley - Reckase COMIRT model. Clearly, more research is needed. Indeed, given the scant amount of research, the same *may* be true of the DGACS. However, provided the correlation between dimensions is not high, it seems the DGACS is more appropriate for assessing approximate unidimensionality for short or moderate length tests than the SNTS for data modeled with either the Sympson PCMIRT model or the McKinley - Reckase COMIRT model.

Finally, the dearth of replication studies in this area (as well as most areas of edumetrics, psychometrics, etc.) is most problematic. This may be due to a culture gripping the field - if it is not *new*, then it is not presentable or publishable; or, it may be due to the fact that IRT is a rapidly changing field. Without the added confidence of replication studies, neither of the theories presented here can be recommended with high confidence.

---

samples of size 2,000 to less than 40,000 “moderate,” and samples of size greater than or equal to 40,000 “large.”

## REFERENCES

Bartlett, M. S. (1950). Tests of significance in factor analysis. British Journal of Psychology, 3, 77 - 85.

Berger, M. P. F. & Knol, D. L. (1990, April). On the assessment of dimensionality in multidimensional item response theory models. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, MA.

Birnbaum, A. F. (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 (pp. 396 - 451). Reading, MA: Addison - Wesley.

DeChamplain, A. F. (1992). Assessing test dimensionality using two approximate chi-square statistics. Unpublished doctoral dissertation, University of Ottawa.

DeChamplain, A. F. (1995, April). Assessing the effect of multidimensionality on IRT true-score equating for subgroups of examinees. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA. (ERIC Document Reproduction Service No. 382 643).

DeChamplain, A. F. (1996, April). Assessing the dimensionality of item response matrices using a goodness-of-fit index based on non-centrality. Paper presented at the annual meeting of the American Educational Research Association, New York, NY. (ERIC Document Reproduction Service No. 397 100).

DeChamplain, A. F. & Gessaroli, M. E. (1991, April). Assessing test dimensionality using an index based on non-linear factor analysis. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL. (ERIC Document Reproduction Service No. 334 235).

DeChamplain, A. F. & Gessaroli, M. E. (1992, April). Using an approximate chi-square statistic to test for the number of dimensions underlying the responses to a set of items. Paper presented at the annual meeting of the American Education Research Association, San Francisco, CA.

DeChamplain, A. F. & Gessaroli, M. E. (1996, April). Assessing the dimensionality of item response matrices with small sample sizes and short test lengths. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY. (ERIC Document Reproduction Service No. 397 099).

DeChamplain, A.F. & Gessaroli, M. E. (1997, March). An empirical comparison of two LISREL chi-square goodness-of-fit statistics and the implications for dimensionality assessment of item response data. Paper presented at the annual meeting of the American Educational Research Association, Chicago. IL. (ERIC Document Reproduction Service No. 411 269).

DeChamplain, A. & Tang, L. (1993, April). The effect of non-normal ability distributions on the assessment of dimensionality. Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

DeChamplain, A. & Tang, L. (1997a). CHIDIM [Computer Program].

DeChamplain, A.F. & Tang, K. L. (1997b). CHIDIM: A fortran program for assessing the dimensionality of binary item responses based on McDonald's non-linear factor analytic model. Educational and Psychological Measurement, 57 (1), 174 - 178.

Douglas, J. Kim, H. R., & Stout, W. (1994). Exploring and explaining the lack of local independence through conditional covariance functions. Unpublished manuscript, Champaign, IL: University of Illinois.

Fraser, C. (1988). NOHARM II: A computer program for fitting unidimensional and multidimensional normal ogive models of latent trait theory. New South Wales, AU.: The University of New England, Center for Behavioural Studies.

Fraser, C. (1998). NOHARM87 [Computer Program] Ontario, Canada: Niagara College.

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

Gao, F. O. (1997). DIMTEST enhancements and some parametric IRT asymptotics. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Gessaroli, M. E. (1994). The assessment of dimensionality via local and essential independence: A comparison of theory and practice. *In* D. Laveault, B. D. Zumbo, M. E. Gessaroli & M. W. Boss (eds.), Modern Theories of measurement: Problems and issues. Ottawa, ON: Edumetrics Research Group.

Gessaroli, M. E. & DeChamplain, A. F. (1996). Using an approximate chi-square statistic to test for the number of dimensions underlying the responses to a set of items. Journal of Educational Measurement, 33 (1), 157 - 179.

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

Haley, D. C. (1952). Estimation of the dosage mortality relationship when the dose is subject to error. Technical Report No, 15. Stanford, CA.: Stanford University, Applied Mathematical and Statistical Laboratory.

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

Hattie, J. (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.

Harrison, D. A. (1986). Robustness of IRT parameter estimation in violation of the unidimensionality assumption. Journal of Educational Measurement, 11, 91 - 115.

Humphreys, L. G. (1982). Systematic heterogeneity of items in tests of meaningful psychological attributes: A rejection of unidimensionality. Unpublished manuscript, Urbana, IL.: University of Illinois.

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

Junker. B. W. & Stout, W. F. (1994). Robustness of ability estimation when multipletraits are present with one dominant trait. *In* D. Laveault, B. D. Zumbo, M. E. Gessaroli & M. W. Boss (eds.), Modern Theories of measurement: Problems and issues. Ottawa, ON: Edumetrics Research Group.

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

McDonald, R. P. (1967). Non-linear factor analysis. Psychometric Monographs (No. 15) New York, NY: The Psychometric Society.

McDonald, R. P. (1979). The structural analysis of multivariate data: A sketch of general theory. Multivariate Behavioral Research, 14, 21 - 38.

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

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

McDonald., R. P. (1988). An index of goodness-of-fit based on non-centrality. Journal of Classification, 6, 97 - 103.

McDonald, R. P. (1994). Testing for approximate dimensionality. *In* D. Laveault, B. D. Zumbo, M. E. Gessaroli & M. W. Boss (eds.), Modern Theories of measurement: Problems and issues. Ottawa, ON: Edumetrics Research Group.

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

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

McDonald, R. P. & Mulaik, S. A. (1979). Determinacy of common factors: A non technical review. Psychological Bulletin, 86 (2), 297 - 306.

McKinley, R. L. & Reckase, M. D. (1983, April). The use of IRT analysis on dichotomous data from multidimensional tests. Paper presented at the annual meeting of the American Educational Research Association, Montréal, Québec. (ERIC Document Reproduction Service No. 228 332).

McLoughlin, M. P. M. M. (2000a). On axiomatic, fundamental, and essential unidimensionality in incomplete item response spaces. Unpublished doctoral dissertation, Georgia State University.

McLoughlin, M. P. M. M. (2000b, October). The effect of non-normal ability distributions on the assessment of approximate unidimensionality in item response spaces. Paper presented at the annual meeting of the Georgia Educational Research Association, Morrow, Georgia.

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

Nandakumar, R. (1991a). Traditional dimensionality versus essential dimensionality. Journal of Educational Measurement, 28 (2), 99 - 117.

Nandakumar, R. (1991b, April). Assessing dimensionality of a set of items - Comparison of different approaches. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Nandakumar, R. (1993). Assessing essential unidimensionality of real data. Applied Psychological Measurement, 17 (1), 29 - 38.

Nandakumar, R. (1994). Assessing dimensionality of a set of items - Comparison of different approaches. Journal of Educational Measurement, 31 (1), 17 - 35.

Nandakumar, R. & Junker, B. W. (1993, April). Estimation of latent ability distributions under essential unidimensionality. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA. (ERIC Document Reproduction Service No. 359 207).

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

Nandakumar, R. & Yu, F. (1994, April). Testing robustness of DIMTEST on non-normal ability distributions. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. 371 011).

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

Roussos, L.A., Stout, W. F., & Marsden, J. I. (1993). Analysis of the multidimensional structure of standardized tests using DIMTEST with hierarchical cluster analysis. Unpublished manuscript, Champaign, IL: University of Illinois.

Spence, I. (1983). Monte carlo simulation studies. Applied Psychological Measurement, 7, 405 - 425.

SPSS, Inc. (1998). SPSS 8.0 [Computer Program]. Chicago, IL: SPSS, Inc.

Stout, W. F. (1987). A non-parametric approach for assessing latent trait unidimensionality. Psychometrika, 52 (3), 589 - 617.

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

Stout, W. F. , Habing, B., Douglas, J., Kim, H. R., Roussos, L., & Zhang, J. (1995). Conditional covariance - based non-parametric multidimensional assessment. Applied Psychological Measurement, 19 (4), 331 - 354.

Stout, W. F., Junker, B., Nandakumar, R., Chang, H. H., & Steidinger, D. (1992a). DIMTEST [Computer Program]. Urbana, IL: Department of statistics, University of Illinois.

Stout, W. F., Junker, B., Nandakumar, R., Chang, H. H., & Steidinger, D. (1992b). DIMTEST: A fortran program for assessing dimensionality of binary item responses. Applied Psychological Measurement, 16, 236.

Wang, Y. L. (1994, April). Robustness of unidimensional IRT calibration in the presence of essential unidimensionality. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. 371 019).

Wang, Y. L. & Hocevar, D. (1994, April). Effects of mathematics text context on essential dimensionality in U. S. and Japan data. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. 372 090).

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