National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file Votre référence

Our file Notre référence

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

Canada

A Comparison of the Approximate χ^2 and DIMTEST in Conditions of Pseudo-Guessing and Correlated Factors

by

Krista J. Breithaupt

Faculty of Education

Thesis submitted to
the school of Graduate Studies and Research
in partial fulfillment of the requirements for the
MA degree in Education

University of Ottawa

ISBN 0-612-07837-X

Canada

UNIVERSITÉ D'OTTAWA
UNIVERSITY OF OTTAWA

ACKNOWLEDGEMENTS

This research project could not have been completed without the timely intervention of several individuals. Thanks to Dr. Marc Gessaroli for his expertise, instruction and criticism. Thanks to Dr. Bruno Zumbo for his early encouragement and inspiration. Thanks to Jack Boulet for pointing me in the right direction.

Abstract

An area of current interest in educational measurement is the assessment of the dimensionality underlying the responses to a set of test items. The determination of the dimensionality, or more specifically the assessment of the lack of unidimensionality, is extremely important. Item Response Theory (IRT) is commonly used to estimate item parameters (difficulty and discrimination) as well as subjects' abilities. However, the validity of common IRT methods depends upon the requirement that the data be unidimensional. When this assumption is false, the abilities of students might be poorly estimated leading to possible serious consequences such as failure on certification or liscensure examinations.

The purpose of this study was to compare two promising methods to assess the dimensionality underlying a set of responses to test items; Stout's T statistic (Stout, 1987) and a form of non-linear factor analysis, the approximate χ^2 statistic (De Champlain, 1992). Response data was simulated to compare the number of rejections of unidimensionality made using the approximate χ^2 statistic and using Stout's T statistic. Two conditions have been identified in the literature as possibly affecting the accuracy of dimensionality assessment techniques; the degree of correlation among abilities (for two-dimensional data) and the amount of guessing present in the data. These two variables were manipulated in the generation of 100 different data sets, each with scores for 1000 examinees under each experimental condition for both two-dimensional and unidimensional response data. The dimensionality underlying each data set was assessed by the program NOHARM II (Fraser & McDonald, 1988) for the approximate χ^2 and by DIMTEST (Stout, Douglas, Junker & Roussos, 1993) for Stout's T statistic.

Both statistics had higher Type I error rates when unidimensional data was simulated with pseudo-guessing present. As has been found in earlier studies, longer tests were also associated with lower Type I error rates for these statistics in the unidimensional data condition. Both statistics performed with acceptable levels of Type I errors (less than 5% in most conditions).

Logit-linear analysis of the number of rejections of unidimensionality made by both statistics showed main effects of pseudo-guessing, test length and the level of correlation of the latent factors in the two-dimensional simulation condition. Longer tests had a greater number of (correct) rejections of unidimensionality. Rejections of unidimensionality were less frequent when the latent traits were highly correlated (r = .7). Both statistics performed well when latent traits were correlated at r = .5. There were more correct rejections of unidimensionality made with response data which contained no pseudo-guessing.

This study suggests that the approximate χ^2 performs at least as well as DIMTEST, with lower Type I error rates in the unidimensional simulation conditions. Both statistics also performed fairly well in two-dimensional simulation conditions, with the exception of response data with highly correlated latent factors (r = .7). Continued study with real and simulated data is needed to determine the effects of complex compensatory two-dimensional latent structure and larger sample sizes (over 1000 subjects) on indices of dimensionality. There is broad application for valid indices of dimensionality wherever decisions about individuals are made based on the interpretation of scores from a test.

Table of Contents

Index of Tables

Introduction

Users of tests commonly wish to assume that examinees' scores from a test, or from a given subsection of a test, reflect some defined ability of that examinee. This underlying ability or *latent trait* is then described as a unitary concept such as verbal aptitude, mathematical ability, or reading skill. The interpretation of scores reflects this assumption of unidimensionality (Hattie, 1984; McDonald, 1981; Reckase, 1986). Unidimensionality is an underlying assumption of item response theory or IRT (Crocker & Algina, 1986). The common logistic item response model (or IRM) can be generally described as a probability function closely approximating the monotonically increasing normal ogive. The probability of a correct or incorrect response is predicted from this model with parameters to represent any or all of : ability (θ), difficulty (b), discrimination (a), and level of pseudo-guessing (c) in response data for items. For an overview of variations on IRT models for dimensionality assessment the reader is referred to McDonald (1982).

The unidimensional IRM does not mathematically represent response probabilities where more than one latent trait determines responses to test items. It has been shown that the assumption of unidimensionality of the latent trait may often be unrealistic in real test data. Recently, researchers have attempted to develop statistical methods to assess the actual dimensionality of a set of items on a test (e.g. De Champlain & Gessaroli, 1992; Hattie, 1985; McDonald, 1994). The alternative item response model is multidimensional, where each of two (or more) traits influence responses to items on the test. The dominance and correlation of the latent traits will vary. For example, on a mathematical ability test we might expect geometry skill to be strongly correlated with mathematical skill, but less so with verbal skill. In this example you

would expect to recover three dimensions of the latent trait, with the degree of correlation and the relative weight of the three suggested factors to be reflected by the multidimensional variation of the common logistic IRM.

Serious consequences for examinees may result if multidimensional data are interpreted using a unidimensional IRM. Researchers have described the effects of multidimensionality on test calibration and interpretation using the (unidimensional) IRM (Camilli, 1995). There are two principal issues of test misinterpretation possible when the assumption of unidimensionality is violated. The first and most important consequence is the possible misclassification of examinees (Hattie, 1984). When the IRM is used to estimate ability, examinee scores may be used to make decisions of consequence. Misclassification of individuals may occur when the test is measuring more than the one ability of interest to the decision makers (De Champlain, 1992). This is very risky when examinees are ranked, selected into special programs, or classified for mastery or non-mastery based on multidimensional data using a unidimensional IRM. This error could result in certification of non-competent professionals, or in misdiagnosis of learning disabilities in the case of children.

The second concern is also related to the misclassification of examinees; errors occurring when the IRM is used to estimate item parameters. The parameters of the response function may be estimated incorrectly due to the application of a unidimensional IRM to multidimensional data. Errors may occur in estimation of any of the parameters of the model; the difficulty, discrimination, or pseudo-guessing (upper asymptote) parameters of items may be estimated incorrectly. It has been assumed in the common IRM that performance on the item was correctly approximated by these variables, given a single latent trait. When multidimensionality is present

the performance of examinees will not be accurately represented by the traditional IRM, particularly when the underlying traits are correlated (Ackerman, 1987; Reckase, 1979). This might invalidate IRT-based techniques such as equating forms of a test, investigating item bias or DIF (Camilli, 1992; Oshima & Miller 1992), or in test design using IRT to calibrate items for item banking. Our desire to make valid decisions based on test performance and to apply an accurate IRM for calibration and interpretations of response data has inspired the search for useful statistical indicators of test dimensionality. This simulation study examined two-dimensionality assessment methods; DIMTEST and the approximate χ^2 statistic.

Dimensionality Assessment Methods

A great variety of methods have been suggested, studied and reviewed in the literature. A brief description of various methods which have been used in dimensionality assessment is needed to lay the foundation for the present study. However, this discussion is limited to those which are relevant to the dimensionality assessment methods applied here. Examples of some methods based on linear factor analysis are followed by a presentation of methods using some form of non-linear factor analysis. The final portion of the literature review describes the methods which were applied in the present study; DIMTEST (which calculates Stout's T statistic), and the approximate χ^2 statistic.

In his review of dimensionality statistics Hattie (1984) studied 87 different techniques, organized into four categories: indices based on answer patterns, indices based on reliability, indices based on component or factor analysis, and indices based on latent trait models. Later,

McDonald reviewed several varieties of dimensionality assessment methods (McDonald, 1989, 1995). A thorough review is also available in Gessaroli and De Champlain (1992).

Several studies have assessed the dimensionality of response data using procedures based on a linear factor analysis or a principal components analysis. These methods include indices which use the size of the first eigenvalue, the number of eigenvalues greater than 1, or an analysis of scree plots. Such methods are based on the analysis of a matrix of phi coefficients and are found to overestimate the true dimensionality of a set of items that had been derived from a logistic model (Berger & Knol, 1990; Hattie, 1984). Spurious or 'difficulty' factors are recovered using this method. These spurious factors have been described as due to the effect of items which have nonlinear regressions on the latent trait(s) (McDonald & Ahlawat, 1974). For example, the relationship between the probability of a correct response and level of ability has been shown to be nonlinear at higher points on the ability scale (McDonald & Ahlawat, 1974). For this reason, many authors caution the use of linear factor analysis or principle components analysis to assess dimensionality (e.g. De Champlain, 1992).

In their simulation study, Knol and Berger (1991) found that linear factor analysis using a tetrachoric correlation matrix recovered the true parameters fairly well. However some numerical problems can occur with this method, such as non-Gramian correlation matrices and Heywood cases (Berger & Knol, 1990; Hattie, 1984). These problems require smoothing to constrain variances to positive values. Estimation of tetrachoric correlations also assumes that a normally distributed variable underlies responses to items. This has been shown to be untenable theoretically, and as a result many authors do not support linear factor analysis of a tetrachoric correlation matrix for dimensionality assessment (e.g. McDonald, 1981; Hattie, 1984).

The conditional association between pairs of items has been used as the basis of Holland and Rosenbaum's approach (Holland and Rosenbaum , 1986). This method is based on their supposition that if ICCs are monotone nondecreasing functions of a single ability, the local independence of item responses lead to positive conditional covariances for all item response pairs. Conditional association for each pair of items is then tested with the Mantel-Haenszel statistic, and significance is determined by referring to the lower tail of the standard normal distribution. A large number of conditionally associated pairs of items would suggest that the test is multidimensional.

Nandakumar (1994) found that the Holland and Rosenbaum method showed fairly good power in a study with simulated and real response data. This result was consistent with an earlier study by Zwick (1987) where the Holland-Rosenbaum method was shown to perform comparably with a method based on nonlinear factor analysis. Nandakumar pointed out a meaningful difference in the hypothesis tested by the Holland and Rosenbaum method, specifically that it "...looks at all item pairs and detects items that are not measuring the same trait as other items of the test" (Nandakumar, 1994 p.33). This differs from recently popular approaches which test unidimensionality through a specific test of local item independence based on zero (or nearly zero) conditional item covariances (e.g. Stout's T statistic as described in Stout, 1987). Problems may be expected when complex structure is present. Nandakumar (1994) found that the Holland and Rosenbaum approach had poor power applied to two-dimensional response data with strongly correlated dimensions (e.g. when r = .7). More research of this method in a variety of conditions of response data is needed.

Several dimensionality statistics and indices have been studied which examine the residuals after some form of non-linear factor analysis has been performed. McDonald (1989) discusses several statistical assessment methods for dimensionality assessment which range from joint-, marginal- and conditional maximum likelihood estimators, to Bayesian estimation and item factor analysis. He summarized the techniques by stating that non-linear tests of fit using common factor analysis are most robust, and have a strong theoretical basis. McDonald later suggested that the most promising methods make use of the fact that IRT and linear common factor analysis are "...two special cases of a general nonlinear factor model, definable by a strong (full information) or weak (bivariate information) form of the principle of local independence..." (McDonald, 1994, p 72). A brief overview of these methods follows.

Statistics based on full-information factor analysis have been proposed by several researchers (e.g. Bock & Aitkin, 1981). Their approach uses all the information in binary responses to items; specifically, higher order associations are calculated to compute a full-information factor analysis (FIFA). McDonald (1994) states that FIFA is a test of the *strong* principle of local item independence, and that FIFA may have a stronger theoretical basis for dimensionality assessment than tests of the *weak* principle of local item independence (nonlinear factor analysis which uses only bivariate information). A disadvantage of FIFA has been the large amount of CPU time required to analyse even a relatively small set of items (Berger & Knol, 1990; Knol & Berger, 1991). Also, there are 2^p distinct response vectors where p is equal to the number of items, and very large sample sizes are required as the number of items analysed increases.

Several statistics and indices have been proposed to assess the fit of models of specified dimensionality obtained with FIFA computer programs such as TESTFACT (Wilson, Wood & Gibbons, 1984). Studies of the effectiveness of such indices (for example the likelihood ratio χ^2 difference test) have used real data where the dimensionality was not known a priori (Bock, Gibbons & Muraki, 1988). These indices need to be studied in conditions where the dimensionality of the response data is known. FIFA may be a theoretically promising alternative for assessing the dimensionality of test items, however, studies of statistics associated with FIFA need to be extended.

Nonlinear factor analysis of bivariate information in the response data has been used with some success for dimensionality assessment by several researchers (for example, Christoffersson, 1975; De Champlain 1992). Two distinct forms of this method are the generalised and the unwieghted least-squares estimation procedures. Christoffersson (1975) and Muthen (1978) use generalised least-squares in the estimation of the parameters. From this they derive a theoretically based χ^2 to test the adequacy of the proposed model. However, because of restrictions of generalised least-squares estimation, the analysis is limited to about a maximum of 25 items.

McDonald's (1967) approach uses unweighted least-squares estimation of the model parameters and has the practical advantage of allowing for the analysis of tests with a large number of items and dimensions. Several researchers have proposed indices based on McDonald's *weak* principle of local independence to assess the number of dimensions underlying a set of item responses (Berger & Knol, 1990; Hattie, 1984; Knol & Berger, 1991; Nandakumar,1994). The most promising of these incorporate some examination of residual covariances after a model of specified dimensionality has been fitted to the response data (e.g.

Berger & Knol, 1990; De Champlain & Gessaroli, 1992; Hattie, 1984). One example is the Incremental Fit Index which is based on the sum of squares of residual covariances obtained after fitting an *m*-factor model (De Champlain & Gessaroli, 1991).

Knol and Berger (1991) have shown that estimates obtained with NOHARM II (Fraser, 1988) which tests bivariate information in the response data, compare very favourably with those obtained from the full-information methods used in TESTFACT. NOHARM II is based on McDonald's *weak* principle of local item independence. This method seems promising but yields only descriptive indices of dimensionality. Nandakumar recognises this limitation in her recent study of several methods of dimensionality assessment including one based on nonlinear factor analysis. She comments that "...some type of adequacy of fit statistics with associated sampling distributions would be necessary to aid in assessing the fit of nonlinear models" (Nandakumar, 1994, p.32).

Recently, two-dimensionality assessment methods have been found favorable. Stout's T statistic (Stout, 1987) and nonlinear factor analysis have received attention in recent literature. In reviews by other researchers (e.g. Berger & Knol, 1990; De Champlain & Gessaroli, 1992) there is agreement that methods based on some form of non-linear factor analysis with an examination of residuals seems promising.

Responding to the need for a test of significance using nonlinear factor analysis based on unweighted least-squares estimation, De Champlain and Gessaroli (1992) have shown that an approximate χ^2 method correctly recovers the dimensionality of the latent trait structure in various conditions. McDonald (1967, 1981, 1985) has provided detailed support for the application of non-linear factor analysis to derive an IRM for binary response data. Alternatively,

Stout's T statistic (Stout, 1987) makes use of the principle of *essential unidimensionality* and has been found accurate in identifying unidimensional data in various conditions (De Champlain & Gessaroli, 1991; Nandakumar, 1994; Nandakumar & Stout, 1993; Stout, 1987). In the present study the approximate χ^2 statistic and Stout's T statistic will be farther investigated to test unidimensionality of simulated binary response data.

Stout's T Statistic

Stout's T statistic is used to test the null hypothesis of essential unidimensionality of the response data (Stout, 1987). This method is based on the principle that the conditional mean absolute residual covariance value will approach zero when (unidimensional) ability has been partialled out, as test length and sample size approach infinite size (Stout, 1987, 1990). This procedure is sensitive to dominant dimensions of the latent trait and less so to minor dimensions. The specifics of computation, performed by the computer program DIMTEST, are available in Nandakumar and Stout (1993) or more recently, in Nandakumar (1994).

The general procedure is to divide the test response data into three subsections, AT1, AT2, and PT. The AT1 section represents unidimensional items (according to content analysis or factor analysis of a tetrachoric matrix); the AT2 section is made up of items which have the same difficulty level as the AT1 item response data but are representative of the remaining items on the test. The T statistic is computed from the AT1 subset; the AT2 items correct the statistic for test length and difficulty levels. The final section, PT is made up of the remaining items from the test. The PT items are used to partition the examinees according to ability level. For each set of response data the AT1 conditional score variance estimate (corrected for bias of difficulty and test length by the AT2 subset) is assessed to determine if it differs from the PT conditional score variance estimate (Stout, Douglas, Junker & Roussos, 1993). The value of this difference is summarized by the T statistic, such that a significant value suggests multidimensionality of the data.

The T statistic obtained is referred to the upper tail of the standard normal distribution. The hull hypothesis of unidimensionality is rejected when the obtained T is significant at the chosen level of α. Small (significant) p values suggest multidimensionality and larger (non-significant) p values suggest unidimensionality (Nandakumar, 1994).

Stout's T statistic has been found to accurately test the assumption of essential unidimensionality (Nandakumar & Stout, 1994; Stout 1987). There is some evidence that when response data contain low discrimination values for items and shorter test lengths Stout's T is less powerful, particularly with smaller sample sizes (Nandakumar, 1987; Stout, 1987). In applications with simulated multidimensional data, De Champlain (1992) found that test length, sample size and the strength of a second dimension influenced the accuracy of Stout's T statistic.

Nandakumar and Stout (1993) have modified Stout's T. The statistic has been revised to account for the poor performance found in Stout's T under conditions of pseudo-guessing combined with high discrimination values in the items. Nandakumar (1994) states that these improvements have made the statistic "...robust against the presence of pseudo-guessing in item responses; (with) better control of the observed level of significance, and greater power; and automation of the size of assessment subtests..." (p 21). There is also some indication that the more powerful version of Stout's T statistic is able to correctly identify dimensionality when the (multidimensional) response data have correlated latent traits, with correlations as high as .7 (Nandakumar & Stout, 1993).

While Stout's T statistic has been investigated to some degree under several conditions, it appears that no systematic study has considered the performance of the statistic in the presence of possible pseudo-guessing. An investigation of this type would seem to be overdue.

Approximate χ^2 Statistic

The approximate χ^2 statistic for assessing the dimensionality of a set of examinee responses to test items uses nonlinear factor analysis and the 'weak' principle of local

independence. Gessaroli & De Champlain (1995) provide a detailed description of the calculation of this approximate χ^2. In general the approximate χ^2 tests the null hypothesis that the off-diagonal elements in a matrix of residual correlations are equal to zero. This examination of residuals can be performed after the fit of a model of specified dimensionality (e.g. *m* dimensions), so that the fit of the *m*-factor model may be evaluated. Computations are available in De Champlain (1992), and in De Champlain and Gessaroli (1994). The five steps in calculating the χ^2 statistic can be summarized as follows:

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 *m*-factor 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 correlations ($r_{ij}^{(r)}$) for each pair of dichotomous items with the following formula:

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

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

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

5. Calculate an approximate χ^2 statistic defined as

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

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,...,p) and N is the number of subjects in the sample. This statistic is approximately distributed as a central χ^2 with df = .5k (k - 1) - t where k is equal to the number of items and t is the total number of independent parameters estimated.

De Champlain (1992) assessed the performance of the statistic using unidimensional and two-dimensional Monte Carlo data. The number of rejections made by the approximate χ^2 were calculated with various test lengths (15, 30, or 45 items), sample sizes (500 and 1000 examinees), and dimension dominance structures. In all two-dimensional conditions, item parameters consistent with a simple structure were used. That is, only one factor influenced the response to any item. Dimension dominance was set at two levels for the (two-dimensional) response data. In the 80:20 condition the first factor loaded on 80% of the items and the second (weaker) factor loaded on the remaining 20% of the items. A second level of dimension dominance was set so that the responses to approximately 50% of the items were uniquely influenced by (loaded on) either one of the latent factors. The correlation between the two latent traits was set to be zero for all the two-dimensional simulations. As well, no pseudo-guessing was introduced in the generation of the item responses.

The only condition which influenced the empirical Type I error rate with unidimensional data was the length of the test. The maximum number of (false) rejections of unidimensionality was 4% in any of the 18 conditions.

In the conditions where two-dimensional data were simulated, the main effects of test length and sample size were related to the number of rejections of the hypothesis of unidimensionality. There was a greater number of rejections with longer test lengths and larger

sample sizes. As well, larger numbers of rejections were found with stronger dimension strengths and less dimension dominance (i.e., the 50:50 condition).

In summary, the approximate χ^2 was able to correctly reject the assumption of unidimensionality in most conditions. The number of correct rejections made was above 95% in 32 of the 36 conditions simulated. However, the temptation to overgeneralize the results must be controlled due to the nature of data simulated. For example, the simple structure in the factor patterns (i.e., discrimination parameters) along with setting the factors to be uncorrelated might have influenced these results. As well, the effect of pseudo-guessing was not considered. It would seem necessary to further explore the performance of the approximate χ^2 under more complex conditions such as allowing for correlated factors and pseudo-guessing.

Research Relating Stout's T and the Approximate χ^2 Statistics

Only one study has been carried out previously wherein the performance of the approximate χ^2 and Stout's T was compared (De Champlain, 1992). The performance of the two procedures was quite similar in almost all conditions. Specifically, both procedures had good Type I error control and both tests correctly rejected the assumption of unidimensionality with multidimensional data with test lengths of 30 and 45. However, as expected (Nandakumar, 1994), Stout's T performed quite poorly in correctly rejecting undimensionality with the 15 item test lengths. The approximate χ^2, however, had rejection rates that were much higher than Stout's T in these conditions. Again, it must be emphasized that these results may not be generalizable to other conditions and further study is warranted.

Advantages of the χ^2 are summarized by De Champlain (1992), including it's derivation from nonlinear factor analysis, of which IRM is a special case. NOHARM II uses an unweighted least-squares estimation procedure for nonlinear factor analysis, and thereore, there is no practical limit on the number of dimensions which can be analyzed. The promising initial findings of De Champlain need to be extended to examine performance in conditions where the latent traits are

correlated and where pseudo-guessing is simulated. Further study would provide more information as to the generalizability of the approximate χ^2.

Purpose of the Study

The purpose of the present study was to assess the performance of an approximate χ^2 and Stout's T statistics in the presence of simulated response data which contains pseudo-guessing, and/or correlated abilities. Specifically the research questions to be addressed were:

1. What are the simple and interactive effect of pseudo-guessing and test length on the Type I error rate of the approximate χ^2 and Stout's T statistics?

2. What are the simple and interactive effect of pseudo-guessing, test length and correlated factors on the number of rejection of unidimensionality made by the approximate χ^2 and Stout's T statistics when two-dimensional data are simulated?

Methods

The independent variables examined in this study were test length and pseudo-guessing for unidimensional data sets, and test length, pseudo-guessing and correlation of abilities for multidimensional data sets. To determine the effect of pseudo-guessing two levels were simulated. The value of the pseudo-guessing parameter ranged between $c = .05$ and $c = .25$ in the pseudo-guessing present condition and was set to $c = 0$ for the no pseudo-guessing condition. Test lengths were varied to be either 30 or 45 items in each condition. The correlation between abilities had three levels for the multidimensional data sets: $r = 0$, $r = .5$, or $r = .7$.

There were 4 unique conditions (2 test lengths X 2 levels of pseudo-guessing) for the unidimensional data sets and 12 unique conditions (2 test lengths X 2 levels of pseudo-guessing X 3 levels of correlated abilities) for the multidimensional data sets. Response data for 1000 examinees were simulated for each data set, and 100 data sets were simulated for each condition of the study.

Item Response Model for Unidimensional Data Sets

Response data were generated using the computer program M2PLGEN, (Ackerman, 1985). The general form of the three-parameter logistic model (after eliminating subscripts for subjects) is:

$$P(i = 1|\underline{\theta}) = c_i + (1 - c_i)\frac{e^{1.7(\underline{a}'\underline{\theta}+d_i)}}{1+e^{1.7(\underline{a}'\underline{\theta}+d_i)}},$$

where $\underline{\theta}$ is the vector of abilities on the latent traits;

$\underline{a}$ is the vector of discriminations of item i on the latent traits;

c_i is the pseudo-guessing parameter for item i; and,

d_i is a scalar related to the difficulty of item i.

The model simplifies to the usual three-parameter logistic IRT model in the unidimensional case. In particular, the unidimensional difficulty parameter (b) for item i is related to d_i by $b_i = -d_i/a_i$.

Ability (θ) was independently generated from a standard normal distribution. The values for discrimination (a) and difficulty (b) were randomly generated from a normal distribution based values derived from the ACT-English Test Battery (Drasgow & Parsons, 1983). The means and variances of the difficulty and discrimination parameters are shown on Table 1.

Table 1

Means and Variances of Items Parameters for Unidimensional Data Simulation

IRM Parameter Distribution	Difficulty (b)	Discrimination (a)
μ	0.00	0.72
σ^2	0.96	0.25

The value of pseudo-guessing (c) for each item was randomly generated within a range of c = 0.05 to c = 0.25 for the pseudo-guessing present condition, and with c = 0 for the no pseudo-guessing condition. The resulting design was fully crossed resulting in a 2 X 2 design (test length by pseudo-guessing).

Item Response Model for Two-dimensional Data Sets

The multidimensional response data were generated using the same computer program and three-parameter logistic model as was used for the unidimensional data. The values for the latent traits (θ_1 and θ_2) were randomly sampled from a bivariate normal distribution.

Dimension dominance was set at 80:20 to represent a dominant first factor and a weaker second factor. For the 30 item test length condition the first factor uniquely determined responses to 24 items, and the second factor uniquely determined responses to 6 items. In tests with 45 items there were 36 items for the first factor, and 9 items for the second factor.

Results

The purpose of this study was to answer two research questions; first, what are the simple and interactive effect of pseudo-guessing and test length on the Type I error rate of the approximate χ^2 and Stout's T statistic? Second, what are the simple and interactive effect of pseudo-guessing, test length, and correlated dimensions on the number of rejections of unidimensionality made by the two statistics when two-dimensional data were simulated?

The first research question was addressed by counting the number of rejections made with the approximate χ^2 and with Stout's T statistic in conditions with various test lengths and/or pseudo-guessing. Type I error rates were equal to the number of false rejections of the null hypothesis of unidimensionality made by each statistic. The second research question was answered by tabulating the number of correct rejections of unidimensionality (or *essential*

unidimensionality) made with each statistic in the presence of two-dimensional response data.. Rejection rates for the approximate χ^2 and for Stout's T statistic were compared in conditions of varying test length, pseudo-guessing, and levels of correlated factors.

A log-linear analysis was used to determine which conditions were significantly associated with rejection decisions made by each statistic. This method has been applied to dimensionality assessment statistics by past researchers (e.g. Gessaroli & De Champlain, 1995). The independent variables were test length, pseudo-guessing and the correlation of the latent traits. The dependent variable was the number of acceptances and rejections of the null hypothesis of unidimensionality. Any predictor was considered to be significant if the z-value associated with the parameter estimate was greater than $|2.0|$. The results for two-dimensional simulated response data are presented after those for data simulated as unidimensional.

Unidimensional Response Data

The approximate χ^2 statistic tests the null hypothesis of unidimensionality whereas Stout's T statistic is a test of the null hypothesis of *essential unidimensionality*. For the purpose of later comparisons, the number of rejections made with each index of dimensionality was considered the Type I error rate when unidmensional response data sets were simulated.

The approximate χ^2 statistic. A significant χ^2 represents a rejection of the null hypothesis of unidimensionality. There were fewer Type I errors with unidimensional data sets having 45 items than with 30 item data sets (see Table 2). Approximately 2.5% of the 200 data sets with 30 items were rejected by the approximate χ^2 test of unidimensionality. There was a Type I error rate of 1% for the 200 data sets simulated with 45 items.

There was a slightly larger number of rejections made with no pseudo-guessing present in the data than when pseudo-guessing was simulated. The Type I error rate for 200 data sets simulated with pseudo-guessing was .5% compared to the 3% error rate found with data sets which contained no pseudo-guessing. The Type I error rate for the approximate χ^2 was below the alpha of .05 in all conditions with unidimensional data.

Stout's T statistic. The original Stout's T statistic did not differ greatly from the revised version of the statistic in the number of rejections of unidimensionality. The number of rejections made using the powerful Stout's T statistic are described here. Results of both the powerful and the original version of Stout's T are presented on Table 2.

Table 2

Unidimensional Response Data: Number Of Rejections Per 100 Data Sets

No Pseudo-Guessing Condition c = 0

TEST LENGTH	APPROXIMATE χ^2 STATISTIC	ORIGINAL STOUT'S T STATISTIC	POWERFUL STOUT'S T STATISTIC
30 ITEMS	4	1	3
45 ITEMS	2	0	0

Pseudo-Guessing Condition $0.25 \geq c \geq 0.05$

TEST LENGTH	APPROXIMATE χ^2 STATISTIC	ORIGINAL STOUT'S T STATISTIC	POWERFUL STOUT'S T STATISTIC
30 ITEMS	1	3	5
45 ITEMS	0	2	7

The null hypothesis of essential unidimensionality was rejected in 3.5% of the 200 data sets simulated with 45 items, and in 4% of the 200 data sets which were simulated with 30 items. Test length was not strongly related to rejection decision for unidimensional data sets with Stout's T statistic.

There were fewer rejections of unidimensionality made for response data which contained no pseudo-guessing than for data which contained pseudo-guessing. The proportions of rejections made by Stout's T statistic were 0.5% for 200 data sets with no pseudo-guessing simulated, and 2.5% for data sets which contained pseudo-guessing. The overall Type I error rate for Stout's T statistic was acceptably low.

Two-Dimensional Response Data

The results for the approximate χ^2 and Stout's T statistic are discussed separately. Rejection frequencies for each simulation condition were tabulated, and a logit linear analysis was performed to indicate which independent variables were significantly associated with rejection decisions. The performance of the approximate χ^2 and Stout's T statistic are compared after the results for each statistic are summarized.

The approximate χ^2 statistic. The logit-linear analysis shows that the number of rejections made with the approximate χ^2 statistic was significantly influenced by the main effects of test length, pseudo-guessing, and the level of correlation of the latent traits with multidimensional response data. No interactions of the independent variables made a significant contribution to the model.

The main effect of level of correlation of the latent traits was quite clear (the logit linear analysis determined $z = 7.22$). The number of (correct) rejections of a unidimensional model made with an approximate χ^2 statistic was consistently lower for highly correlated factors and more rejections were made for weaker or zero correlations. The null hypothesis of unidimensionality was rejected in only 16% of the 400 simulated data sets when $r = .7$. This was much less than the rejection rates of 85% and 98% in cases where $r = .5$ and $r = .0$, respectively.

The effect of test length was also evident ($z = -2.04$). There were more (correct) rejections of unidimensionality when data sets of 45 items were simulated when compared to data sets consisting of 30 items. Unidimensionality was rejected in 69% of the 600 simulations of 45 item data sets, compared to 63% of the 600 simulations of 30 item data sets. The presence of pseudo-guessing in the response data was associated with fewer rejections of unidimensionality for the approximate χ^2 ($z = 3.43$). The null hypothesis of unidimensionality was rejected in 71% of the 600 data sets simulated with no pseudo-guessing, compared with rejections in 62% of the 600 data sets simulated with pseudo-guessing.

Stout's T statistic. A significant T statistic indicates a rejection of essentially unidimensional latent structure for the data set analyzed. The frequency of rejections made by the original Stout's T statistic and the revised Stout's T were very similar in most conditions. To simplify the presentation of results, only the rejections of unidimensionality made using the more powerful Stout's T (revised) statistic are described. Logit linear analysis of the number of rejections of essential unidimensionality made by the powerful Stout's T statistic indicated that test length, pseudo-guessing and the level of correlation of the latent traits were significantly associated with rejection decisions.

Table 3

Two-Dimensional Response Data: Number of Rejections Per 100 Data Sets.

No Pseudo-Guessing Condition, c = 0

TEST LENGTH	CORRELATION OF FACTORS	APPROXIMATE χ^2 STATISTIC	ORIGINAL STOUT'S T STATISTIC	POWERFUL STOUT'S T STATISTIC
30 ITEMS	0	100	92	93
30 ITEMS	.5	88	76	78
30 ITEMS	.7	18	42	46
45 ITEMS	0	100	99	99
45 ITEMS	.5	99	94	95
45 ITEMS	.7	19	61	65

Pseudo-Guessing Condition, $0.25 \geq c \geq 0.05$

TEST LENGTH	CORRELATION OF FACTORS	APPROXIMATE χ^2 STATISTIC	ORIGINAL STOUT'S T STATISTIC	POWERFUL STOUT'S T STATISTIC
30 ITEMS	0	93	70	74
30 ITEMS	.5	69	54	62
30 ITEMS	.7	12	16	18
45 ITEMS	0	99	88	94
45 ITEMS	.5	82	80	82
45 ITEMS	.7	16	32	38

The main effect of correlation of the latent traits ($z = 7.52$) was evident in the greater number of rejections made when the correlation between the latent factors was lower. The null hypothesis of essential unidimensionality was rejected in 90% of 400 data sets simulated with correlations of $r = 0$. Seventy-nine percent of the 400 data sets simulated with latent traits correlated at $r = .5$ were rejected by Stout's T statistic. The lowest proportion of rejections of essential unidimensionality, 42 %, was found with the 400 data sets simulated with correlated traits of $r = .7$.

Rejection decisions made by Stout's T was significantly associated with the number of items simulated in the data sets ($z = -6.12$). The number of rejections generally increased with the main effect of test length for both versions of Stout's T statistic. When there were 45 items simulated 79% of 600 data sets resulted in significant T statistics. The null hypothesis of essential unidimensionality was rejected in a smaller proportion, 62%, of the 600 data sets simulated with 30 items.

The main effect of pseudo-guessing ($z = 6.01$) influenced rejections of the null hypothesis so that fewer rejections were made when pseudo-guessing was simulated in the response data. Sixty-one percent of the 600 data sets simulated with pseudo-guessing resulted in significant T statistics. There were more correct rejections of unidimensionality made with response data which contained no pseudo-guessing. The assumption of essential unidimensionality was rejected in 79% of 600 data sets simulated with no pseudo-guessing present.

Discussion

The research questions have been answered in the above description of the results of the study. The following discussion attempts to situate these findings relative to past research and to consider some possible implications. The unidimensional data simulations will be examined first followed by a presentation of issues relating to two-dimensional data simulations.

Unidimensional Data Simulation

The effects of pseudo-guessing and test length on Type I error rates in this study seem consistent with the findings of past research. Gessaroli and De Champlain (1995) showed with unidimensional simulated response data that the main effect of test length predicted rejections of unidimensionality (Type I errors) made by the approximate χ^2 statistic. There was no clear relationship between rejections of unidimensionality and test length for Stout's T statistic in the present study, which is also consistent with earlier research (De Champlain, 1992; Nandakumar & Stout, 1993).

Pseudo-guessing increased the number of Type I errors made with both the approximate χ^2 and with Stout's T statistic in unidimensional data simulations, although the empirical α was close to 5% in all conditions. The revised version of Stout's T statistic had a higher Type I error rate than did the original version of Stout's T statistic in both the pseudo-guessing and no guessing conditions. It is not clear if this is consistent with the assertion made by Nandakumar and Stout (1993) that the (revised) statistic has been made more robust to the presence of pseudo-guessing.

Pseudo-guessing could be expected to have the effect of increasing model misfit in the tails of the item characteristic curve (ICC). This value is defined as *c*, or the height of the lower

asymptote in the three parameter logistic or normal ogive model. This represents the minimum probability of correctly answering the item for all examinees, notably those at the lower end of the latent trait distribution (McDonald, 1989). It is conceivable that the height of the lower asymptote would be related to both the item format (e.g. guessing purely by chance there is a 25% probability of a correct answer on a multiple choice item with 4 options), and to the characteristics of the examinee (e.g. gender or ability). The slope of the ICC (or discrimination value of the item) may be fairly consistent within a set of item responses, but the height of the lower asymptotes would vary according to the qualities of the item and the examinee described above. Hambleton (1993) has suggested a four-parameter model which takes into account the empirical probability of correct responses at the high end of the latent trait distribution (even those with a high ability level may make errors). Variation in ICC's in the upper and lower asymptotes which is not random may lead to errors in estimation using a two-parameter IRM.

When *c* is set to be non-zero (a value for each item was randomly assigned ranging from c = .05 to c = .25 in this study), the result would be non-random variance in the response data for that item. In the unidimensional case the misfit would be similar to the 'artifact' of an additional dimension found when a linear model is fitted to binary response data. Dimensionality indices sensitive to correlated errors, such as those based on local item dependence would then fail to reject unidimensionality in the presence of pseudo-guessing. This study seems to support this interpretation of the effects of pseudo-guessing on rejections made with the approximate χ^2 in unidimensional data sets.

Nandakumar and Stout (1993) reported some attempts to correct for guessing in DIMTEST. They concluded that "...even with this reduction of guessing levels, the items selected

for the AT1 subtest did not differ significantly from those selected without correction for guessing" (Nandakumar & Stout, 1993. p.51). The present study indicates that Stout's T (falsely) rejects unidimensionality more often when pseudo-guessing is present. However, the Type I error rates still seem to be acceptable.

Two-Dimensional Data Simulation

It is interesting to consider the conditions of simulation in the two-dimensional data sets and their relationship to rejections of unidimensionality. The presence of pseudo-guessing had the effect of decreasing the number of (correct) rejections of unidimensionality made with Stout's T and the approximate χ^2. In this condition, the presence of correlated errors for examinees seems to mask the two-dimensional latent structure. This may explain why there were fewer rejections of unidimensionality when the two-dimensional data contained pseudo-guessing.

The approximate χ^2 and Stout's T both performed well for multidimensional data sets with factors correlated 0 and at .5, but the higher correlation of r =.7 was associated with fewer (correct) rejections of unidimensionality. Highly correlated factors in a two-dimensional latent structure have been linked to poor performance in earlier applications of Stout's T statistic (Nandakumar, 1994). More (correct) rejections of unidimensionality occurred when there was no correlation or a correlation of r = .5. In these conditions the data were simulated with a more distinct two-dimensional structure.

The implications of possible failure to reject unidimensionality (with two-dimensional data having correlated latent factors) with the approximate χ^2 and Stout's T statistic must be considered. For example, in a test of reading ability one general factor (reading comprehension) may influence responses to all items, while subgroups of items may reflect specific factors

(knowledge of the topic in a given reading passage). Good performance on the general trait or latent factor would be a necessary but not a sufficient condition to correctly answer all items. This structure could be described as *compensatory* at the item level, whereas the present study used data simulated as *compensatory* in terms of the total score for examinees.

Two-dimensional data were simulated in this study with a large proportion of items loading on the first factor and a smaller proportion of items determined by the second factor (80:20 dominance structure). An alternative complex compensatory structure is described in Camilli, Wang and Fesq (1995) where a relatively large proportion of items may be determined by the second (and third, etcetera) factors. It is clear that this structure may be expected in many real data sets, particularly those with content dependent subsets of items. Those authors suggest a distinction between *functional dimensionality* and *statistical dimensionality*. Expert review of content should be considered as a necessary aid in interpreting the results of dimensionality indices.

The structure of two-dimensional response data simulated here was chosen in order to extend the research of De Champlain (1992). The parameters were based on ACT English response data in the difficulty and discrimination means and variances. This structure was selected because it did not favor the approximate χ^2 method over Stout's T statistic in De Champlain's (1992) study. The relatively low discrimination values (μ_a= .72) resulted in latent factors loading poorly on the majority of items on the test (weak factor structure). Studies of two-dimensional response data which use higher discrimination values, or a range of values, would result in higher item-trait correlations (Nandakumar & Stout, 1993). This may be expected to result in more (correct) rejections of unidimensionality for the approximate χ^2 statistic.

Systematic study is needed to examine the performance of the approximate χ^2 and Stout's T statistic in a wide range of conditions with simulated and real data. The possible effect of complex compensatory two-dimensional structure (where responses to individual items may be determined by more than one latent trait) on rejections of unidimensionality by the approximate χ^2 and Stout's T statistic is unknown. It is interesting to note that the sample size selected for study here (1000) is considered as 'small' by Stout in the computer program DIMTEST (Nandakumar & Stout, 1993).

Summary

Methods based on nonlinear factor analysis of binary response data seem promising in dimensionality assessment. The approximate χ^2 method, which is based on an examination of the residual correlation matrix after non-linear factor analysis, performed about as well as the popular Stout's T statistic in most conditions of simulated response data studied here. Stout's T statistic was more accurate when longer test lengths were simulated in both the unidimensional and two-dimensional conditions. The presence of pseudo-guessing increased Type I errors with unidimensional data simulation, and was associated with fewer (correct) rejections of unidimensionality with two-dimensional simulated data for both statistics. When two-dimensional structure was simulated with a correlation of $r = .7$ the approximate χ^2 failed to reject unidimensionality more often than Stout's T statistic.

Practitioners must consider a variety of methods, and the problems associated with these methods, when selecting a valid indicator of dimensionality. Inferential methods based on a χ^2 distribution suffer due to limitations associated with larger sample sizes. As stated by McDonald (1995) "...one might almost say that the probability under the restrictive hypothesis is an indirect

measure of sample size" (p 32). For this reason the practitioner may wish to consider a descriptive index of dimensionality. Many such indices are weak in their theoretical foundation or are not effective (Hattie, 1984).

McDonald (1995) points out that some popular descriptive indices (e.g. Akaike's Information Criterion) also depend on the χ^2 distribution as a test of model fit, and may be inadequate for use with real data. He cautiously recommend methods based on some examination of the residual covariance matrix after some form of nonlinear factor analysis has been performed.

Nonlinear factor analysis based on bivariate information in the response data has been compared to methods using information from the higher joint moments of the data. Researchers have found that "...the values of the approximate χ^2 do not differ appreciably from those obtained from other limited or full-information methods" (Gessaroli & De Champlain, 1994, p.24). McDonald's recent study (MacDonald, 1995) confirms those results and he adds that "...an alternative basis for judgment of dimensionality rests on inspection of residual covariances" (p.23).

The advantages of an approximate χ^2 statistic consist of a strong foundation in nonlinear factor analysis, the possibility of confirmatory dimensionality assessment, and a statistical test of the hypothesized dimensionality. The sample size, suspected dimensional structure, and intended use of the test (e.g. stakes of possible decisions about examinees) need to be considered when selecting a dimensionality assessment method. A test of unidimensionality should be the first step before the application of a unidimensional IRM is used to estimate ability or item parameters.

This study provides some evidence for nonlinear factor analysis in determining the dimensionality of binary response data. Research which examines items of complex compensatory two-dimensional structure, with larger sample sizes (2000 or more), and with a variety of

dimension strengths seems overdue. Dimensionality assessment has broad application across disciplines where high stakes decisions rest on specific interpretations of the meaning of scores from a test.

References

Ackerman, T. A. (1985). M2PLGEN: A computer program for generating thetas and response strings corresponding to the M2PL model. Iowa City, Iowa: The American College Testing Program

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

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

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

Camilli, G. (1995). The effects of dimensionality on equating the law school admission test. Journal of Educational Measurement. 32, (1) 79-96.

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

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. Orlando, FLA: Holt, Rinehart, & Winston, Inc.

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

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

De Champlain, A., & Gessaroli M.E. (1991, April). Assessing test dimensionality using an index based on non-linear factor analysis. Paper presented at the meeting of the American Educational Research Association, Chicago, IL.

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

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

Gessaroli, M.E., & De Champlain, A. F. (1995). Assessing test dimensionality using an approximate χ^2 statistic. Manuscript submitted for publication.

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

Hattie, J. (1984). An empirical study of various indicies 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.

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

McDonald, R.P. (1967). Nonlinear factor analysis. Psychometrika Monograph no. 15, 32 (4, Pt.2).

McDonald, R.P. (1981). The dimensionality of tests and items. British Journal of Mathematical and Statistical Psychology, 32, 212-228.

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 Educational 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 of measurement: problems and issues, p.63-86. Ottawa, Canada: Edumetrics Research Group, University of Ottawa.

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

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

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

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

Oshima T.C., & Miller M.D. (1992). Multidimensionality and item bias in item response theory. Applied Psychological Measurement, 16, 237-248.

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. (1986). The interpretation of unidimensional IRT parameters when estimated from multidimensional data. Paper presented at the meeting of the Psychometric Society, Toronto, Ontario.

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.

Wilson, D.T., Wood, R. & Gibbons, R.T. (1984) TESTFACT: Test scoring, item statistics, and factor analysis. Pub. Scientific Software, Mooresville, IN.

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

Appendix A

The Approximate χ^2 Test of Two Dimensions

Although not directly posed as a research question, a two-dimensional χ^2 test of fit was available in the program NOHARM II, and was applied to the data sets simulated for this study. The results of this version of the approximate χ^2 are presented briefly here. In this case the null hypothesis of two-dimensional structure is tested, and when applied to two-dimensional simulated response data the number of rejections is the Type I error rate for the statistic.

As can be seen in Table 4, the total number of (false) rejections was not more than 5% for any data set in all levels of variables. The proportion of data sets where the two-dimensional null hypothesis was rejected was 1.33% in the 1200 data sets simulated. There were no rejections of two-dimensional structure made when tests were 45 items.

There was no apparent difference in the frequency of rejections associated with the presence of pseudo-guessing in the response data. Type I errors occurred when data sets were simulated with 30 items and had latent traits that were correlated at $r = .5$ or $r = .7$. Generally, rejections made by the two-dimensional χ^2 were well within an acceptable range for empirical Type I errors in all conditions.

Table 4

Test of Two Dimensions With the Approximate χ^2 Statistic, Number of Rejections Per 100 Data Sets.

No Pseudo-Guessing Condition c = 0

TEST LENGTH	CORRELATION OF FACTORS	APPROXIMATE χ^2 STATISTIC
30 ITEMS	0	0
30 ITEMS	.5	5
30 ITEMS	.7	4
45 ITEMS	0	0
45 ITEMS	.5	0
45 ITEMS	.7	0

Pseudo-guessing condition $0.25 \geq c \geq 0.05$

TEST LENGTH	CORRELATION OF FACTORS	APPROXIMATE χ^2 STATISTIC
30 ITEMS	0	0
30 ITEMS	.5	4
30 ITEMS	.7	3
45 ITEMS	0	0
45 ITEMS	.5	0
45 ITEMS	.7	0