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.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

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.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

Canada

ASSESSING TEST DIMENSIONALITY USING TWO APPROXIMATE CHI-SQUARE STATISTICS

By

André F. De Champlain

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

Ottawa, Canada, 1992

© Andre De Champlain, Ottawa, Canada, 1992

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

The author has granted an irrevocable non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of his/her thesis by any means and in any form or format, making this thesis available to interested persons.

L'auteur a accordé une licence irrévocable et non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de sa thèse de quelque manière et sous quelque forme que ce soit pour mettre des exemplaires de cette thèse à la disposition des personnes intéressées.

The author retains ownership of the copyright in his/her thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without his/her permission.

L'auteur conserve la propriété du droit d'auteur qui protège sa thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

ISBN 0-315-80060-7

Canada

UNIVERSITÉ D'OTTAWA
UNIVERSITY OF OTTAWA

ACKNOWLEDGMENTS

This dissertation was completed under the supervision of Dr. Marc E. Gessaroli, Associate professor, Faculty of Education, Measurement and Evaluation concentration, University of Ottawa. His patience, suggestions and availability were greatly appreciated throughout the past three years. I would like to take this opportunity to thank him for not only being there as an educator but also as a friend.

I am also grateful to Dr. Marvin Boss and Bruno Zumbo for their helpful comments and criticisms. Their input was invaluable and contributed a great deal to this dissertation.

I would also like to thank the following individuals for their help regarding computer programs used in this research project: Chris Carruthers at the Computing Center of the University of Ottawa; Ratna Nandakumar, University of Delaware, for providing the software needed to compute Stout's T statistic and Martin Berger, University of Twente, the Netherlands, for his help with NOHARMII.

Also, thanks to other graduate students in Lamoureux 333 and those crazy Psych. girls: Leanne, Veronica, Celyne and others. You guys are great!

A very, very special thank you and hug to the Jersey girl for her support.

And, yes Mom and Dad, I'm finally getting a job.

CURRICULUM STUDIORUM

Andre F. De Champlain was born in Ottawa, Ontario, Canada, May 9th, 1963. He completed his B.A. (Honours-Psychology) and M.A. (Education-Measurement and Evaluation) at the University of Ottawa.

TABLE OF CONTENTS

LIST OF TABLES

APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Statement of the Problem

The numerous studies dealing with Item Response Theory (IRT) that have dominated the measurement literature in the past decade attest to its importance in the development and analysis of tests and items. The many advantages of IRT models, namely that they provide sample-free item parameter estimates, test-free ability estimates as well as the possibility of supplying the test developer with information pertaining to a wide range of examinee abilities, have generated considerable interest in the area of educational testing. Following earlier work published by Lord (Lord, 1952; Lord, 1980; Lord & Novick, 1968), who is generally acknowledged to be one of the main contributors in this area, researchers have continued to improve the various IRT models as well as apply them in a host of situations. Indeed, IRT models are currently being utilized by large test publishers (Yen, 1983) as well as departments of education (Pandey & Carlson, 1983) for a variety of purposes such as norm- and criterion-referenced test development, test equating, the detection of differential item functioning, etc,. Warm (1978) summarizes the importance of IRT as follows:

> "Item Response Theory (IRT) is the most significant development in psychometrics in years. It is, perhaps, to psychometrics what Einstein's relativity is to physics. I do not doubt that during the next decade it will sweep the field of psychometrics". (p.11).

The widespread application of IRT, however, has been

hindered by the many strong assumptions and principles underlying the majority of the models, namely local independence of examinee responses as well as unidimensionality of the latent space. Briefly stated, local independence entails that for subsets of examinees of identical ability, the (conditional) distributions of item scores are statistically independent.

The unidimensionality assumption implies that item response probabilities are a function of a single latent trait (Hulin, Drasgow, & Parsons, 1983). Most IRT models assume that response probabilities can be estimated in a unidimensional latent space. Unidimensionality is, however, rarely met in practice (Traub, 1983). A mathematics test, for example, might entail not only mathematical ability but also the capability to read and understand the problems being presented. With regards to this issue, Lord (1980) states that:

> "It seems plausible that tests of spelling, vocabulary, reading comprehension, arithmetic reasoning, word analogies, number series, and various types of spatial tests should be approximately one-dimensional. We can easily imagine tests that are not. An achievement test in chemistry might in part require mathematical training or arithmetic skill and in part require knowledge on nonmathematical facts". (Lord, p.20).

McDonald (1981) has suggested that the principle of local independence be replaced by the following more tenable, "weak" principle of local independence,

$$Cov(U_j, U_k|\theta) = 0 \quad j \neq k \tag{1}$$

that is, item response probabilities are a function of a single

latent trait that would entirely explain covariances among all item pairs. McDonald (1981) states that the presence of a single latent trait can be assumed only when the "weak" principle of local independence has been met for a particular set of item responses. In other words, the latent trait is defined by this "weak" principle of local independence.

This type of consideration has led researchers to question the accuracy of estimates derived from unidimensional models when applied to multidimensional data sets. Authors who have estimated the robustness of IRT item and ability parameter estimates obtained from multidimensional data have generally shown that these multidimensional values are poorly recovered by unidimensional models, most notably, when several equally important abilities are required to correctly answer an item (Ackerman, 1987; Ansley & Forsyth, 1985; Drasgow & Parsons, 1983; Reckase, 1979; Reckase, Carlson, Ackerman & Spray, 1986).

These results have led to the development of statistical techniques to assess test dimensionality or, more commonly, departure from the assumption of unidimensionality. The majority of the research in this field has focused primarily on the evaluation and/or development of indices based on principal component analysis(PCA) / linear factor analysis(LFA), the Holland-Rosenbaum procedure, Stout's essential dimensionality and residual covariance analyses. The two most popular approaches at present appear to be Stout's work on essential dimensionality as well as the analysis of a residual covariance matrix after

fitting a nonlinear (one-)factor analytic model.

Stout's procedure is based on his new definition of dimensionality, *essential dimensionality* (Stout, 1987). He argued that it is unrealistic to believe that a test can truly be unidimensional (i.e., zero conditional residual covariances between pairs of items after fitting a one-factor model), a fact previously hinted at by Lord & Novick (1968). *Essential dimensionality* corresponds to the number of dimensions necessary to satisfy the assumption of *essential independence* (i.e., the mean residual conditional covariance which tends towards zero as the number of items increases to infinity). The assumption of essential independence is then tested using the T statistic that Stout (1987) developed. In addition, Nandakumar (1987) proposed a correction method for the procedure used to calculate the T statistic in order to reduce bias due to homogeneous item difficulties ("easy items") being solely retained in the item assignment step. Results indicate that the T statistic appears to be able to accurately determine if correctly responding to a set of item responses requires *essentially* one or more than one ability (Stout, 1987), especially when Nandakumar's modification is utilized (Nandakumar, 1987; 1988; 1989). However, the power of the statistic seems to decrease with shorter test lengths and smaller sample sizes which prompted Nandakumar to advise against using the T statistic, in some instances, with less than 25 items and 750 examinees (Nandakumar, 1987).

Another approach quickly gaining popularity is one that

treats IRT as a special case of nonlinear factor analysis (see McDonald, 1967, for some of the first work in this area). Takane and De Leeuw (1987) have shown that some of the models used in IRT and nonlinear factor analysis are mathematically equivalent, a point previously alluded to by McDonald (1967). Using this IRT-factor analysis relationship, some authors have suggested that the most logical method of assessing dimensionality would have to be based on an analysis of the residual covariance matrix after some type of nonlinear factor analysis (Hambleton & Rovinelli, 1986; Hattie, 1984; McDonald, 1989). Indeed, as was previously stated, unidimensionality of the latent trait would theoretically imply zero residual covariances between all pairs of items at fixed ability levels (i.e., the "weak" principle of local independence). Results show that various indices such as the mean absolute residual covariance (Berger & Knol, 1990; Hattie, 1984) as well as the incremental fit index (IFI) (De Champlain & Gessaroli, 1991) tend to be related to the number of dimensions underlying a set of test items. However, these types of indices possess the following major disadvantage: they are purely descriptive in nature and hence offer no objective criteria on which to assess the number of traits underlying a set of item correlations. These results led Goldstein and Wood (1989) to state that:

> "A number of alternative procedures for judging unidimensionality have been proposed, but these tend to be statistically unsound, lack sensitivity or adopt an unsatisfactory definition of unidimensionality. The factor analytic approach of McDonald (1982) provides one of the most useful formulations of this problem". (p.151).

Future research in this area should therefore focus on the development and testing of an inferential statistic that would allow researchers to determine, with greater confidence, the actual number of abilities underlying a set of item responses with various test structures.

1.2 Purpose of the Research

The purpose of the present study is to investigate two approximate χ^2 statistics that are based on the "weak" principle of local independence and therefore might possibly be used to determine if one or more than one latent trait underlies a set of item responses. The first approximate χ^2 statistic was initially proposed by Bartlett (1950) and recently outlined by Steiger (1980a; 1980b) whereas the second is an original contribution of this research. Both approximate χ^2 statistics test the null hypothesis that the off-diagonal elements of a residual correlation matrix are equal to zero. The rationale underlying the use of these test statistics lies in the assertion that the off-diagonal elements in a matrix of residual correlations after a nonlinear factor analysis should theoretically be zero if the correct number of factors (dimensions) are specified in the model. Of course, in practice this is rarely the case. However, it may be possible to simultaneously test whether all the

residual correlations are significantly different from zero. The advantages of these procedures over existing techniques are threefold:

1. The assessment of dimensionality is based on a general model on which common IRT models are derived (nonlinear factor analysis).
2. The procedures involve actual hypothesis testing and are not merely descriptive indices. Hence, their values are not intrinsically linked to specific parameters used in a study (sometimes with little relationship to actual achievement test data) and they (possibly) can be used for a variety of data sets with greater confidence.
3. The statistics can perhaps be used to assess not only departure from unidimensionality but also the fit of successively more complex models, i.e., two-, three-, etc. dimensional structures to a set of item responses.

PURPOSES OF THE STUDY

The purposes of this study are twofold:

(1) Examine Type I error rates obtained with the two approximate χ^2 statistics, the χ^2 difference tests and Stout's T-statistic with unidimensional test structures.

(2) Examine the power of the approximate χ^2 statistics, the χ^2 difference tests and the T-statistic in rejecting the assumption of unidimensionality with two-dimensional test structures.

CHAPTER 2

THEORETICAL FRAMEWORK

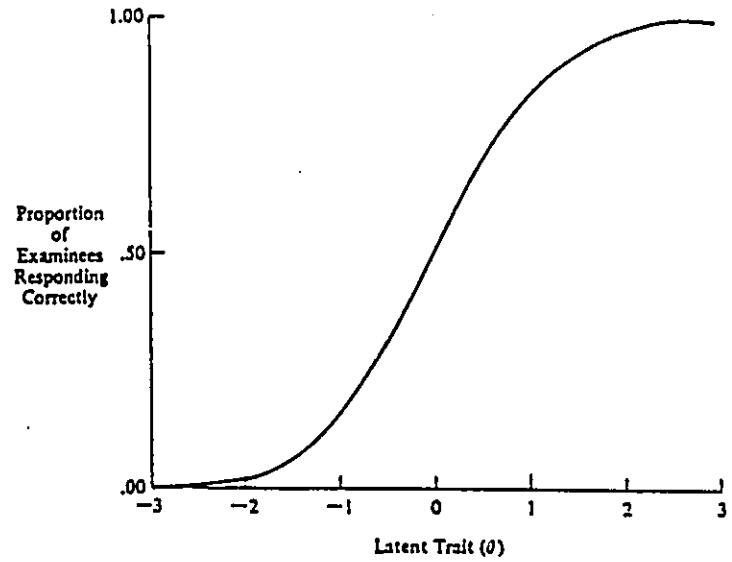
In this chapter, the theoretical framework underlying the research will be presented. First, the two-parameter logistic IRT model will be briefly outlined. Second, the relationship between logistic IRT models and nonlinear factor analytic models will be closely examined. Third, McDonald's nonlinear factor analytic model will be presented as an example of a general model on which common IRT logistic approaches are based. Fourth, the principle of local independence and assumption of unidimensionality underlying the latter IRT models will be explicated. Fifth, the effects of violating the assumption of unidimensionality with respect to the estimation of item and ability parameters derived from a unidimensional logistic IRT model will be examined. Finally, procedures that have been proposed to assess the assumption of unidimensionality will be presented. More precisely, indices and/or statistics based on the following techniques will be outlined: linear factor analysis/principal component analysis, nonmetric multidimensional scaling, Tucker's procedure, Humphrey's procedure, modified parallel analysis, order analysis, Bejar's procedure, the Holland-Rosenbaum procedure, Stout's essential dimensionality procedure and nonlinear factor analysis.

2.1 The two-parameter logistic IRT model

Item response theory models hypothesize that the performance of an examinee on a set of (dichotomous) test items can be explained or predicted in terms of one or several (unobservable) latent traits. In the context of educational measurement, these latent traits are usually considered to be various abilities that are required to correctly answer items on a test. For example, the probability of correctly answering a multiple-choice item on the GRE verbal scale is assumed to be a function of a hypothesized latent trait that we refer to as "*verbal ability*". At this point it is important to re-emphasize the fact that these latent trait values cannot be directly measured and must be estimated using observed item scores. Therefore, at the core of item response theory is a mathematical model which specifies the relationship between the observable examinee item performance and the unobservable latent trait (ability) underlying test performance. This relationship can be detailed using several mathematical functions. One popular model that will be more closely examined is the two-parameter logistic model proposed by Birnbaum (1968).

The two-parameter logistic model (as well as other IRT models) hypothesizes that the relationship between observable item performance and the underlying latent trait can be depicted graphically in an *item characteristic curve* (ICC). Specifically, an ICC plots the probability of correctly answering an item given a certain ability score denoted by θ. An example of an ICC

specified by a two-parameter logistic model is presented below (taken from Crocker & Algina, 1985; p. 347).



In an ICC, the abscissa corresponds to points along the ability scale whereas the ordinate indicates the probability of correctly answering a given item. It is important to point out that the item response function can be plotted with an ICC when only one ability underlies performance on a set of test items. When more than one ability is required to correctly answer test items, the relationship is plotted in a multidimensional item response function (IRF). In addition, the general "shape" of the ICC is dependent upon the mathematical function used. The use of a two-parameter logistic model will result in a monotonically increasing (logistic) ogive curve. However, changing the mathematical function will result in a different ICC. In fact, Hambleton & Cook (1977) have looked at the ICCs obtained with seven different mathematical functions (Guttman scale, latent

linear curves, logistic curves, etc.). McDonald (1982b) has also shown that several item response models can be used, each resulting in its own distinctive ICC or IRF.

The probability of obtaining a correct response on a given dichotomous item *i* can be represented by the following two-parameter logistic model:

$$P_i(i=1|\theta) = \frac{e^{Da_i(\theta-b_i)}}{1+e^{Da_i(\theta-b_i)}} \quad (2)$$

Hence, the probability that a randomly selected examinee of ability level θ will correctly answer an item is dependent upon two item parameters, *a* and *b*.

The first item parameter, *b*, is usually interpreted as an item difficulty parameter. The value of the difficulty parameter corresponds to the ability score where the slope of the ICC is at its maximum, i.e., at the inflexion point. It may also, for the one- and two-parameter models only, be interpreted as the location along the ability scale at which an examinee has a probability of .50 of correctly answering an item. Theoretically, the item difficulty parameter can assume any value ranging from $-\infty$ to $+\infty$. However, in practice, values typically vary from -2 to +2 (Hambleton & Swaminathan, 1985). A low b estimate (for example, from -2 to 0) is indicative of an easy item whereas a high value (for example from 0 to +2) would suggest that the item is difficult.

The second item parameter, *a*, corresponds to the item discrimination value. The discrimination parameter corresponds to

the value of the slope of the ICC at its point of inflexion. Item discrimination values can also, theoretically, range from $-\infty$ to $+\infty$. However, in practice, items with negative discrimination values are discarded given that this would indicate that low ability examinees have a higher probability of correctly answering an item than high ability examinees. Therefore, with actual achievement data, discrimination parameter values are typically found in the interval (0,+2) (Hambleton & Swaminathan, 1985). High values of *a* would result in a very steep ICC whereas low values of *a* would result in a "flat" ICC. Finally, an item having a low *a* value would discriminate poorly over a wide range of abilities whereas with a high *a* value, the item would discriminate well, but over a small range of abilities.

As previously mentioned, the ability level of an examinee is denoted as θ. Again, ability can be theoretically defined on the scale $(-\infty, +\infty)$. In practice, however, most ability scores are usually found in the interval ranging from -3 to +3.

Finally, two constants are included in the two-parameter logistic function. The first, *D*, is a scaling factor that is used to approximate a two-parameter normal ogive model. A value of 1.7 is usually assigned to this factor. The second constant, *e*, corresponds to the natural logarithm and is approximately equal to 2.71828.

McKinley (1983) and McKinley and Reckase (1983) have also proposed a multidimensional extension of the two-parameter logistic IRT model. This mathematical function, which enables the

estimation of the probability of a correct response given more than one ability, can be represented as follows:

$$P(x_{ij}=1|\underline{\theta}_j)=\frac{e^{(d_i+\sum_{k=1}^{m}a_{ik}\theta_{jk})}}{1+e^{(d_i+\sum_{k=1}^{m}a_{ik}\theta_{jk})}} \quad (3)$$

where x_{ij} is the response to item i by person j;

$\underline{\theta}_j$ is the ability vector for person j;

θ_{jk} is the ability parameter for person j on dimension k;

a_{ik} is the discrimination parameter for item i on dimension k;

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

This model is said to be *compensatory* given that ability on the first dimension can "compensate" for lack of ability on the second dimension. Indeed, a small value of θ_{j1} could be "compensated" for by a larger value of θ_{j2}.

With a two-dimensional compensatory model, Reckase (1985; 1986a) has provided multidimensional analogues of the usual discrimination and difficulty parameters. Specifically, the multidimensional discrimination parameter can be defined as follows:

$$MDISC=\sqrt{a_1^2+a_2^2} \quad (4)$$

Finally, the multidimensional difficulty parameter can be defined

as follows:

$$MDIF = \frac{-d}{MDISC} \quad (5)$$

where d has previously been described as a scalar related to multidimensional difficulty.

The function specifying the relationship between the probability of correctly responding to an item and ability level is plotted in a three-dimensional item response surface. Ackerman (1987) provides excellent examples of item response surfaces that can be obtained with various multidimensional data sets. Although this model and other (multidimensional) logistic models do show promise for the estimation of abilities in a multidimensional space, they are rarely used in practice due mainly to their unfamiliarity to most practitioners and until recently, the lack of available computer programs. Indeed, unidimensional models, such as the previously presented two-parameter logistic IRT model, are still mainly used by researchers who wish to analyze a set of test items and estimate examinee abilities.

2.2. The relationship between IRT and nonlinear factor analytic models

An approach which is currently gaining popularity in educational measurement is the one that treats item response theory as a special case of nonlinear factor analysis. Several authors have shown that these models are mathematically equivalent (Goldstein & Wood, 1989; McDonald, 1989a; McDonald, 1991) (see McDonald, 1967, for some of the first work in this area). Muthen (1978, 1983, 1984) has also shown that commonly used models in IRT (e.g. the two-parameter normal ogive model) are really specific cases of a more general factor analytic model for categorical variables with multiple indicators (i.e. items). McDonald (1982b), starting from Spearman's common factor model, also shows that IRT models are a special case of nonlinear factor analysis and provides a general framework which includes unidimensional/multidimensional, linear/nonlinear models as well as dichotomous and polychotomous models.

Bartholomew (1983) has also provided a general latent class model on which several IRT models as well as factor analytic models for dichotomous variables are founded. This general latent class model is of the form,

$$G(\pi_i(\mathbf{(y)})) = \alpha_{i0} + \sum_{j=1}^{q} \alpha_{ij} H(y_j), \; i = 1, 2 \ldots, p. \qquad (6)$$

In the case of a unidimensional IRT model, the parameters in equation 6 would correspond to the following:

$G(\pi_i)$ = the response function outlining the probability of obtaining a correct response to item i;

$(\mathbf{y})$ = a vector of ability (in this case, a scalar, given that q=1);

α_{i0} = the difficulty parameter of item i;

α_{ij} = the discrimination parameter item i on latent trait j;

$H(y_j)$ = The density function for a given latent trait j.

Takane and De Leeuw (1987) have also shown that IRT models as well as nonlinear factor analytic models are mathematically equivalent. These authors have provided a systematic series of proofs that show the equivalence of these models with dichotomous as well as polychotomous models. Mathematical proofs of the equivalence of these two models when faced with binary item responses are presented in Appendix A (n.b.: these are taken directly from Takane & De Leeuw, 1987 and reproduced for the reader's benefit).

Thus, it appears as though IRT and nonlinear factor analytic models represent two equivalent formulations of the same general (latent class) model. Indeed, the two terms are often used interchangeably. For example, the model proposed by Bock (1984) has been synonymously referred to as full-information factor analysis (Bock, Gibbons, & Muraki, 1988) and multidimensional IRT (McKinley, 1988). Given the equivalence of IRT and nonlinear

factor analytic models, it would appear reasonable to make use of the latter models to examine a multitude of educational measurement problems which had been, until quite recently, looked at solely with IRT models. One nonlinear factor analytic model which has promising applications in the field of educational measurement is McDonald's polynomial approximation to a normal ogive (McDonald, 1967; 1982b). Indeed, it has been used in the recent past to examine problems particular to the area of educational testing (Miller & Hirsch, 1991; De Champlain & Gessaroli, 1991). The next section of the second chapter of this dissertation will briefly outline McDonald's model for the analysis of dichotomous item responses.

2.3. McDonald's nonlinear factor analytic model

McDonald (1967; 1982a; 1982b, 1989; 1991) and McDonald and Ahlawat (1974) have provided a general framework that enables the organization of existing unidimensional as well as multidimensional IRT models based on a more general nonlinear factor analytic approach. This framework also enables the researcher to generate many new models. Specifically, generalizing from Spearman's common (unidimensional) factor model, McDonald (1982b) has presented the following three classes of models which could be used in educational measurement:

i. models that are strictly linear (in both their coefficients and latent traits),

ii. models that are linear in their coefficients but not in their latent traits,

iii. models that are strictly nonlinear.

i. Models that are strictly linear

The first type of model discussed by McDonald (1982b) is both linear in its coefficients and latent trait(s). This would correspond to Spearman's common single factor model,

$$\hat{y}_g = \mathbf{E}(y_g|\theta=\theta_a) = f_g\theta_a + m_g \qquad (7)$$

where y_g = the conditional mean of y_g;

θ = the single common factor (or latent trait);

θ_a = any fixed value of the common factor;

f_g = the regression coefficient of y_g on the common factor;

m_g = a uniqueness component unexplained by the common factor.

ii. Models that are linear in their coefficients but nonlinear in their latent trait(s)

McDonald and Ahlawat (1974) have also proposed a group of regression functions that are linear in their coefficients (i.e. their item parameters) but nonlinear in their traits, of the general form,

$$f_j(x_1,\ldots,x_t) = a_{j0} + \sum_{l=1}^{t}\sum_{p=1}^{s} a_{jlp}h_p(x_l)\ (j=1,\ldots n) \qquad (8)$$

where,

$f_j(x_1,\ldots,x_t)$ = a function that represents the probability that an examinee with latent trait values $x_1,\ldots,x_t$ will correctly respond to the jth binary item;

a_{j0} = An intercept parameter of the regression function for item j;

a_{jlp} = A regression coefficient for item j on latent trait l of the p-th polynomial degree;

$h_p(x_l)$ = a general polynomial function of the form,

$$f_{j1}\theta + f_{j2}\theta^2 + \ldots + f_{jk}\theta^k. \qquad (9)$$

iii. Models that are strictly nonlinear

The final class of models discussed by McDonald (1982b) corresponds to the familiar normal ogive and logistic models commonly used in IRT. For example, the two-parameter normal ogive model (Lord, 1952; 1953) can be presented as follows:

$$P_i(\theta) = \int_{-\alpha}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \qquad (10)$$

where z is often substituted by L, that is, a logistic function involving two item parameters, $a_g(\theta - b_g)$ and where t is equal the normal frequency function.

These models are considered strictly nonlinear, that is, in their latent trait(s) as well as their coefficients.

iv. A final word on the three classes of models

Starting from Spearman's general factor analytic model, McDonald (1982b) has provided a broad framework that allows the researcher to classify current IRT as well as nonlinear factor analytic models. His unique and significant contribution to the area, however, lies with the second class of models presented above, that is, a series of functions linear in the coefficients but nonlinear in the latent trait(s). McDonald's model is outlined below.

As was previously hinted at, McDonald (1967; 1982a; 1982b, 1989; 1991) has elaborated a nonlinear factor analytic model that can be used, among other things, to analyze dichotomous test items. Specifically, this author has proposed a polynomial approximation to a normal ogive model. As McDonald (1991) states, a common parameterization (in the unidimensional case) of function z for item *j* on latent trait *i* in equation (10) is

$$z_{ji}=a_j(\theta_i-b_j) \qquad (11)$$

However, alternative parameterizations have been proposed such as,

$$z_{ji} = (\lambda_{j0} + \lambda_{j1}\theta_i) / (1 - \lambda_j^2)^{1/2} \quad (12)$$

as well as

$$z_{ji} = f_{j0} + f_{j1}\theta_i \quad (13)$$

McDonald (1991) has also shown that these two forms can be generalized to the following multidimensional models,

$$z_{ji} = (\lambda_{j0} + \underline{\lambda}_j'\underline{\theta}_i) / (1 - \underline{\lambda}_j'\Phi\underline{\lambda}_j)^{1/2} \quad (14)$$

$$z_{ji} = f_{j0} + \underline{f}_j'\underline{\theta}_i \quad (15)$$

where $\underline{f}_j$, $\underline{\lambda}_j$ and $\underline{\theta}_i$ are k-component vectors; k being the number of dimensions. In fact, McDonald (1991) states that substituting function 14 in equation 10 would yield Christoffersson's factor analytic model for dichotomized variables (Christoffersson, 1975). Replacing function 15 in equation 10 gives McDonald's polynomial approximation to a normal ogive model. McDonald (1967) has shown, using harmonic analysis, that the normal ogive could be approximated by a polynomial series of the general form,

$$z_{ji} = f_{j0} + f_{j1}\theta + f_{j2}\theta^2 + \ldots + f_{jk}\theta^k \quad (16)$$

For example, a one-factor (or unidimensional) cubic model can be written as

$$Y_i = b_{i0} + b_{i1}\theta + b_{i2}\theta^2 + b_{i3}\theta^3 + e_i \quad (17)$$

where,

b_{ijk} = The factor loading of factor j on item i of polynomial degree k;

e_i = The uniqueness component of item i.

McDonald (1982b) has shown that this linear model (in its coefficients) provides a very good approximation to the normal ogive model. Nonetheless, some authors have noted that a major problem with the model lies in the absence of an index that would indicate the appropriate number of polynomials to retain in a series (Hambleton & Rovinelli, 1986). Recent work in this area, however, seems to indicate that terms beyond the cubic can generally be disregarded (McDonald, 1982b, Nandakumar, 1991b). Having said this, several researchers have stated that McDonald's model appears to be one of the most promising approaches for the analysis of educational test items as well as problems typically encountered in educational measurement, such as those related to the assessment of dimensionality, which will be presented in the next section of the chapter (Goldstein & Wood, 1989; Hambleton & Rovinelli, 1986; Hambleton & Swaminathan, 1985; Hattie, 1984).

2.4. The principle of local independence and assumption of unidimensionality

The majority of IRT models require that examinee item responses be statistically independent and the latent (ability) space be unidimensional. The principle of local independence and assumption of unidimensionality are outlined below.

(a) Local independence

This principle, initially stated by Lazarsfeld (1950), entails that responses provided by an examinee to a set of items are statistically independent. In other words, the conditional distributions of the item scores are independent (Lord & Novick, 1968). Local independence generally implies that:

$$P(u_1=1, u_2=1, \ldots, u_n=1|\theta) = \prod_{i=1}^{n} P(u_i=1|\theta) \qquad (18)$$

Therefore, at any given value of θ, the probability of correctly answering all items on a test is equal to the product of the separate probabilities of correctly answering each item. Lord and Novick (1968) point out, however, that local independence does not imply that item responses are uncorrelated for the total group of examinees. Items will be correlated to a certain extent when examinees vary with regards to the ability being measured by the test. However, at a fixed ability level, the correlations between pairs of items are expected to be equal to zero.

(b) Unidimensionality

According to the classical definition of dimensionality provided by Lord and Novick (1968), if there are *k* traits (referred to as latent variables) influencing examinee performance on *n* items in a test, i.e., each of the traits influences performance on at least one item in the test, then *k* is the dimensionality of the latent space. The latent space is thus said to be complete when all traits affecting the test scores of a population of examinees have been delineated. Consequently, dimensionality would be viewed as the number of dimensions required to satisfy the assumption of local independence. In other words, a test is said to be unidimensional if only one ability is required to satisfy the assumption of local independence. However, the equivalence of these two assumptions is a contentious issue that has been seriously challenged by several authors (Goldstein, 1980; Goldstein & Wood, 1989; McDonald, 1979; 1981; 1982a; 1991). These arguments will be presented shortly. Most IRT models assume that only one latent ability underlies performance on a set of test items. In practical testing situations, however, this assumption is rarely met. Indeed, the multitude of factors that can affect the performance on a given item, some of which are test-taking factors such as anxiety, motivation, as well as cognitive skills other than the one (theoretically) underlying item responses, will invalidate this assumption in the majority of testing situations. Some authors have therefore suggested that a more

realistic definition of unidimensionality, reflecting the fact that a test typically requires knowledge of a major ability and possibly minor abilities, be adopted.

Stout (1987; 1990) has provided a less restrictive formulation of both local independence and unidimensionality. This author states that *essential dimensionality* corresponds to the number of dimensions that are required to satisfy the assumption of *essential independence*. A test consisting of items U_j, (j=1,...,N) of length N is said to be *essentially unidimensional* if there exists a latent trait θ such that for all values of θ,

$$\frac{1}{N(N-1)} \sum_{1 \le i \ne j \le N} |Cov(U_i, U_j|\theta)| = 0 \qquad (19)$$

Thus, Stout proposes that a test is *essentially unidimensional* if the mean residual absolute covariance tends towards zero, as the number of items increases to infinity, after partialling out the effect of the common factor.

McDonald (1981) has also proposed an alternative view of unidimensionality which rests on the "weak" principle of local independence. According to this author, in practice it is probably more realistic to specify the dimensionality k that is sufficiently small to satisfy the "weak" principle of local independence, that is,

$$Cov(U_j, U_k|\theta) = 0 \quad j \ne k \qquad (20)$$

The advantage of McDonald's conceptualization of unidimensionality is that it allows the practitioner to determine if one or more than one latent trait underlies a set of item responses based on a falsifiable principle, that is, "weak" local independence. In other words, showing that a set of (conditional) residual correlations is equal to zero after fitting a one-factor model would lead the researcher to conclude that one latent trait is required to correctly answer a set of test items.

(c) A final comment on the principle of local independence and assumption of unidimensionality

As was previously presented, the principle of local independence entails that responses are statistically independent for a group of examinees at a certain ability level. The assumption of unidimensionality, on the other hand, entails that only one latent trait underlies performance on a set of test items. Several researchers have stated that these assumptions are distinct, i.e., a unidimensional model does not necessarily explain all observed item dependencies as was incorrectly pointed out by Lord and Novick (1968) (Goldstein, 1980; Goldstein & Wood, 1989; McDonald, 1981; 1982a; 1982b; 1991). The latter author points out that the (strong) assumption of local independence is not falsifiable since it would entail that the latent trait explain not only item covariances but also their higher joint moments. Hence, it is usually substituted for the (falsifiable) assumption of "weak" local independence presented previously. However, a recent full-information factor analytic model (Bock,

Gibbons, & Muraki, 1988), which analyzes response vectors instead of item pairs, might lead to the development of dimensionality assessment methods based on the "strong" principle of local independence.

2.5. The effects of violating the assumption of unidimensionality on the estimation of item and ability parameters

In this section of the second chapter, a review of studies that have examined the effect of estimating unidimensional item and ability parameters when obtained from multidimensional data sets will be outlined.

As was previously reported, the assumption that only one latent trait (usually, a particular ability) underlies performance on a set of test items is one of the critical postulates associated with the use of the majority of IRT models. Hence, one of the first problems that was examined by several researchers in this field was the robustness of unidimensional item and ability estimates obtained from multidimensional data.

Results obtained from these studies generally caution against the use of unidimensional models with multidimensional data sets (Hsu & Yu, 1989). Nevertheless, some researchers have reported that unidimensional models could be used to estimate item and ability parameters from multidimensional response strings only when the correlation between abilities was moderate to high (generally, above .4) (Doody, 1985; Drasgow & Parsons, 1983; Harrison, 1986). In fact, unidimensional item and ability

estimates appear to be fairly accurate when a dominant first factor is present (Drasgow & Parsons, 1983; Harrison, 1986; Reckase, 1979). On the other hand, these unidimensional estimates are not robust when several equally important abilities are required to correctly answer a set of items (Doody, 1985; Drasgow & Parsons, 1983; Harrison, 1986). Most unidimensional estimates appear to be a function of the original multidimensional item and ability parameters, dependent upon the particular model used to generate the data and the characteristics of initial (multidimensional) parameter values (Ackerman, 1987; Reckase, 1986b). Using a compensatory model, some researchers have noted that unidimensional difficulty, discrimination and ability estimates tended to correspond to the mean of their true difficulty, discrimination and ability values (Ansley & Forsyth, 1985). Others have shown, using a noncompensatory model, that the unidimensional discrimination parameter estimate corresponded to the sum of the true values (i.e. a_1, a_2) whereas the unidimensional difficulty and ability estimates appeared to be equal to mean of their respective true values (Way, Ansley, & Forsyth, 1988). Finally, some researchers have reported, again using a noncompensatory model, that unidimensional difficulty parameter estimates were consistent overestimates of the true value on the first dimension (i.e. b_1) whereas discrimination and ability estimates appeared to be the averages of their respective true values (Ansley & Forsyth, 1985; Way, Ansley, & Forsyth, 1988).

Based on these studies, what can we conclude with regards to the robustness of unidimensional item and ability parameter estimates of multidimensional data? In the majority of the studies having examined this problem, these unidimensional estimates are highly dependent upon the characteristics of the simulated data sets. Hence, violating the assumption of unidimensionality could result in the very inaccurate estimation of item and ability parameters. This would prohibit the use of IRT models for a host of applications. Reckase, Carlson, Ackerman and Spray (1986) suggest that violating the assumption of unidimensionality could have serious consequences in the equating of several test forms as well as in adaptive testing. In fact, some authors have shown that the quality of equatings is worse when the forms show evidence of multidimensionality (Doody-Bogan & Yen, 1983; Dorans & Kingston, 1985). One author has also shown that the violation of unidimensionality could also affect the accuracy of differential item functioning indices (Wang, 1988). Although multidimensional IRT models have been proposed (Doody-Bogan & Yen, 1983; Mckinley & Reckase, 1983; Samejima, 1974; Sympson, 1978), the lack of computer programs available to estimate parameters derived from these models make them unappealing to the majority of researchers. Multidimensional IRT models that do provide software (e.g. Carlson, 1987 for McKinley and Reckase's model) still require the user to specify the dimensionality of the latent ability space. Hence, the researcher must make use of a given statistical technique to determine the

number of underlying abilities.

These results and considerations have therefore led researchers to develop a multitude of statistical techniques to assess test dimensionality or, more commonly, departure from the assumption of unidimensionality. These methods will be presented in the next section of this second chapter of the dissertation.

2.6. Review of literature: Methods commonly used to assess the dimensionality of a test

Commonly used and new methods of assessing the dimensionality of a test are presented below.

(a) Indices based on linear factor analysis and principal component analysis

In this first group of studies, researchers have examined the extent to which indices derived from either a principal component analysis or a linear factor analysis based on phi and tetrachoric correlation matrices could be helpful in assessing the dimensionality of dichotomous item responses generated from a logistic model (Berger & Knol, 1990; De Ayala & Hertzog, 1989; Hambleton & Rovinelli, 1986; Hattie, 1984; Zwick & Velicer, 1986). An excellent review of these indices can be found in Hattie (1982; 1984; 1985). For the most part, these indices have been found to be poor predictors of the actual number of abilities underlying a set of item responses and their values tend to be highly related to simulation conditions examined (Berger & Knol, 1990; Nandakumar, 1991b). Indices such as the magnitude of the first eigenvalue, percentage of variance accounted for by the first factor or component, the number of eigenvalues greater than 1, scree plots and the ratio of the first to the second eigenvalue generally overestimate the actual number of latent traits underlying a set of item responses

(Collins, Cliff, McCormick, & Zatkin, 1986; De Ayala & Hertzog, 1989; Hambleton & Rovinelli, 1986; Nandakumar, 1991b; Zwick & Velicer, 1986). A factor analysis of a phi correlation matrix often leads to the identification of spurious factors (Green, 1983; Hattie, 1984; McDonald & Ahlawat, 1974; Mislevy, 1986). These are often ambiguously referred to in the literature as *"difficulty"* factors because it is assumed that these extraneous factors load on homogeneous subsets of items with regards to their difficulty. However, McDonald and Ahlawat (1974) offer a more theoretically sound explanation of this phenomenon. According to these two researchers, the superfluous factors are actually composed of items that have nonlinear regressions on the latent trait. Specifically, the relationship between observed item responses and the latent trait at the upper end of the latter scale is nonlinear and hence would account for the poor fit of the linear model. Fitting a nonlinear function would therefore seem more appropriate.

The results from performing a linear factor analysis of a tetrachoric correlation matrix have shown some promise. Knol and Berger (1991) in their simulation study found that the true parameters are recovered quite well. However, a factor analysis of a tetrachoric matrix can lead to some numerical problems, not least of which are non-Gramian correlation matrices and Heywood cases (Collins et al., 1986; De Ayala & Hertzog, 1989; Hattie, 1984; Mislevy, 1986; Nandakumar, 1991a). In such cases one would need to smooth the tetrachoric matrix to be non-negative (e.g.,

MINRES, Zegers & Ten Berge, 1983). In addition, the computation of tetrachoric correlations requires that a normally distributed variable underlie each of the dichotomous items. Violation of this assumption can lead to tetrachoric correlation coefficient values that do not reflect the true degree of association between the two dichotomous variables (Gourlay, 1951). Also, spurious factors can be obtained with tetrachoric correlations if guessing is present (Carroll, 1945; Hulin, Drasgow & Parsons, 1983). Although some methods have been proposed to correct tetrachorics for guessing (Carroll, 1945), they are generally ineffective (Zwick, 1987). Lord (1980) strongly advises against computing tetrachoric correlation coefficients in the presence of a guessing factor. Finally, Hulin, Drasgow & Parsons (1983) have shown that large samples are required in order to obtain fairly accurate estimates of tetrachoric correlations. Clearly, further research into the effects of employing these correction procedures as well as the testing of appropriate indices needs to be done before one can make strong conclusions regarding the linear factor analysis of a tetrachoric matrix.

Thus, results obtained in past studies seem to prohibit the use of indices to assess test dimensionality that are based on either a linear factor analysis or a principal component analysis of a phi or tetrachoric correlation matrix. However, these results are not surprising given the degree of misfit that would be expected when attempting to fit a linear model to item responses generated from a nonlinear (logistic) model. Linear

factor analytic models assume that the relationship between item performance and the underlying ability is linear, which is clearly not the case for high ability examinees. Hence, the linear model poorly accounts for the overall item-ability relationship which does contain an important nonlinear component.

(b) Indices based on nonmetric multidimensional scaling

Some researchers have also investigated the extent to which nonmetric multidimensional scaling could be useful in assessing the dimensionality of a given test.

Multidimensional scaling procedures (MDS) are based on a distance model. Typically, measures of similarity and dissimilarity are used as inputs in the analysis. The goal of MDS is to locate objects in a *k*-dimensional space such that distances between the points (initial measures of dissimilarity, for example) are reproduced as closely as possible. Hence, the more accurate the fit is between the number of dimensions in the reproduced space and the number of dimensions in the space which contains the items, the closer the approximated distances will match the original distances. MDS procedures that have been used to assess test dimensionality are nonmetric in nature. In nonmetric MDS (NMDS), the scaling is done such that only the rank orders of the interpoint distances are maintained. In other words, NMDS will generate a solution that will attempt to preserve the monotonic relationship between (original) proximity

measures and the distances among the items in the final configuration. The fit of the (reproduced) spatial configuration is then usually assessed with the STRESS value (Kruskal & Wish, 1978). Typically, a STRESS value of .15 or less is indicative of a "satisfactory" fit between original and reproduced distances.

Results obtained following a NMDS analysis of a set of test items generally indicate that no single approach is appropriate and effective for determining the number of dimensions underlying item responses (Jones, Sabers, & Trosset, 1987; Jones, 1988; Koch, 1983). The spatial configuration retained tends to vary according to the similarity coefficient used (Jones, Sabers, & Trosset, 1987). Also, the spatial configurations seem to be adversely affected by guessing and vary according to difficulty parameter values (Reckase, 1981; Koch, 1983). Last, and more serious, NMDS analyses often yield solutions which are uninterpretable (De Ayala & Hertzog, 1989; Reckase, 1981).

McDonald (1981) has also questioned the use of NMDS from a theoretical standpoint. As was previously stated, the author states that the number of dimensions underlying a set of item responses corresponds to the *k*-dimensional space that satisfies the principle of (weak) local independence. However, multidimensional scaling analysis does not provide a principle equivalent to (weak) local independence. Indeed, the fit of a spatial configuration to an (original) set of distances between test items does not in any way entail that the conditional covariance between the latter group is equal to zero. It is

simply a descriptive technique which will provide one representation of the distances between items in a multidimensional space. Therefore, it would seem theoretically inappropriate to use NMDS to assess test dimensionality.

(c) Tucker's procedure for assessing dimensionality

Tucker (Roznowski, Tucker, and Humphreys, 1991) proposed an index to assess test dimensionality that is based on the principle of local independence. Specifically, this approach of assessing dimensionality is based on a definition of local independence similar to that proposed by McDonald (1981), that is, "weak" local independence. According to this principle, (conditional) covariances among the item scores should be zero. Tucker states that in fallible data where true level of ability is unknown, the assumption of local independence can be approximated for examinees possessing the same total score. Tucker's procedure is comprised of the following six steps:

i. Separate variance-covariance matrices are obtained for samples of examinees having the same total test score.

ii. A weighted aggregate variance-covariance matrix C is computed based on these separate matrices; the weights attributed to each separate matrix being their (respective) sample sizes.

iii. Signs of the aggregate covariances are changed according to the sign-changing procedure of centroid factor analysis in order to maximize the algebraic sum of the elements in matrix C.

Because the elements in matrix C (i.e. the covariances) are conditional upon total test score, there are approximately equal numbers of positive and negative signs in the aggregate matrices.

iv. The ratio of the algebraic sum of covariances following the sign change procedure to the absolute sum is computed. Ratios that approach unity are indicative of the presence of more than one factor among the original item covariances.

v. The ratio outlined in step iv is formed in the covariance matrix of raw scores in which total test score is not held constant. Regardless of the dimensionality, this ratio will approach unity for reliable cognitive items.

vi. The final step involves subtracting the ratio obtained in step iv from the ratio in step v. According to the authors, small differences would indicate that more than one ability is required to answer a given set of items.

Results obtained with regards to this index show that it is accurate in determining the correct number of dimensions underlying a variance-covariance matrix only when sample size is large (500 or more) and the range of item difficulties is narrow (Tucker, Humphreys, & Roznowski, 1986).

(d) Humphreys' procedure for assessing dimensionality

Humphreys (Roznowski *et al.* 1991) has proposed two procedures for assessing dimensionality that are based mainly on the analysis of the pattern of factor loadings associated with

the first two factors of a principal axis factor analysis. These two indices are based on the premise that fitting a linear factor analytic model to a set of item intercorrelations comprising a perfect Guttman scale forms a distinctive pattern. More precisely, the first component will be comprised of items that all have positive loadings with the largest values being associated with moderately difficult items and the lowest values with the easiest as well as most difficult items. Second component loadings, on the other hand, will form a curve that closely resembles an ogive with easy and difficult items having high loadings of opposite sign and with moderately difficult items having loadings close to zero. The authors argue that since a Guttman scale requires that the response patterns be unidimensional, obtaining the above mentioned factor loadings with a particular set of item correlations would be indicative of a unidimensional data set. The first index is based on the signs of the second (principal) factor of the R-matrix of product-moment correlations and involves the following four steps:

i. The correlation matrix is factored after replacing the unities in the diagonal with squared multiple correlations.

ii. Items are ranked according to difficulty and the preponderance of signs of the loadings on the second principal factor in the easy and difficult halves of the items is determined.

iii. Each aberrant item, for example, one with a positive loading in the difficult half where most of the items have negative

loadings, is given a numerical value based on the number of ranks by which it is removed from the centre of the distribution of difficulties.

iv. These numerical values are summed over all aberrant items. Small sums would therefore be associated with unidimensional data sets.

The second index proposed by Humphreys (Rosnowski *et al.*, 1991) is based on both first and second factor loadings. The first two steps are identical to those previously presented with the first index. The last two steps are as follows:

iii. Each aberrant item is given a numerical value representing the product of its first and second factor loadings.

iv. These numerical values are summed without regard to sign over all aberrant items. Again, small sums are denotative of a unidimensional structure.

Results obtained show that the first index outperforms the second in conditions examined (Tucker *et al.*, 1986). However, the first index tends to only be effective with items having a wide range of difficulties. Thus, preliminary results indicate that neither index appears to be effective in a large number of conditions.

(e) Modified parallel analysis

Hulin, Drasgow and Parsons (1983) as well as Drasgow and Lissak (1983) have proposed a procedure for assessing test

dimensionality that is an extension of parallel analysis (Horn, 1965). Specifically, this approach is referred to as modified parallel analysis and combines factor analysis as well as item response theory. The procedure involves four steps:

i. Compute the eigenvalues of the matrix of item tetrachoric correlations (obtained with real data).

ii. Using LOGIST, estimate item parameters.

iii. Based on the item parameter values obtained in step 2, generate a data set that is truly unidimensional. This simulated data set should have the same number of items and examinees as the real data.

iv. Compute the eigenvalues of the matrix of tetrachorics obtained with the simulated data set.

The assumption of unidimensionality is then assessed by comparing the second eigenvalues of the real and simulated data. A large difference would indicate that more than one dimension is required to correctly answer a set of test items. This is usually accomplished through the visual inspection of scree plots obtained with the real and simulated data sets (Drasgow & Lissak, 1983). Results obtained using this procedure suggest that it is useful in determining the number of dimensions underlying a set of item responses (Drasgow & Lissak, 1983; Hulin, Drasgow, & Parsons, 1983). However, the procedure does have definite shortcomings, not the least of which are those associated with a linear factor analysis of a tetrachoric matrix. In addition, the authors do not provide a firm quantitative criterion to help the

researcher determine if the difference between the second eigenvalues is indicative of either a unidimensional or multidimensional data set.

(f) Order analysis

An order analysis algorithm proposed by Wise (1981) that enables the researcher to extract unidimensional item chains from multidimensional tests has been reported as an alternative procedure for the assessment of dimensionality (Eddins, 1984). Briefly stated, order analysis is used to investigate logical relationships between binary test items. For example, it can be applied to assess the relationship between dimensionality of a data set and its underlying cognitive processes. The technique assumes that elements measuring a single dimension show characteristics of a strong simple order; that is, the relations between elements are transitive, asymmetric and connected. Hence, if only one ability underlies performance on a set of test items, the item relations will be consistent across examinees. Based on this assumption, Wise (1981) has assessed the effectiveness of three order analytic procedures. Only one of these (c_3, originally proposed by Reynolds, 1976) could reproduce the correct factor structure for all simulated data sets whereas none could correctly identify the factor structure of a mathematics achievement test.

However, there are serious drawbacks associated with the use

of these types of algorithms which limit their usefulness. Indeed, the accuracy with which these procedures can assess dimensionality was determined by comparing results to those obtained with a linear factor analysis (Birenbaum & Tatsuoka, 1982). More distressingly, the (linear) factor analytic indices used to determine the *"correct"* number of dimensions have been shown to be ineffective (number of eigenvalues exceeding unity and % of variance accounted for by the first factor). Hence, it is difficult to draw any definite conclusions on the merits of order analysis based on the limited number of studies available.

(g) Bejar's procedure for the assessment of dimensionality

Bejar (1980) has proposed two procedures that are based on item parameter estimates to assess the extent to which a data set deviates from the assumption of unidimensionality. The first procedure involves comparing content-based versus total-test-based item parameter estimates. Specifically, Bejar argues that if a set of test items is unidimensional, then grouping items into subcomponents for the purpose of calibration will be irrelevant. The procedure entails the following three steps:

i. Identify subsets of items that appear to be measuring distinct dimensions.

ii. Calibrate item difficulties (separately) for items of a given subtest as well as for the total test.

iii. Plot the subtest-based item difficulties against the total-

test-based item difficulties.

If the test is unidimensional, item difficulty values will cluster about a straight line with a slope of 1 and an intercept of 0.

A second ancillary procedure is based on the computation of mean squared distances of each content area to the theoretical axis (i.e. a straight line with slope equal to one and intercept equal to zero). In other words, the greater distance between pairs of item difficulties and the theoretical axis, the greater the deviation from the assumption of unidimensionality (Bejar, 1980). Initial findings had suggested that the procedures were helpful in assessing departure from the assumption of unidimensionality (Bejar, 1980; Kingsbury, 1985). However, more recent results obtained with both procedures generally indicate that they are ineffective in correctly identifying departure from that assumption (Hambleton & Rovinelli, 1986; Liou, 1988). Indeed, the two methods are often unable to identify departure from unidimensionality in data sets that have clearly been generated to be two-dimensional (Hambleton & Rovinelli, 1986). Others have suggested that the procedures are ineffective with short test lengths (less than 60 items) (Liou, 1988). Bejar (1988) has recently shown in a Monte Carlo study that the procedures work well when each dimension is defined by an equal number of items. However, its accuracy can be seriously questioned when the dimensions are defined by an unequal number of items.

Finally, the descriptive nature of the procedures limits their usefulness in many testing situations.

(h) The Holland-Rosenbaum procedure

A theorem based on conditional association between pairs of items, that can be used to assess test dimensionality, has been proposed by Holland (1981), Rosenbaum (1984) as well as Holland and Rosenbaum (1986).

Specifically, Rosenbaum (1984) states that if ICCs are monotone nondecreasing functions of a single ability, the local independence of item responses implies nonnegative conditional covariance between all pairs of item responses, that is,

$$COV(X_j, X_k, | \sum_{i \neq j,k} X_i) \geq 0 \qquad (21)$$

where ΣX_i is the number-right score on the remaining n-2 items, taken as an estimate of ability. Hence, the researcher can develop a statistical test to assess the following related assumptions: local independence, unidimensionality and monotonicity of the latent trait. Conditional association for each pair of items is usually tested with the Mantel-Haenszel statistic (Mantel & Haenszel, 1959). The computation of this statistic involves representing (dichotomous) item responses in 2 x 2 contingency table such as,

ITEM K				
ITEM J		1	0	
	1	a	b	a + b
	0	c	d	c + d
		a + c	b + d	N

The Mantel-Haenszel statistic is then given by

$$Z=\frac{\Sigma a_i-E(\Sigma a_i)+.5}{\sqrt{VAR(\Sigma a_i)}} \tag{22}$$

where $E(\Sigma a_i)$ is given by,

$$E(\Sigma a_i)=\sum_{i=1}^{i}\frac{(a_i+b_i)\,(a_i+c_i)}{N_i} \tag{23}$$

and $VAR(\Sigma a_i)$ is equal to,

$$VAR(\Sigma a_i)=\sum_{i=1}^{i}\frac{(a_i+b_i)\,(c_i+d_i)\,(a_i+c_i)\,(b_i+d_i)}{N_i^2\,(N_i-1)} \tag{24}$$

The MH(z) is normally distributed and the approximate significance level can be obtained by referring to the lower tail of the standard normal distribution. For example, rejecting the null hypothesis for a given pair of items would indicate that they have negative partial association. A large number of conditionally associated pairs of items would suggest that the

test is multidimensional.

Results obtained with regards to the usefulness of the procedure are somewhat conflicting. Zwick (1987) applied the Holland-Rosenbaum procedure to assess the dimensionality of the NAEP data and showed that results were consistent with those obtained with full-information factor analysis (i.e. TESTFACT, Wilson, Wood, & Gibbons, 1987). However, two major limitations associated with this study cast a doubt over these interpretations. Firstly, due to CPU time restrictions, only a sample of items was subjected to TESTFACT and the Holland-Rosenbaum procedure. Rosenbaum (1984) states that *all* item pairs should be subjected to the test unless a priori knowledge of the structure would justify looking at only a subset. Finally, the usefulness of the Holland-Rosenbaum procedure was determined by comparing it to FIFA using two statistics that have been shown to often be ineffective in TESTFACT, namely the likelihood-ratio χ^2 and the χ^2 difference test (c.f. Berger & Knol, 1990). On the other hand, Nandakumar (1991b), using Bonferoni bounds, and Ben-Simon and Cohen (1990) have shown that the procedure is not very powerful in correctly rejecting the assumptions of local independence, unidimensionality and monotonicity of the latent trait. Indeed, Ben-Simon and Cohen's results indicate that the procedure correctly identified multidimensionality in only 56% of item pairs simulated to be two-dimensional. Nevertheless, the latter authors did obtain encouraging results using a modified version of the procedure (APSN index) that incorporates parallel

analysis. However, their results are based on a very limited number of analyses (only 16 data sets were considered) and more research on the effectiveness of the procedure under varying conditions should be done before judging the merits of their approach more conclusively.

(i) Stout's procedure for assessing dimensionality

Stout (1987; 1990) has elaborated a nonparametric statistical procedure to assess the dimensionality of the latent space. This procedure is based on a new conceptualization of dimensionality as well as local independence, namely *essential* dimensionality and *essential* independence. Specifically, *essential* dimensionality corresponds to the number of latent traits required to satisfy the assumption of *essential* independence, that is, a conditional mean absolute residual covariance value which tends towards zero after ability has been partialled out,

$$\frac{1}{N(N-1)} \sum_{1 \le i \ne j \le N} |Cov(U_i, U_j | \theta)| = 0. \qquad (25)$$

A test $(U_1, U_2, \ldots, U_N)$ is said to be *essentially* unidimensional if for <u>all</u> subsets $\{(U_1, U_2, \ldots, U_M)\}$ of length M (<N) and all values of Y_p,

$$S_{M,N} = \frac{1}{M(M-1)} \sum_{1 \le i \ne j \le N} |Cov(U_i, U_j | Y_p)| = 0 \qquad (26)$$

where Y_p is the proportion correct score on the longer subtest and (U_1, U_2, ..., U_M) are shorter subtests with length n = N - M.

Stout's procedure, therefore, is sensitive to dominant dimensions and is only minimally influenced by trivial (minor) dimensions. Consequently, it is suited to the analysis of cognitive test data which are often characterized by a dominant ability and several minor dimensions. The assumption of essential unidimensionality is tested with the T statistic (Stout, 1987; 1990). The steps involved in the computation of the T statistic as well a modified version of the procedure (Nandakumar, 1987) are presented in Appendix B.

Results indicate that the T statistic appears to be accurate in assessing the assumption of *essential* unidimensionality (Stout, 1987), especially when Nandakumar's bias correction modification is utilized (Nandakumar, 1987; 1988; 1989, 1991b). However, the power of the statistic seems to decrease with certain test structures (specifically, tests that contain many items with low discrimination values) and as test length and sample size decrease (Nandakumar, 1987; Stout, 1987). Therefore, Nandakumar (1987) does not generally recommend using this procedure with a small number of items (less than 25) and small sample sizes (less than 750). Recent work by Junker (1991) and Junker and Stout (1991) was undertaken in order to develop and investigate an index to correct the standard error of the estimate of unidimensional ability when various degrees of multidimensionality are present. This work appears to be

important because it enables the researcher to assess the degree of accuracy of the unidimensional estimate of ability in the presence of some multidimensionality.

(j) Nonlinear factor analysis

Another approach which is gaining popularity in the assessment of test dimensionality is the one that treats IRT as a special case of nonlinear factor analysis. As was previously stated, several authors have shown that these two models are mathematically equivalent (Bartholomew, 1983; McDonald, 1989, 1991; Goldstein & Wood, 1989). Using this IRT-nonlinear factor analytic (NLFA) relationship, some researchers have suggested that the most suitable method of assessing dimensionality should be based on the analysis of the residual covariance matrix after fitting a *k*-factor NLFA model (Goldstein, 1980; Goldstein & Wood, 1989; McDonald, 1981, 1989). With regards to the development of indices and statistics to assess test dimensionality within this framework, the majority of the research has been based upon the *full-information factor model* (Bock, Gibbons, & Muraki, 1988) as well as *limited-information models* such as McDonald's joint-proportions (pairwise) model (McDonald, 1967). This research is summarized below.

Full-information factor analysis

Statistics based on full-information factor analysis (FIFA) have also been proposed as a means of assessing test

dimensionality (Bock & Aitkin, 1981; Bock, Gibbons, & Muraki, 1988). Most factor analytic models presented so far are based solely on low-order joint occurrence frequencies of item scores (Christoffersson, 1975; McDonald, 1967; Muthen, 1978). FIFA, on the other hand, is based on the distinct item response vectors of all examinees and tests the "strong" principle of local independence, that is, the statistical independence of higher-order joint moments (Bock & Aitkin, 1981). Since it uses all available information in a matrix of dichotomous item scores it is often referred to as *full-information factor analysis*. FIFA is based on an iterative marginal maximum likelihood estimation procedure derived from the EM algorithm (Dempster, Laird, & Rubin, 1977). Some authors, however, have recently suggested that expected a posteriori scores (EAP) provided a better estimation of factor scores than maximum likelihood (Muraki & Engelhard, 1985). The procedure was implemented in the computer program TESTFACT (Wilson, Wood, & Gibbons, 1987). Readers who wish to obtain more information on the procedure should refer to Bock, Gibbons, and Muraki (1988) for a concise presentation of the approach. FIFA possesses the following distinct advantages:

- It uses all available information in a matrix of item responses;
- It does not analyze pairwise item correlations and hence avoids problems associated with the factor analysis of tetrachoric correlation coefficients.

These advantages have led several researchers to use this

approach for the analysis of test items (Berger & Knol, 1990; Dorans & Lawrence, 1988; Kingston, 1986; Kingston & McKinley, 1988; Morgan, 1989; Muraki & Engelhard, 1985). From a practical standpoint, however, a distinct disadvantage of the procedure is the immense amount of CPU time required to analyze a relatively small set of items. Indeed, in the context of Monte Carlo studies, several researchers have stated that the model as well as the accompanying computer program (TESTFACT) could only be used for small data sets (15 items or less) and a very small number of replications (typically, less that 25) (Berger & Knol, 1990; Knol & Berger, 1991). Also, there are 2^p distinct response vectors where p is equal to the number of items. Hence, in order to use the "full information" contained in the data, there should be no empty cells which is usually not feasible unless some collapsing is done.

Several statistics and indices have been proposed to assess the fit of *k*-dimensional models obtained with TESTFACT. The two most popular statistics appear to be the likelihood-ratio χ^2 goodness-of-fit test as well as the χ^2 likelihood-ratio difference test. Most studies have shown that these two statistics are helpful in determining the number of dimensions underlying a set of item responses (Bock, Gibbons, & Muraki, 1988; Kingston, 1986; Muraki & Engelhard, 1985). However, these studies analyzed "real" data sets with no a priori knowledge of the actual number of underlying dimensions. Studies that have examined the effectiveness of these two statistics with true

(i.e. simulated) unidimensional and multidimensional data have shown that they are not very powerful in detecting the correct number of abilities required to correctly answer a set of test items (Berger & Knol, 1990).

One index, based on maximum likelihood estimation, which seems more promising for the assessment of dimensionality is Akaike's information criterion (AIC) (Akaike, 1987). The AIC can be defined, in its simplest form, as,

$$AIC(H) = \chi^2_{d.f.} - 2(d.f.) \tag{27}$$

The measure is interpreted as a "badness-of-fit" index, that is, the higher its value, the less adequate the fit of a given model is. In the context of dimensionality, a large AIC value would suggest that more than one ability is required to correctly answer a set of test items. Though the criterion seems to be helpful in correctly identifying the dimensionality of an item response matrix (Berger & Knol, 1990), McDonald (1989b) has shown it to be dependent upon sample size.

In conclusion, although FIFA appears to be a theoretically promising alternative for the assessment of test items, more research should be directed at assessing current fit indices and statistics before making any definite conclusions about their effectiveness. In addition, researchers should attempt to elaborate indices that are consistent over variations of sample size and other pertinent factors such as the one provided by McDonald (1989b).

Limited-information factor analysis

Other factor-analytic models use the information present in the pairwise relationships between the items (Christoffersson, 1975; McDonald, 1967; Muthen, 1978). Given that only second-order relationships (pairwise) among items are used, McDonald (1981; 1991) labels these models as being based on "bivariate" information and suggests that underlying factors are defined by the "weak" principle of local independence. Although the weak principle of local independence is employed, McDonald (1991) and Muthen (1978) suggest that very little information should be lost by not using the higher-order relationships among items. A systematic investigation of this issue would seem necessary to address this problem.

Two general approaches to the *limited-information* non-linear factor analysis have been proposed. Christoffersson (1975) and Muthen (1978) employ generalized least-squares in the estimation of parameters. From this, one is able to derive a theoretically based χ^2 on which to test the adequacy of the proposed model. Unfortunately, due to the nature of the generalized-least squares estimation, the procedure is limited to the analysis of approximately no more than 25 items.

On the other hand, as was previously shown, McDonald's (1967) approach to non-linear factor analysis uses unweighted least-squares estimation of the model parameters. This procedure has the practical advantage of allowing for the analysis of tests with a large number of items and/or dimensions.

Numerous authors have suggested that McDonald's nonlinear factor analytic model and accompanying principles (namely, "weak" local independence) provide a sound theoretical framework on which indices as well as statistics could be developed to assess test dimensionality (Goldstein, 1980; Goldstein & Wood, 1989; Hattie, 1984, 1985; Hambleton & Rogers, 1986; Traub & Lam, 1985). Several authors have, in fact, proposed a host of indices based on McDonald's model to assess the number of abilities underlying a set of item responses (Berger & Knol, 1990; Hambleton & Rovinelli, 1986; Hattie, 1984; Knol & Berger, 1991; Nandakumar, 1991b).

Indeed, results obtained by Hambleton and Rovinelli (1986) and Nandakumar (1991b) show that the value of the mean absolute residual as well as the standardized mean absolute residual tends to be related to the number of dimensions underlying a set of item responses. Hattie (1982; 1985) also demonstrated that the absolute sum of squares of residual correlations appears to be useful in determining the number of dimensions underlying examinee responses. Berger and Knol (1990) also suggested depicting the mean absolute residual values obtained after fitting models of varying dimensionality using scree plots. From a practical perspective, however, the unrealistic test length of the data sets generated (15 items) as well as the small number of replications (10) indicate that the authors' conclusions should be interpreted cautiously and that the index should be assessed in more varied situations before any definite judgment is made

about its effectiveness. Finally, the Incremental Fit Index (IFI), based on the sum of squares of the residual covariances (SSRes) was proposed and investigated by De Champlain & Gessaroli (1991). In the context of assessing the dimensionality of a set of test items, we can define the IFI as:

$$IFI_m = \frac{SS_{Res}(m\text{-}factor) - SS_{Res}((m+1)\text{-}factor)}{SS_{Res}(m\text{-}factor)} \qquad (28)$$

The IFI calculates the proportion of the sum of squares of the residual covariances from the m-factor solution that is accounted for by the (m+1)-factor. If the (m+1)-factor is important in explaining the structure of the items, then the IFI should be quite large.

Knol and Berger (1991) found that the estimates of the factor analytic parameters obtained from NOHARMII (Fraser & McDonald, 1988), a computer program based on McDonald's non-linear factor analysis, compared very favourably with those obtained from the full-information methods used in TESTFACT. Though results using this procedure in the assessment of dimensionality are promising, they are limited by their purely descriptive nature. Indeed, they offer no objective criteria to the researcher on which to assess the number of dimensions underlying a set of item correlations. Knol and Berger (1991) realize this problem when they state that

> "A possible drawback of all common FA methods and NOHARM is that no statistical goodness of fit tests are available, hence the assessment of the dimensionality of the model can be problematic" (p. 475).

(k) Summary of procedures for the assessment of dimensionality

The previous sections of the theoretical framework outlined the main procedures that have been proposed in the recent past to assess the number of dimensions (abilities) underlying a set of educational test items. At present, the two most promising approaches to this problem appear to be Stout's T statistic for the assessment of *essential* dimensionality as well as indices based on the analysis of a residual correlation matrix obtained after fitting a nonlinear factor analytic model. However, the power of Stout's statistic is questionable with short tests (less than 25 items) and moderately small sample sizes (less than 750 examinees).

Also, nonlinear factor analytic indices are descriptive and offer no solid criterion on which to determine dimensionality. Therefore, with regards to the latter procedure, future research in this area should focus on the development of an inferential statistic that would allow practitioners to determine, with greater confidence, the number of abilities required to answer a set of item responses obtained from various test structures. Two approximate chi-square statistics for assessing dimensionality are outlined in the next section of the second chapter.

2.7. Description of two approximate χ^2 statistics for the assessment of dimensionality

In this section, two approximate χ^2 statistics, based on McDonald's weak principle of local independence, that might be applicable to the problem of dimensionality assessment are presented. The research problem and questions are also outlined.

a. χ^2_1

The first procedure for assessing the dimensionality of a set of items that is examined is based on the computation of an approximate χ^2 test statistic initially proposed by Bartlett (1950) and outlined by Steiger (1980a, 1980b). It is hereinafter be referred to as χ^2_1. As was stated in the introduction, the statistic tests the null hypothesis that the off-diagonal elements in a matrix of residual correlations after a nonlinear factor analysis are equal to zero. The statistic involves the following five computational steps after estimating parameters for an *m*-factor model using the nonlinear factor analytic approach outlined by McDonald (1967) and implemented by Fraser and McDonald (1988) in the computer program NOHARMII.

1. For a given pair of items, determine the proportion of examinees who correctly answered item i, item j, as well as both items. These quantities will be referred to as p_i, p_j, and p_{ij}.

2. For the same pair of items, determine the expected as well as residual proportions of examinees who correctly answered items i and j. The residual joint-proportions estimates are provided by the computer program NOHARMII (Fraser & McDonald, 1988) and will be referred to as $p_{ij}^{(r)}$.

3. Calculate the residual correlations for each pair of dichotomous items with the following formula (product-moment correlation coefficient or phi coefficient).

$$r_{ij}^{(r)} = \frac{p_{ij}^{(r)}}{\sqrt{p_i(1-p_i)\,p_j(1-p_j)}} \tag{29}$$

4. Transform each of the residual correlations to a Fisher z using the following formula,

$$z_{ij}^{(r)} = .5\log_e(1+r_{ij}^{(r)}) - .5\log_e(1-r_{ij}^{(r)}) \tag{30}$$

5. Calculate an approximate χ^2_1 statistic defined as

$$\chi_1^2 = (N-3)\sum_{i=1}^{p}\sum_{j=1}^{i-1} z^{2(r)}_{ij} \tag{31}$$

where $z_{ij}^{2(r)}$ is the square of the Fisher z corresponding to the residual correlation between items *i* and *j*, ($i,j = 1 \ldots p$) and *N* is the number of subjects in the sample. This statistic is possibly distributed as an approximate 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.

b. χ^2_2

The second procedure, hereinafter referred to as χ^2_2, also involves the calculation of residual correlations. Specifically, the procedure entails the following steps:

1. Step 1 involves the computation of $p_{ij}^{(r)}$ as was described with χ^2_1.

2. Compute the asymptotic variance estimate for joint proportions with the following equation (*Source*: Bishop, Fienberg, & Holland, 1975, p. 381).

$$\sigma^2_{\rho_{ij}} = \frac{1}{N}\{1-\rho^2+(\rho+\frac{1}{2}\rho^3)\frac{(P_{1\cdot}-P_{2\cdot})(P_{\cdot 1}-P_{\cdot 2})}{\sqrt{P_{1\cdot}P_{2\cdot}P_{\cdot 1}P_{\cdot 2}}}-\frac{3}{4}\rho^2[\frac{(P_{1\cdot}-P_{2\cdot})^2}{P_{1\cdot}P_{2\cdot}}+\frac{(P_{\cdot 1}-P_{\cdot 2})^2}{P_{\cdot 1}P_{\cdot 2}}]\} \qquad (32)$$

where

ρ= the correlation between a given pair of items (c.f. equation 30);

p_{1+}= the proportion of examinees who incorrectly answered item 2;

p_{2+}= the proportion of examinees who correctly answered item 2;

p_{+1}= the proportion of examinees who incorrectly answered item 1;

p_{+2}= the proportion of examinees who correctly answered item 1.

3. Calculate the standardized residual for each pair of items,

$$z_{ij}^{(r)} = \frac{p_{ij}^{(r)}}{\sigma_{\rho_{ij}}} \tag{33}$$

4. Compute χ^2_2 with the following formula,

$$\chi_2^2 = \sum_{i=1}^{p} \sum_{j=1}^{i-1} z_{ij}^{2(r)} \tag{34}$$

where $z_{ij}^{2(r)}$ is the square of the z corresponding to the residual correlation between items *i* and *j*, ($i,j = 1, \ldots, p$) computed in step 3. This statistic is also distributed approximately 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.

There are several possible advantages to these statistics as methods of assessing test dimensionality:

1. The assessment of dimensionality is made using a model on which IRT is based (nonlinear factor analysis).
2. The statistics involve actual hypothesis testing and are not merely descriptive indices. Hence, the criterion values on which to base decisions regarding dimensionality are perhaps not intrinsically linked to a set of simulation conditions and possibly could be used for a variety of data sets with greater confidence.

3. The statistics can conceivably be used to assess not only departure from unidimensionality but also the fit of successively more complex models, i.e., two-,three-, etc. dimensional structures to a set of item correlations.
4. The statistics have the desirable trait of being based on the discrepancy function (i.e., the discrepancy between the observed and fitted item-covariance matrices) which is consistent with the unweighted least-squares estimation procedure.
5. Because of the unweighted least-squares estimation procedure, these statistics are derived from a model that has no severe limitation on the number of items or dimensions that can be analyzed.

There are, however, some important limitations associated with these statistics. First, these approximate χ^2 statistics are based on unweighted least-squares estimation and, of course, are weak in their theoretical foundation. Browne (1977), however, has indicated that in many cases these χ^2s are formally equivalent to a χ^2 obtained from generalized least-squares estimation. In fact, Browne states that in general, they differ only slightly. Therefore, from a practical perspective, this statistic has the potential to be a useful tool for the assessment of dimensionality.

Finally, it is important to note that these two χ^2 statistics have the same limitations as other χ^2 indices.

Specifically, one would expect to often falsely reject the correct *m*-factor model with large sample sizes and, conversely, fail to reject inappropriate models with small samples (Marsh, Balla, & McDonald, 1988). It is therefore important to not only look at the values of these statistics when assessing the dimensionality of a set of test items but also ancillary methods that are commonly used in covariance structure modelling and possibly not affected by sample size such as a χ^2 difference test (Hayduk, 1987; Loehlin, 1987). Chi-square difference tests, based on χ^2_1 and χ^2_2 were therefore also computed for both unidimensional and multidimensional data sets.

2.8. Research problem and questions

One of the main reasons for the use of tests in education is to obtain an indication of the level of ability of students in a particular area. In the simplest situation, individual (usually dichotomous) item scores are summed and the resulting total score is interpreted as a measure of ability. If this total score is to have any meaning whatsoever, all items must measure a common ability. If the assumption of unidimensionality doesn't hold, the total test score cannot be legitimately interpreted. For example, assume that a mathematics test requires knowledge of mathematics as well as reading comprehension. Two students could theoretically have the same ability score in mathematics but different total scores due to unequal ability in reading comprehension. Hence, interpreting the total score as an ability score in mathematics would unfairly penalize some students.

As was previously stated, one of the main assumptions of the majority of IRT models that are currently used to analyze test items is that only one dimension (ability) underlies item scores.

Violation of this assumption can lead to serious problems, that is, lack of robustness in the estimation of multidimensional parameters by unidimensional models. Hence, the use of a unidimensional or multidimensional IRT model to estimate item and ability parameter values requires that the researcher assess the dimensionality of a given set of test items. At the present time, nonlinear factor analysis is one of the most promising avenues for assessing the dimensionality of an item correlation matrix.

Indeed, McDonald (1967) has provided a model which allows the researcher to determine the smallest k-dimensional space that satisfies the (falsifiable) principle of "weak" local independence. Though several descriptive indices based on McDonald's model have been proposed, none seems to be ideal beyond a somewhat rigid set of simulation conditions.

The purpose of this study is therefore to examine two inferential procedures for assessing the dimensionality of an item correlation matrix that are based on the computation of approximate chi-square statistics and to compare their performance to that of another promising method, Stout's T statistic. Results that will be obtained in the present study will allow us to answer the following three research questions.

(1) How accurate are the approximate χ^2statistics, χ^2 difference tests and T statistic in identifying unidimensional test structures?

(2) How accurate are the approximate χ^2 statistics, χ^2 difference tests and T statistic in correctly rejecting unidimensional test structures?

(3) Are there either first- or higher-order associations of test length, sample size, dimension strength, and dimension dominance on the performance of the approximate χ^2 and T statistics?

CHAPTER 3

METHODS

A simulation study was carried out to assess the effectiveness of two approximate chi-square statistics and chi-square difference tests for the assessment of dimensionality. Given that this was a preliminary investigation of the two procedures, it was felt important to select a type of research design that would allow us to control all threats to the internal validity of the study and hence allow us to attribute results obtained to the variables that were manipulated. In this chapter, a description of the model used for the simulations and the conditions utilized in the unidimensional as well as multidimensional studies will be presented. Finally, the computer programs used in both studies will be outlined.

3.1. Model used for the simulations

There were two parts to the study. In the first part, the rejection rates of χ^2_1, χ^2_2, χ^2_1DF (chi-square difference test based on χ^2_1), χ^2_2DF (chi-square difference test based on χ^2_2) and the T statistic with various unidimensional data sets were examined. The purpose of the second part was to determine the level of accuracy of the same five statistics in rejecting unidimensionality. In the first part of this study, binary response strings were randomly generated using the 2-parameter logistic model proposed by Birnbaum (1968) and outlined in section 2.1. In the second part of the study, two-dimensional

dichotomous response strings were generated using the general two-parameter compensatory multidimensional IRT model proposed by McKinley and Reckase (1983) and also described in section 2.1.

The two-parameter model was selected over the three-parameter model because it was felt that the somewhat modest sample sizes simulated (500 and 1000 examinees) would not allow for a precise estimate of the lower asymptote parameter ("c" parameter) included in the latter model. Therefore, this parameter was set at zero in both parts of the study; that is, "pseudo-guessing" was not taken into account in the generation of the data.

3.2. Unidimensional data sets

The first part of the study entailed comparing empirical and theoretical Type I error rates of the two χ^2 statistics, χ^2 difference tests and Stout's T statistic with varying unidimensional test structures. The theoretical Type I error rate (α) selected was .05. In order to carry out the first part of this study, unidimensional data sets were generated with a modified version of M2PLGEN (Ackerman, 1985; modification by Gessaroli, 1990), a computer program designed to simulate binary response strings based on a two-parameter logistic model.

(a) Test length

Test length was set at either 15, 30 or 45 items. Hence, the number of items retained reflect three approximate test lengths that are typically found with actual achievement tests. It was very difficult to simulate tests containing more than 45 items

due to computer time restrictions.

(b) Sample size and ability distribution

Examinee abilities were randomly generated from a unit normal distribution. Data sets containing 500 and 1000 examinees were generated for this part of the study. Past research has indicated that item and ability estimates obtained with IRT models are reasonably stable with 500 or more examinees (Swaminathan & Gifford, 1983). In addition, very little research has been done to assess the usefulness of Stout's T-statistic with sample sizes containing less than 750 examinees. The sample size of 500 examinees retained in this study will therefore give us some preliminary information regarding this problem.

(c) Dimension strength

The mean and variance of the item parameters used to generate response strings were similar to those reported with the ACT English Usage Test Battery (Drasgow, 1987), the SAT Verbal Test Battery (Drasgow & Parsons, 1983) and the Armed Services Vocational Aptitude Battery for Arithmetic Reasoning (Bock, 1984) (cited from Nandakumar, 1987). The means and variances of the item parameters for these three test structures (i.e. dimension strengths) are presented in Table 1.

In is important to stress that the selection of the item parameter means and variances was done somewhat arbitrarily; that is, values were chosen so as to reflect three distinct test structures varying in terms of their dimension strength. Indeed, data sets generated according to the third dimension strength can

Table 1

Mean and variance of item parameters used to generate unidimensional data sets

	ACT ENG	SAT V	ASVAB AR
μ_a	0.72	1.07	1.46
σ^2_a	0.06	0.16	0.26
μ_b	0.00	0.00	0.00
σ^2_b	0.92	0.77	0.71

be viewed as being comprised of (mainly) *highly* discriminating items whereas the second and first structures include, respectively, *moderately* and *low* discriminating items. It would be expected that the accuracy rates of all three statistics would increase as the strength of the dimension increases.

(d) Final design: Unidimensional study

In the first part of this study, data sets in each cell of this 2 x 3 x 3 design (i.e. sample size by test length by dimension strength) were replicated 100 times for a total of 1800 unidimensional data sets.

Each of these data sets was analyzed with both a 1-factor (i.e. unidimensional) and 2-factor (i.e. two-dimensional) model specification. This enabled us to determine if one or more than one ability was required to correctly answer the items on the (simulated) tests. Stout's T statistic was also computed for each

unidimensional data set to determine if they were *essentially* unidimensional or not.

3.3. Two-dimensional data sets

The second part of the study entailed examining the power of the χ^2 and T statistics in rejecting the assumption of unidimensionality with two-dimensional simulated test structures. As was the case in the first part of the study, two-dimensional binary response strings were generated (based on a two-parameter logistic compensatory IRT model) using M2PLGEN (Ackerman, 1985; modification by Gessaroli, 1990).

(a) Test length

Test length was again set at either 15, 30 or 45 items.

(b) Sample size and ability distribution

Examinee abilities were also randomly generated from a unit normal distribution. Data sets containing 500 and 1000 examinees were once more simulated for this second part of the study.

(c) Dimension dominance

Two test structures reflecting distinct dimension dominance conditions were simulated in the second part of the study. These structures are shown in Table 2.

Data sets generated according to the first structure (i.e. dimension dominance A) correspond to a test having one fairly strong (dominant) dimension. Specifically, 80% of the items require knowledge of the first ability whereas 20% of the items require knowledge of the second ability only. On the other hand,

Table 2

Two dimension dominance structures

	Dimension Dominance A	Dimension Dominance B
% of items on θ_1	80%	≃ 50%
% of items on θ_2	20%	≃ 50%

data sets simulated according to the second structure (i.e. dimension dominance B) reflect a more distinct two-dimensional test where approximately half of the items require knowledge of the first ability only and the remaining items require knowledge of only the second ability. The number of items measuring each dimension was not equally divided for data sets generated according to the second structure (dimension dominance B) due to the test lengths chosen (15, 30 and 45 items). The exact number of items requiring knowledge of each dimension according to the second structure is as follows:

15 item data sets: 8 items on θ_1 and 7 items on θ_2

30 item data sets: 15 items on θ_1 and 15 items on θ_2

45 item data sets: 23 items on θ_1 and 22 items on θ_2

Hypothetical tests corresponding to the above mentioned two structures are presented in Appendix C for the reader's benefit.

(d) Dimension strength

The mean and variance of the item parameters used to generate multidimensional response strings were identical to those reported previously with unidimensional data sets. Specifically, M2PLGEN requires that the user supply discrimination parameter values and a difficulty scalar value for each item (see d parameter in equation 2). Since we are dealing with two-dimensional data sets, values of a_1, a_2 and d were randomly generated according to the means and variances of the unidimensional discriminations and difficulties provided in Table 1. Using these same values will result in identical mean and variance MDISC and MDIF values for both unidimensional and multidimensional data sets. As was previously outlined, with two-dimensional data, MDISC can be defined as,

$$MDISC=\sqrt{a_{j1}^2+a_{j2}^2}$$

where a_{jk} is the discrimination parameter of item j on dimension k, $(k=1,2)$.

Given the two-dimensional test structures examined, MDISC for a given item will always be reduced to,

$$MDISC=\sqrt{a_{j1}^2+0}$$

or,

$$MDISC=\sqrt{0+a_{j2}^2}$$

Hence, the MDISC value can always be simplified to

$$MDISC=\sqrt{a_j^2}$$

or a_i, the unidimensional discrimination parameter value for item i.

Similarly, MDIF was previously defined as,

$$MDIF=\frac{-d}{MDISC}$$

In the unidimensional case, the MDIF value is computed as,

$$MDIF=\frac{-b}{MDISC}$$

Hence, using the same mean and variance values to generate the (unidimensional) difficulties and d scalars will result in the same mean and variance MDIF values for the unidimensional and two-dimensional data sets.

(e) Correlation between abilities

The correlation between the two latent abilities was fixed at zero. Although this may be atypical of actual test data, we felt that it was important to control this variable (ability intercorrelations) given that this was an initial investigation of the two approximate χ^2 statistics. Also, computer time restrictions made it difficult to generate the data sets that would allow us to properly examine this problem. It was felt that it was more important, in this preliminary investigation of the approximate χ^2 statistics, to assess their performance with two

very distinct (simple) structures, that is, one requiring a dominant as well as weak second ability and the other necessitating two stronger dimensions. In other words, we did not, at this point, want to confound the problem of dimensionality assessment with the whole issue of correlated abilities. However, as will be discussed in the conclusion of this dissertation, this is obviously a crucial issue that will have to be addressed in further studies of these approximate χ^2 procedures.

(f) Final design: multidimensional study

In the final part of this study, 100 data sets in each cell of this 2 x 3 x 3 x 2 design (i.e. sample size by test length by dimension strength by dimension dominance) were generated for a total of 3600 two-dimensional data sets.

Each of these data sets was also analyzed with both a 1-factor (i.e. unidimensional) and 2-factor (i.e. two-dimensional) model specification. This once more allowed us to determine if one or more than one ability was required to correctly answer the items on the (simulated) tests. Stout's T statistic was also computed for each unidimensional data set to determine if it was *essentially* unidimensional or not.

In addition, results obtained with the approximate χ^2 and T statistics were compared using logit linear analysis. More precisely, the effects of test length, sample size, dimension strength and dimension dominance on the rejection rates of the

assumption of unidimensionality were assessed.

3.4. Computer programs

As was previously stated, unidimensional as well as multidimensional data sets were generated using a modified version of M2PLGEN (Ackerman, 1985; modification by Gessaroli), a program that simulates dichotomous response strings based on a two-parameter logistic model. Specifically, the program was modified in the following two ways:

- The two-parameter (unidimensional) logistic formula was incorporated in the program in order to enable the generation of data sets for the first part of the study.
- The program was also modified using the IMSL library in order to allow the researcher to generate unidimensional as well as multidimensional data sets according to specific item difficulty, discrimination and ability means and variances. The accuracy with which the program generated item parameters was assessed by computing the actual means and variances of the difficulties (d) and discriminations (a_1, a_2) obtained in the "45 item, 1000 examinee, dimension dominance B" condition and comparing these values to the expected values (see Table 1). These results are presented in Appendix D.

Unidimensional and multidimensional data sets were also analyzed using NOHARMII (Fraser & McDonald, 1988), a program based on McDonald's polynomial approximation to a normal ogive model. Specifically, the program approximates a normal ogive

model using a four-term polynomial series of the form,

$$Y(i) = b_{i0} + b_{i11}\theta_1 + b_{i12}\theta_1^2 + b_{i13}\theta_1^3 \quad (41)$$

in the unidimensional case and,

$$Y(i) = b_{i0} + b_{i11}\theta_1 + b_{i12}\theta_1^2 + b_{i13}\theta_1^3 + b_{i12}\theta_2 + b_{i22}\theta_2^2 + b_{i23}\theta_2^3 \quad (42)$$

in the two-dimensional case, where

b_{i0} = an intercept term,

b_{ijk} = the factor loading of factor j on item i of polynomial degree k.

An unweighted least squares function is minimized in the program using an iterative conjugate gradients minimization algorithm. In addition, given that starting values for the estimation of the parameters are essential for the minimization of the fit function, factor loadings obtained from a linear factor analysis of the item-correlations matrix (phi coefficients) were used as initial values.

Stout's T statistic (1987; 1990), implementing Nandakumar's (1987) bias correction procedure, was computed for these same unidimensional and multidimensional data sets using a program written by Junker (1988).

Unidimensionality of the data sets was tested at the .05 level of significance for the five statistics examined.

Finally, the logit-linear analyses were performed using SPSSX (Norusis, 1988).

CHAPTER 4

RESULTS

Results obtained in the present study are presented in this section of the dissertation. First, empirical Type I error rates obtained for both approximate χ^2 statistics, the χ^2 difference tests as well as Stout's T statistic are presented for unidimensional data sets. Second, the number of rejections of the assumption of unidimensionality are given for χ^2_1, χ^2_2, χ^2_1 difference test, χ^2_2 difference test as well as Stout's T statistic with regards to both dimension dominance conditions examined in this study. Finally, in order to investigate the effects of the independent variables on the rejection rates, separate logit-linear analyses were performed for the approximate χ^2 and Stout T statistics for each of the unidimensional and multidimensional conditions. Specifically, logit-linear analyses were performed with the objective of fitting the most parsimonious model to the response frequencies. The independent variables were test length, sample size, dimension strength and dimension dominance. The dependent variable was the number of acceptances and rejections of the null hypothesis. This variable was labelled "rejection decision". The logit-linear analysis was done in a forward hierarchical manner (i.e., starting with the simplest main effect and then fitting incrementally more complex models while adhering to the principle that lower-order effects are also included in the model). The likelihood-ratio χ^2 was employed as the fit statistic. A model was deemed to be

acceptable if the corresponding p-value was equal to or greater than 0.15. Any individual effect was considered to be significant if the size of the absolute z-value was greater than 2.0. Results are presented for the simulated unidimensional and multidimensional data sets separately. It should be noted that, for the sake of simplicity, these effects will be presented with respect to the impact of the independent variable(s) only. For example, if the test length by rejection decision effect was significant, it is referred to as the effect of test length.

4.1. Unidimensional data sets

a. Approximate χ^2_I statistic

Table 3 shows the empirical Type I error rates for the approximate χ^2_I statistic with the simulated unidimensional data sets. Results obtained indicate that the assumption of unidimensionality was not rejected very often and that empirical α values were lower than the nominal Type I error rate (.05) for all simulated (unidimensional) test structures. In fact, the maximum number of rejections based on the 100 replications in any one condition was four for the cases having 45 items, 1000 examinees and either "moderate" (SAT-V) or "strong" (ASVAB-AR) dimension strengths. There were no incorrect rejections of the assumption of unidimensionality for the 15-item data sets whereas the number of rejections ranged from 0/100 to 2/100 for the 30-item test structures. Finally, empirical Type I error rates

Table 3

Empirical Type I error rates for the approximate χ^2_1 statistic: Unidimensional data sets

Test length	Sample size	Dimension strength	Empirical α
15 items	500 examinees	ACT-E	.00
		SAT-V	.00
		ASVAB-AR	.00
	1000 examinees	ACT-E	.00
		SAT-V	.00
		ASVAB-AR	.00
30 items	500 examinees	ACT-E	.01
		SAT-V	.00
		ASVAB-AR	.00
	1000 examinees	ACT-E	.02
		SAT-V	.01
		ASVAB-AR	.00
45 items	500 examinees	ACT-E	.00
		SAT-V	.03
		ASVAB-AR	.02
	1000 examinees	ACT-E	.02
		SAT-V	.04
		ASVAB-AR	.04

obtained with the longer (45-item) data sets were very close to the expected α, most notably for the 1000-examinee conditions where they were, as stated previously, only slightly below the nominal α for data sets simulated according to SAT-V and ASVAB-AR mean and variance item parameter values. The results of the logit-linear analysis indicate that a model including only the independent variable "test length" was sufficient to adequately explain the frequencies of rejection (and acceptance) rates ($\chi^2(16)=14.45$, $p=.565$). The effect of test length is quite clear. In data sets simulated with 15 items, there were zero rejections of the assumption of unidimensionality (from a total of 600 tests); with 30 items the null hypothesis was rejected four times; and with 45 items there were 15 rejections. Sample size and dimension strength did not significantly affect the probability of rejecting the null hypothesis of unidimensionality.

b. Approximate χ^2_2 statistic

Empirical Type I error rates obtained in each unidimensional condition for the second approximate χ^2 statistic are shown in Table 4. Results outlined in this table are nearly identical to those obtained with the previous χ^2 statistic in that the rejection of the assumption of unidimensionality was very infrequent and empirical Type I error rates were lower than the nominal alpha level (.05) for the 15- and 30-item data sets whereas they approached the expected α value for the 45-item

Table 4

Empirical Type I error rates for the approximate χ^2, statistic: Unidimensional data sets

Test length	Sample size	Dimension strength	Empirical α
15 items	500 examinees	ACT-E	.00
		SAT-V	.00
		ASVAB-AR	.00
	1000 examinees	ACT-E	.00
		SAT-V	.00
		ASVAB-AR	.00
30 items	500 examinees	ACT-E	.01
		SAT-V	.00
		ASVAB-AR	.00
	1000 examinees	ACT-E	.01
		SAT-V	.01
		ASVAB-AR	.00
45 items	500 examinees	ACT-E	.00
		SAT-V	.03
		ASVAB-AR	.02
	1000 examinees	ACT-E	.02
		SAT-V	.04
		ASVAB-AR	.04

simulated tests. In fact, the only difference between the two χ^2 statistics occurred with 30-item/1000 examinee data sets generated according to ACT-E item parameter means and variances where the empirical Type I error rate obtained with χ^2_2 (.01) was lower than the one obtained with χ^2_1 (.02). Logit-linear analysis results were also very similar to those obtained with the previous approximate χ^2 statistic. Indeed, a model incorporating test length as the sole independent variable was sufficient to adequately explain frequencies of acceptance and rejection rates ($\chi^2(16)=12.00$, $p=.744$). Specifically, there were no rejections of the assumption of unidimensionality with 15 item data sets whereas with 30 items the null hypothesis was rejected three times and with 45 items there were 15 rejections. Again, sample size and dimension strength did not significantly affect the probability of rejecting the null hypothesis of unidimensionality.

c. Approximate χ^2 difference tests

Empirical Type I error rates obtained with both χ^2 difference tests are presented in Table 5. Again, these difference tests were obtained by computing χ^2_1 and χ^2_2 values after fitting one- as well as two-factor models to unidimensional data sets. As is indicated in this table, the empirical Type I error rates are close to the nominal α values for the short data sets (15 items) only. Indeed, the empirical α increases as the number of items becomes larger. In fact, the assumption of

Table 5

Empirical Type I error rates for the approximate χ^2 difference tests: Unidimensional data sets

Test length	Sample size	Dimension strength	Empirical α: χ^2_1 DF test	Empirical α: χ^2_2 DF test
		ACT-E	.10	.10
	500	SAT-V	.02	.01
		ASVAB-AR	.02	.03
15 items				
		ACT-E	.06	.08
	1000	SAT-V	.05	.04
		ASVAB-AR	.06	.05
		ACT-E	.27	.30
	500	SAT-V	.17	.15
		ASVAB-AR	.30	.23
30 items				
		ACT-E	.17	.26
	1000	SAT-V	.29	.34
		ASVAB-AR	.62	.58
		ACT-E	.57	.64
	500	SAT-V	.38	.49
		ASVAB-AR	.58	.56
45 items				
		ACT-E	.51	.60
	1000	SAT-V	.65	.67
		ASVAB-AR	.93	.92

unidimensionality was incorrectly rejected in almost all instances (93 and 92 out of 100 trials for χ^2_1 and χ^2_2 difference tests, respectively) for data sets generated to contain 45 items, 1000 examinees and item parameter means and variances similar to those reported with the ASVAR-AR subtest.

c. Stout's T statistic

Empirical Type I error rates for Stout's T statistic with the same simulated unidimensional data sets are presented in Table 6. It is clear from these results that the actual Type I error rate was close to the nominal α level (.05) in most conditions that were simulated. In fact, the T-statistic yielded rejection rates in all conditions that did not fall outside of two standard deviations of a proportion based on a population proportion of .05. The results of the logit-linear analysis revealed that a model that includes none of the independent variables (i.e., a model that incorporates only the dependent variable "rejection decision") was sufficient to explain the observed frequencies ($\chi^2(10)=22.92$, $p=.152$). Test length, sample size and dimension strength had no effect on the probability of falsely rejecting the assumption of unidimensionality.

4.2. Multidimensional data sets

a. *Approximate χ^2_1 statistic*

Table 7 presents the frequency of rejection rates of the assumption of unidimensionality for the approximate χ^2_1 statistic

Table 6

Empirical Type I error rates for Stout's T statistic: Unidimensional data sets

Test length	Sample size	Dimension strength	Empirical α
15 items	500 examinees	ACT-E	.07
		SAT-V	.08
		ASVAB-AR	.04
	1000 examinees	ACT-E	.08
		SAT-V	.02
		ASVAB-AR	.02
30 items	500 examinees	ACT-E	.05
		SAT-V	.07
		ASVAB-AR	.03
	1000 examinees	ACT-E	.06
		SAT-V	.07
		ASVAB-AR	.02
45 items	500 examinees	ACT-E	.05
		SAT-V	.03
		ASVAB-AR	.02
	1000 examinees	ACT-E	.02
		SAT-V	.08
		ASVAB-AR	.08

Table 7

Number of rejections of the assumption of unidimensionality

for the approximate χ^2_1 statistic per 100 data sets

Test Length	Sample Size	Dimension Strength	Dimension Dominance	Rejections of one-dimensional model Approx. χ^2_1	Rejections of two-dimensional model Approx. χ^2_1
15	500	ACT-E	A	55	2
15	500	SAT-V	A	86	1
15	500	ASVAB-AR	A	97	1
15	500	ACT-E	B	99	0
15	500	SAT-V	B	100	0
15	500	ASVAB-AR	B	100	1
15	1000	ACT-E	A	100	3
15	1000	SAT-V	A	100	3
15	1000	ASVAB-AR	A	100	1
15	1000	ACT-E	B	100	4
15	1000	SAT-V	B	100	1
15	1000	ASVAB-AR	B	100	0
30	500	ACT-E	A	77	0
30	500	SAT-V	A	96	1
30	500	ASVAB-AR	A	100	0
30	500	ACT-E	B	100	3
30	500	SAT-V	B	100	2
30	500	ASVAB-AR	B	100	0
30	1000	ACT-E	A	100	2
30	1000	SAT-V	A	100	4
30	1000	ASVAB-AR	A	100	3
30	1000	ACT-E	B	100	5
30	1000	SAT-V	B	100	0
30	1000	ASVAB-AR	B	100	1
45	500	ACT-E	A	87	2
45	500	SAT-V	A	100	0
45	500	ASVAB-AR	A	100	2
45	500	ACT-E	B	100	0
45	500	SAT-V	B	100	0
45	500	ASVAB-AR	B	100	1
45	1000	ACT-E	A	99	1
45	1000	SAT-V	A	100	0
45	1000	ASVAB-AR	A	100	0
45	1000	ACT-E	B	100	0
45	1000	SAT-V	B	100	0
45	1000	ASVAB-AR	B	100	2

when data sets conformed to the dimension dominance A condition, that is, where 80% of the items require knowledge of the first ability and the remaining 20% necessitate knowledge of the second ability as well as the dimension dominance B condition, where 50% (or approximately 50%) of the items require knowledge of ability one and the remaining half require knowledge of the second latent trait. In addition, rejection rates obtained after fitting a two-dimensional model to the data sets are presented. Again, this allows us to determine if one or more than one ability is required to correctly answer a set of test items.

Results obtained from the logit-linear analysis using test length, sample size, dimension strength and dimension dominance as the independent variables and rejection decision based upon χ^2_1 as the dependent variable yielded a model that included the following effects (interacting with rejection decision): test length, sample size, dimension strength and dimension dominance, $\chi^2(31)=13.71$, $p<.997$. All of the effects had absolute z-values greater than or equal to two. The interpretation of these effects is quite straightforward. Summing across the levels of the other independent variables, it appears that the number of failures to reject unidimensionality decreased as test length increased.

Specifically, for the 1200 two-dimensional data sets that were generated for each test length, there were 63 failures to reject unidimensionality with 15 item tests, 27 failures, for 30 item tests, and 14 non significant tests of unidimensionality with the 45 item test length.

The approximate χ^2_1 statistic failed to reject the assumption of unidimensionality significantly more often for data sets containing 500 cases (103 out of 1800 tests) than data sets having 1000 examinees (once out of 1800 trials).

As well, with respect to dimension strength, the approximate χ^2_1 statistic failed to reject the assumption of unidimensionality for data sets generated according to the "weak" (i.e. ACT-E) dimensional structure 83 times (out of a total of 1200 trials) whereas it produced 18 non-significant tests for data sets simulated according to a "moderate" (i.e. SAT-V) structure and finally, three non significant tests for data sets generated according to ASVAB-AR item parameter means and variances.

Finally, dimension dominance was significantly related to the failure to reject unidimensionality given that the approximate χ^2_1 statistic yielded 103 (out of 1800 trials) non-significant tests of unidimensionality for data sets generated according to the dimension dominance A condition and only one non-significant test for those simulated according to the dimension dominance B condition.

Results that were obtained clearly show that the approximate χ^2_1 statistic performs well in most conditions. Indeed, based on this statistic, we were able to correctly reject the assumption of unidimensionality for 85% or more of dimension dominance A data sets in 16 out of 18 conditions outlined in Table 7. As expected, the approximate χ^2_1 statistic did not perform as well

with the 15-item/500 examinee and 30-item/500 examinee data sets generated according to ACT-E item parameter means and variances as well as the dimension dominance A condition where the assumption of unidimensionality was correctly rejected for 55 and 77 data sets, respectively.

Results obtained after fitting a two-factor model to the data sets also reveal that the approximate χ^2_1 statistic correctly identified the multidimensional nature of the dimension dominance A data sets in virtually all conditions. Indeed, the greatest number of incorrect rejections of the two-dimensional structure was quite low (4/100) and occurred with 30-item/1000 examinee data sets simulated according to SAT-V item parameter means and variances. Hence, it would appear as though the approximate χ^2_1 statistic was capable of confirming the multidimensional nature of these data sets.

It is also clear from the results in Table 7 that the approximate χ^2_1 statistic is extremely consistent in its ability to correctly reject the assumption of unidimensionality for dimension dominance B data sets across test lengths, sample sizes as well as dimension strengths. In fact, with the exception of 15-item/500 examinee data sets simulated according to ACT-E item parameter means and variances, where the assumption of unidimensionality was correctly rejected for 99/100 data sets, the approximate χ^2_1 statistic was able to correctly identify the multidimensional nature of all tests simulated to conform to the dimension dominance B condition.

In addition, results obtained after fitting a two-dimensional model to the latter data sets again show that the statistic was able to accurately identify the multidimensional nature of these simulated test structures. Indeed, decisions based on the approximate χ^2_1 statistic would overwhelmingly lead us to reject the null hypothesis that one dimension underlies these data sets, regardless of the condition examined. Indeed, the highest number of incorrect decisions based on the statistic (5/100), occurring with the 30-item/1000 examinee data sets generated according to ACT-E item parameter means and variances, is still acceptable to most practitioners.

b. *Approximate χ^2_2 statistic*

Table 8 presents the frequency of rejections of the assumption of unidimensionality for the approximate χ^2_2 statistic.

Logit-linear analysis results obtained with the second approximate χ^2 statistic are again very similar to those obtained with χ^2_1. Indeed, the model retained also revealed the same four significant two-way associations, that is, test length, sample size, dimension strength and dimension dominance, $\chi^2(31)=18.84$, $p<.925$. Once more, all of the effects had absolute z-values greater than or equal to two.

Results show that the number of failures to reject unidimensionality decreased as the number of items in the data

Table 8

Number of rejections of the assumption of unidimensionality for the approximate χ^2_2 statistic per 100 data sets

Test Length	Sample Size	Dimension Strength	Dimension Dominance	Rejections of one-dimensional model Approx. χ^2_2	Rejections of two-dimensional model Approx. χ^2_2
15	500	ACT-E	A	55	2
15	500	SAT-V	A	87	1
15	500	ASVAB-AR	A	97	1
15	500	ACT-E	B	99	0
15	500	SAT-V	B	100	0
15	500	ASVAB-AR	B	100	1
15	1000	ACT-E	A	100	3
15	1000	SAT-V	A	100	3
15	1000	ASVAB-AR	A	100	2
15	1000	ACT-E	B	100	4
15	1000	SAT-V	B	100	1
15	1000	ASVAB-AR	B	100	0
30	500	ACT-E	A	70	0
30	500	SAT-V	A	96	1
30	500	ASVAB-AR	A	100	0
30	500	ACT-E	B	100	3
30	500	SAT-V	B	100	2
30	500	ASVAB-AR	B	100	0
30	1000	ACT-E	A	100	2
30	1000	SAT-V	A	100	4
30	1000	ASVAB-AR	A	100	3
30	1000	ACT-E	B	100	5
30	1000	SAT-V	B	100	0
30	1000	ASVAB-AR	B	100	1
45	500	ACT-E	A	86	2
45	500	SAT-V	A	100	0
45	500	ASVAB-AR	A	100	2
45	500	ACT-E	B	100	0
45	500	SAT-V	B	100	0
45	500	ASVAB-AR	B	100	1
45	1000	ACT-E	A	98	1
45	1000	SAT-V	A	100	0
45	1000	ASVAB-AR	A	100	0
45	1000	ACT-E	B	100	0
45	1000	SAT-V	B	100	0
45	1000	ASVAB-AR	B	100	2

sets increased. More precisely, the assumption of unidimensionality was incorrectly accepted for 62 (out of 1200) 15-item data sets whereas there were, respectively, 34 and 16 non-significant χ^2_2 statistics for the 30- and 45-item data sets. Also, the approximate χ^2_2 statistic would lead us to incorrectly accept the assumption of unidimensionality more often for data sets containing 500 cases (110 out of 1800 trials) than data sets having 1000 cases (1 data set only).

Regarding dimension strength, the approximate χ^2_2 statistic failed to reject the assumption of unidimensionality for data sets simulated according to a "weak" (ACT-E) dimensional structure 92 times (out of 1200 trials) whereas it yielded 17 non-significant tests for data sets generated according to a "moderate" (SAT-V) structure and finally, produced three non-significant tests for data sets generated according to ASVAB-AR item parameter means and variances.

Finally, dimension dominance also played a role with regards to the failure to reject unidimensionality in that the approximate χ^2_2 statistic produced 111 (out of 1800 trials) non-significant tests of unidimensionality for data sets simulated according to the dimension dominance A condition and only one non-significant test for those generated according to the dimension dominance B condition.

As was the case with the unidimensional conditions, results obtained with the second approximate χ^2 statistic are similar to those reported with χ^2_1. The major difference between the two

statistics arose with the 30-item/500 examinee data sets generated according to ACT-E item parameter means and variances where χ^2_1 rejected the assumption of unidimensionality for a greater number of data sets (77/100) than χ^2_2 (70/100).

Also, the very small number of rejections obtained after fitting a two-dimensional factor structure to the data sets would again lead the researcher to accept, in 95% or more of cases, the assumption that more than one latent trait underlies responses on the simulated tests.

c. *χ^2 difference tests*

As was previously mentioned, difference tests were also computed for each approximate χ^2 statistic. The number of rejections of the assumption of unidimensionality based on these two difference tests is shown in Table 9.

Results indicate that the χ^2 difference tests rejected the assumption of unidimensionality for a larger number of data sets than either approximate χ^2 statistics. Indeed, the lowest rejection rate (93/100), occurring with 15-item/500 examinee dimension dominance A data sets generated according to ACT-E item parameter means and variances is high and would confirm the multidimensional nature of the data sets, regardless of the simulation condition. The high number of rejections obtained with these difference tests clearly confirm the multidimensional nature of the data sets. Indeed, the χ^2 difference tests were able to correctly reject the assumption of unidimensionality for every dimension dominance B data set generated, regardless of the

Table 9

Number of rejections of the assumption of unidimensionality for the approximate χ^2 difference tests per 100 data sets

Test Length	Sample Size	Dimension Strength	Dimension Dominance	Number of Rejections $\chi^2_{1.05}$	Number of Rejections $\chi^2_{1.10}$
15	500	ACT-E	A	93	95
15	500	SAT-V	A	99	99
15	500	ASVAB-AR	A	100	100
15	500	ACT-E	B	100	100
15	500	SAT-V	B	100	100
15	500	ASVAB-AR	B	100	100
15	1000	ACT-E	A	100	100
15	1000	SAT-V	A	100	100
15	1000	ASVAB-AR	A	100	100
15	1000	ACT-E	B	100	100
15	1000	SAT-V	B	100	100
15	1000	ASVAB-AR	B	100	100
30	500	ACT-E	A	100	100
30	500	SAT-V	A	100	100
30	500	ASVAB-AR	A	100	100
30	500	ACT-E	B	100	100
30	500	SAT-V	B	100	100
30	500	ASVAB-AR	B	100	100
30	1000	ACT-E	A	100	100
30	1000	SAT-V	A	100	100
30	1000	ASVAB-AR	A	100	100
30	1000	ACT-E	B	100	100
30	1000	SAT-V	B	100	100
30	1000	ASVAB-AR	B	100	100
45	500	ACT-E	A	100	100
45	500	SAT-V	A	100	100
45	500	ASVAB-AR	A	100	100
45	500	ACT-E	B	100	100
45	500	SAT-V	B	100	100
45	500	ASVAB-AR	B	100	100
45	1000	ACT-E	A	100	100
45	1000	SAT-V	A	100	100
45	1000	ASVAB-AR	A	100	100
45	1000	ACT-E	B	100	100
45	1000	SAT-V	B	100	100
45	1000	ASVAB-AR	B	100	100

simulation condition.

d. Stout's T statistic

The number of rejections of the assumption of unidimensionality using Stout's T statistic are presented in Table 10. It is important to note that since Stout's T-statistic is used only to test the assumption of unidimensionality, it was impossible to determine the number of times that the (correct) two-dimensional hypothesis was rejected.

Logit-linear analysis results indicate that the simplest model adequately explaining decision frequencies included the following effects of the independent variables (interacting with rejection decision): test length, sample size, dimension strength, dimension dominance, test length by dimension dominance, and dimension strength by dimension dominance, $\chi^2(25)=31.62$, $p<.169$.

It is clear that Stout's T-statistic yielded many more failures to reject unidimensionality for data sets generated according to the dimension dominance A condition (320 out of 1800 trials) than those simulated according to the dimension dominance B condition (89 out of 1800 trials). As well, many more decision errors resulted in the 15 item condition (356 out of 1200 trials) than in either the 30 item (34 out of 1200) or 45 item (19 out of 1200) data sets.

The interaction between dimension dominance and test length can be explained by the difference in the number of failures to reject the null hypothesis of unidimensionality between the 30

Table 10

Number of rejections of the assumption of unidimensionality for the approximate Stout's T-statistic per 100 data sets

Test Length	Sample Size	Dimension Strength	Dimension Dominance	Number of Rejections
15	500	ACT-E	A	19
15	500	SAT-V	A	51
15	500	ASVAB-AR	A	68
15	500	ACT-E	B	69
15	500	SAT-V	B	90
15	500	ASVAB-AR	B	89
15	1000	ACT-E	A	32
15	1000	SAT-V	A	64
15	1000	ASVAB-AR	A	86
15	1000	ACT-E	B	85
15	1000	SAT-V	B	95
15	1000	ASVAB-AR	B	96
30	500	ACT-E	A	79
30	500	SAT-V	A	98
30	500	ASVAB-AR	A	100
30	500	ACT-E	B	97
30	500	SAT-V	B	100
30	500	ASVAB-AR	B	99
30	1000	ACT-E	A	96
30	1000	SAT-V	A	100
30	1000	ASVAB-AR	A	100
30	1000	ACT-E	B	99
30	1000	SAT-V	B	99
30	1000	ASVAB-AR	B	99
45	500	ACT-E	A	91
45	500	SAT-V	A	99
45	500	ASVAB-AR	A	100
45	500	ACT-E	B	100
45	500	SAT-V	B	97
45	500	ASVAB-AR	B	99
45	1000	ACT-E	A	97
45	1000	SAT-V	A	100
45	1000	ASVAB-AR	A	100
45	1000	ACT-E	B	99
45	1000	SAT-V	B	99
45	1000	ASVAB-AR	B	100

item and 45 item test lengths. Indeed, the number of false acceptances of the null hypothesis decreased from 27 (for the 30 item tests) to 13 (for the 45 item tests) for dimension dominance A data sets but remained stable for dimension dominance B data sets (seven and six incorrect decisions for the 30 and 45 item test lengths, respectively).

With regards to the dimension dominance by dimension strength interaction, it appears that, generally, the weak dimension strength (ACT-E) yielded many more false acceptances of the assumption of unidimensionality than did the other two dimension strengths (SAT-V and ASVAB-AR). As was previously the case, it seems that this interaction is attributable to a small difference in the number of failures to reject unidimensionality between the moderate and strong dimension strengths for the dimension dominance B condition as compared to the dimension dominance A condition. In the latter condition, the number of false acceptances of the assumption of unidimensionality decreased from 186 (out of 600 data sets) for data sets generated according to ACT-E item parameter means and variances to 88 for SAT-V data sets and finally, 46 for tests simulated to conform to the ASVAB-AR configuration. The trend is, however, much less noticeable for data sets in the former condition. Indeed, the number of false acceptances of the assumption of unidimensionality was 51 (out of 600 data sets) for "weak" dimension strength data sets whereas it was 20 for "moderate" dimension strength data sets and finally, 18 for tests simulated

according to ASVAB-AR item parameter means and variances.

Finally, there was an overall difference in the accuracy of the T-statistic between data generated with 500 cases and 1000 cases. The assumption of unidimensionality was falsely accepted 255 times (from a total of 1800) with sample sizes of 500 as compared to 154 incorrect decisions when sample sizes of 1000 were used.

Results outlined in Table 10 show that the T statistic performs quite poorly with regards to being able to correctly reject the assumption of unidimensionality for the majority of 15-item dimension dominance A data sets. Indeed, with the exception of 1000 examinee data sets simulated according to ASVAB-AR item parameter means and variances, the T statistic was only able to correctly reject the assumption of unidimensionality in less than 70% of instances. Also, Stout's T statistic did not perform well with 30-item/500 examinee data sets simulated according to ACT-E item parameter means and variances. Here, the assumption of unidimensionality was correctly rejected for 79 data sets. Finally, as was the case for the approximate χ^2 statistics, results indicate that the T statistic has a high degree of accuracy in rejecting the assumption of unidimensionality for the remaining 30-item data sets and all 45-item simulated test structures.

Also, the accuracy of the T-statistic varies from 69/100 to 100/100 correct rejections of the assumption of unidimensionality for data sets simulated according to the dimension dominance B

condition. As was the case with the approximate χ^2 statistics, Stout's T statistic was also generally quite accurate in rejecting the assumption of unidimensionality across conditions. However, contrary to the former statistics, its performance was less than adequate in a few conditions, most notably with 15-item/500 examinee data sets simulated according to ACT-E item parameter means and variances, where the assumption of unidimensionality was correctly rejected in only 69/100 instances.

CHAPTER 5

DISCUSSION

The purpose of this research was to investigate the usefulness of two approximate χ^2 statistics with regards to assessing the assumption of unidimensionality, central to not only modern (i.e. IRT) but also classical test theory. Specifically, these approximate χ^2 statistics were based on McDonald's "weak" principle of local independence and derived from his nonlinear factor analytic model. According to this author, if the "weak" principle of local independence has been met after fitting a unidimensional model (i.e. one-factor model), that is, zero residual correlations, we can assume that a single latent trait underlies responses to a set of items. The two approximate χ^2 statistics conformed to this view of dimensionality in that they tested the null hypothesis that a set of residual correlations were equal to zero following a nonlinear factor analysis of a set of item responses. In addition, difference tests, based on both approximate χ^2 statistics, were also computed as an additional means of determining if fitting a more complex model (one- versus two-factor model) would significantly improve fit and hence suggest that a second latent trait was required to account for item responses. Finally, the performance of these approximate χ^2 statistics and difference tests was compared to that of Stout's T statistic, another promising technique in the area of dimensionality assessment. Results obtained in the previous chapter of the dissertation are

discussed below. First, findings pertaining to the three dimensionality assessment techniques examined, that is, the approximate χ^2 statistics, the χ^2 difference tests as well as Stout's T statistic will be outlined for both the unidimensional and multidimensional studies. Second, a comparison of these techniques will be made. Specifically, which statistic(s) seem(s) to function better in a given condition? Third, a broader discussion of the results obtained with these procedure with regards to other dimensionality assessment statistics/indices will be presented. Finally, the implications of the findings of this study to the practitioner will be examined. More precisely, recommendations regarding the use of the statistics investigated in this study will be offered.

5.1. Approximate χ^2 statistics

Results that were obtained with both approximate χ^2 statistics were virtually identical and hence will be discussed concurrently.

a. Unidimensional study

Logit-linear analysis results suggest that the only variable that seemed to affect the accuracy with which the assumption of unidimensionality was correctly accepted, based on either the first or second approximate χ^2 statistic, was test length. Specifically, it appears as though the empirical Type I error rate approached the expected α value as the number of items, that is, degrees of freedom, increased. The most plausible explanation

for the lower than expected Type I error probability levels encountered with most unidimensional data sets could be related to the the fact that the statistics examined in this study follow an *approximate* χ^2 distribution. Hence, it would appear as though the fit between the empirical and theoretical (i.e. chi-square) distributions is closer with the longer (45-item) data sets that were simulated.

b. Multidimensional study

As was the case with the unidimensional conditions, results obtained with data sets simulated according to multidimensional test structures A and B were virtually identical for both approximate χ^2 statistics. Again, data sets generated to conform to the dimension dominance A condition reflected tests that were composed of one dominant dimension and a weaker second dimension whereas dimension dominance B data sets denote a clearer two-dimensional structure.

Logit-linear analyses that were performed in order to assess how rejection rates obtained from both approximate χ^2 statistics were affected by factors such as test length, sample size, dimension strength as well as dimension dominance were quite similar for the latter two procedures. Indeed, the models retained for the two approximate χ^2 statistics were nearly identical.

The two approximate χ^2 statistics performed very well in correctly rejecting the assumption of unidimensionality and hence correctly identifying the multidimensional nature of data sets

simulated according to the dimension dominance A condition. Indeed, findings show that both procedures correctly rejected the assumption of unidimensionality for over 85% of the data sets simulated in 16 out of 18 multidimensional conditions. Logit-linear analysis results also reveal that the proportion of rejections of the assumption of unidimensionality increased as more information became available in the simulated data sets. Specifically, the proportion of rejections based on χ^2_1 and χ^2_2 was higher for data sets that contained a larger number of items (45) and examinees (1000) as well as those that were comprised mainly of high discriminating items (referred to as ASVAB-AR item parameter means and variances in the text). Indeed, the performance of both approximate χ^2 statistics was less powerful with 15- as well as 30-item data sets containing 500 examinees and generated according to ACT-E item parameter means and variances, where the assumption of unidimensionality was rejected for less than 80% of the data sets. This poorer performance is probably attributable to the limited amount of information available in the data sets (only 15/30 items and 500 examinees) as well as the (generally) low discriminating items that were generated in this particular condition. In a factor analytic framework, this would translate into factors loading poorly on the majority of the items in the test. Hence, χ^2_1 results for these particular data sets were not entirely unexpected. Indeed, it is probably unrealistic to believe that any procedure would be able to correctly identify the presence of a second dimension

when the latter trait is defined by only three items which is the case with the 15-item data sets simulated according to dimension dominance A. Also, it is important to note that the number of rejections of the assumption of unidimensionality improved considerably with data sets generated according to SAT-V and ASVAB-AR item parameter means and variances, even with as few as 15 items and 500 examinees. One would expect (or at least, hope) that most "real" test structures are generally comprised of items that have higher discrimination parameter values than those simulated within the "weak" dimension strength condition in this study. Finally, the only marked difference between the two approximate χ^2 statistics arose with the 30-item/500 examinee data sets based upon ACT-E item parameter means and variances where χ^2_1 slightly outperformed χ^2_2.

Regarding dimension dominance B, results show that the power of the two approximate χ^2 statistics was identical with regards to rejecting the assumption of unidimensionality. Indeed, χ^2_1 and χ^2_2 correctly rejected the assumption of unidimensionality for every data set with the exception of one test simulated to contain 15 items, 500 examinees as well as item parameter means and variances similar to those reported with the ACT-E subtest. Also, as was previously suggested, the poorer performance of both approximate χ^2 statistics is not totally unexpected in this condition given the small number of items as well as sample size and more importantly, the weak factor loadings. In essence, this would correspond to the (probably) unrealistic situation whereby

two ill-defined abilities would be required to correctly answer a set of test items.

5.2. χ^2 difference tests

Given that some authors have suggested that χ^2 statistic values tend to be related to sample size (Balla, Marsh, & McDonald, 1990), difference tests were also computed for each of the above mentioned procedures as an ancillary means of assessing the number of dimensions underlying a set of item responses. Results obtained with both approximate χ^2 statistics were again quite similar for both unidimensional as well as multidimensional data sets. Regarding the unidimensional study, results clearly show that the difference tests were unable, for the majority of data sets, to correctly accept the assumption that only one latent trait was accounting for item responses. Indeed, with the exception of 15-item data sets, these two procedures overwhelmingly rejected the assumption of unidimensionality in all conditions.

Results obtained in the multidimensional study were again quite comparable in that rejection rates obtained with the two procedures were quite high (93% or more of the data sets were correctly identified as being multidimensional). In fact, the difference tests were always at least as accurate as either χ^2_1 or χ^2_2, regardless of the simulation condition examined. Indeed, rejection rates obtained with the difference tests were high, even in conditions that yielded less favourable results with the approximate χ^2 statistics. However, these results must be

interpreted very cautiously in light of findings obtained with unidimensional data sets. The high rejection rates associated with the two χ^2 difference tests are probably attributable to their very liberal nature. As was previously pointed out, these two procedures performed quite poorly with unidimensional data sets due to their inflated Type I error probabilities. Hence, it is likely that the high rejection rates obtained with the multidimensional data sets merely reflect the over "sensitivity" of the difference tests as dimensionality assessment techniques. This might explain the unusually high number of rejections obtained with these two methods.

Therefore, preliminary findings would suggest that χ^2 difference tests not be used to determine if one or more than one latent trait underlies a set of item responses. Though the procedures had very high rejection rates with multidimensional data sets, the inflated empirical Type I error rates with unidimensional data sets cast a serious doubt over the usefulness of both tests. Hence, future research should be undertaken to examine the usefulness of the difference tests in a larger number of conditions before recommending their use as one (of several) dimensionality assessment techniques.

5.3. Stout's T statistic

a. Unidimensional study

Logit-linear analysis results showed that the accuracy with which Stout's T statistic was able to correctly accept the assumption of unidimensionality was not affected by any of the

predictors examined in this study. Indeed, the statistic performed just as well with short, moderate or long tests; small or larger samples; and data sets comprised mainly of either low, average or highly discriminating items. Thus, results obtained with Stout's T statistic parallel those from previous studies (Nandakumar, 1987; 1988; 1989; Stout, 1987). With unidimensional test structures, the empirical Type I error probabilities were generally very close to the nominal α values. Hence, the T statistic is quite accurate in correctly identifying various unidimensional test structures.

Indeed, the T statistic was able to correctly determine that *essentially* one ability was required to answer a set of test items even in conditions that had not previously been looked at such as certain data sets containing 15 items and 500 examinees.

b. Multidimensional study

Logit-linear analysis results show that the model retained with regards to Stout's T statistic was less parsimonious than the one fitted to the two approximate χ^2 statistics' rejection rates. The model contained two three-way associations which suggests that the effects of the predictors (test length, sample size, dimension strength and dimension dominance) on the performance of the T statistic are considerably more difficult to model than those based upon the approximate χ^2 statistics.

The accuracy with which Stout's T statistic is capable of rejecting the assumption of unidimensionality for data sets generated according to the dimension dominance A condition seems

to be highly influenced by test length. Indeed, with the exception of 15 item data sets containing 1000 examinees and simulated to have item parameter means and variances similar to those reported with the ASVAB-AR subtest, the technique was unable to reject the assumption of unidimensionality in 32% or more of the shorter (15-item) data sets. However, it is also important to point out that the performance of the T statistic improved considerably with longer tests (30 or more items) containing 1000 examinees. Again, these findings are consistent with those reported by Nandakumar (1987; 1988; 1989) who strongly advised against using the procedure with less than 25 items. The accuracy with which Stout's T statistic was able to correctly reject the assumption of unidimensionality was also high for the majority of data sets simulated according to the dimension dominance B condition. However, the effectiveness of the T statistic was questionable in some conditions, most notably with 15-item data sets generated according to ACT-E item parameter means and variances where the assumption of unidimensionality was correctly rejected for only 69% of the data sets. In other words, the T statistic appeared to be unable to correctly identify the multidimensional nature of two-factor data sets simulated to contain mainly low discriminating items even though each latent trait was defined by approximately the same number of items. However, with longer and larger data sets containing items with higher discrimination parameter values, the T statistic was able to correctly reject the assumption of unidimensionality in nearly

all instances. It is also important to point out that results obtained with dimension dominance B data sets are somewhat conflicting with Nandakumar's (1987) conclusions which warned against using the procedure with tests that contain less than 25 items. Indeed, our findings show that the T statistic was quite accurate in rejecting the assumption of unidimensionality for most 15-item data sets that conformed to the dimension dominance B condition. In conclusion, it appears as though the performance of the T statistic is not solely affected by test length or sample size, as was evident from the logit-linear analysis results. In addition, it appears to be highly dependent upon dimension strength, at least more so than was the case with the approximate χ^2 statistics. A comparison of the various procedures examined in this study is presented in the next section of the discussion.

5.4. A comparison of the approximate χ^2 statistics, the difference tests and Stout's T statistic

a. Unidimensional study

Results obtained with both approximate χ^2 statistics as well as Stout's T statistic were very similar in that empirical Type I error rates were close to (expected) nominal values for unidimensional conditions examined. However, as was previously pointed out, empirical Type I error probabilities associated with Stout's T statistic were generally closer to expected values than those obtained with the two approximate χ^2 statistics. The most plausible explanation for the low Type I error probabilities

noted for the two latter statistics lies in the fact they that are *approximately* distributed as a central χ^2 distribution. Hence, it is possible, most notably for the 15-item data sets, that the fit between empirical and theoretical distributions is poor. Indeed, as one might expect due to the asymptotic nature of the χ^2 distribution, the fit appears to improve as the number of items increases as is reflected with the 45-item data sets where empirical and nominal α values are much closer. Finally, it also appears that the χ^2 difference tests are not recommended due to the inflated Type I error rates encountered with 30- and 45-item data sets.

b. Multidimensional study

With regards to data sets that were generated to conform to the dimension dominance A condition (i.e. one dominant and one minor ability) or B (two distinct dimensions), it appears fairly obvious that the most accurate procedures, overall, were the two approximate χ^2 statistics. Indeed, with the exception of two conditions (15- and 30-item data sets containing 500 examinees and generated according to ACT-E item parameter means and variances), the two techniques were able to correctly reject the assumption of unidimensionality for nearly all data sets. Again, the poorer performance of the two approximate χ^2 statistics in the latter two conditions is somewhat predictable given the very low discrimination parameters associated with the majority of the items in these conditions. In essence, given these very low factor loadings, the program is unable to fit a satisfactory two-

factor solution to the latter data sets. Finally, the number of rejections of the assumption of a two-dimensional structure was also quite low for all (multidimensional) data sets which again, corroborates the usefulness of the two procedures in that they appear to be able to correctly fail to reject the null hypothesis that specifies the correct number of factors that underlies the item responses.

The performance of Stout's T statistic was, however, more difficult to assess than that of the approximate χ^2 tests. On the one hand, results obtained in this study indicate that the procedure is generally not very accurate in rejecting the assumption of unidimensionality with data sets that contain few items (15) and examinees (500) as well as a "weaker" two-dimensional structure. These findings are consistent with Nandakumar's (1987; 1988; 1989; 1991a; 1991b) results. Again, its performance is poorer with data sets generated according to ACT-E item parameter means and variances which was also the case with the two approximate χ^2 statistics. However, Stout's T statistic was very accurate in rejecting the assumption of unidimensionality with data sets simulated according to a distinct two-dimensional structure (dimension dominance B) which contradicts past findings that it should not be applied with short test lengths. It seems fairly clear from these results, and most notably from logit-linear analysis findings, that the accuracy of the T statistic is highly dependent upon the particular dimensional structure of the items, at least more so

than for the other procedures examined. Indeed, though the performance of the approximate χ^2 statistics was poorer in one or two conditions (15 item/500/examinee/ACT-E/dimension dominance A), it was generally quite accurate in all other conditions examined. Stout's T statistic, on the other hand, was generally ineffective with 15 item data sets simulated to conform to the dimension dominance A condition.

In conclusion, the more parsimonious logit-linear models fitted to the approximate χ^2 statistics seem to suggest that the usefulness of these procedures may generalize to a larger number of test structures than Stout's T statistic. In other words, the accuracy with which the latter procedure rejects the assumption of unidimensionality seems to be dependent upon more complex interactions between factors than the approximate χ^2 statistics. Roznowski, Tucker and Humphreys (1991) in discussing the desirable characteristics of indices used to assess the dimensionality of a set of binary items state,

> "...it is important to have an index that is both robust to changes in levels of parameters and lacks substantial interactions among parameters" (p.124).

Although none of the procedures examined in this study exhibited these properties, results suggest that the two approximate χ^2 statistics were less affected by interactions among independent variables than Stout's T statistic. Hence, the former two statistics might be useful in a greater number of conditions than the latter statistic.

However, it is important to once more point out that these findings are preliminary and that more research should be done with varying dimensional structures before any definite conclusions are made regarding any of the procedures investigated in the present study.

5.5. Comparing the approximate χ^2 statistics and difference tests to other dimensionality assessment procedures

In the second chapter of this dissertation, a review of methods that have been proposed to assess test dimensionality was outlined. Based on results obtained in this research, how do the two approximate χ^2 statistics and difference tests compare with these techniques?

First, results obtained with the approximate χ^2 statistics were generally very favourable and seem to suggest, at the very least, that these procedures warrant further attention. Although the usefulness of these techniques was assessed within a somewhat restrictive framework (i.e. a Monte Carlo study), findings nonetheless show that the statistics might provide valuable information to the practitioner interested in investigating the dimensional structure of a test.

Second, the two approximate χ^2 statistics are derived from a very sound theoretical model, namely, McDonald's polynomial approximation to a normal ogive model and his accompanying principle of "weak" local independence. Hence, contrary to several dimensionality assessment procedures that have been proposed in the literature (e.g. indices based on principal

components analysis, NMDS, etc.) the approximate χ^2s are based on a very clear (theoretical) definition of dimensionality. Also, given that they (might) be approximately distributed as a central χ^2 distribution, it is possible to objectively investigate their accuracy in a host of conditions, unlike descriptive indices.

In addition, contrary to Stout's T statistic, the approximate χ^2 statistics could conceivably be used to assess not only departure from unidimensionality but also the optimal number of latent traits that are required to correctly answer a set of test items.

Also, the approximate χ^2 statistics are derived from a (nonlinear factor analytic) model which has been shown to be equivalent to common IRT approaches. Given that NOHARMII provides Lord's reparameterized difficulty and discrimination values, this model should be familiar to most practitioners and can be easily applied to not only assess test dimensionality but also obtain the latter item parameter estimates. Indeed, McDonald's model as well as the accompanying computer program NOHARMII can enable the practitioner to not only assess the assumption of unidimensionality but also the fit of a given IRT model. In this particular study, for example, two-parameter models were fitted to each data set, that is, the lower asymptote parameter was not taken into account. It would have been possible to assess the fit of a model with constrained factor loadings (discrimination parameters equal to one) in order to examine the feasibility of a Rasch-type model. Hence, the usefulness of McDonald's model is

much broader than the issue of dimensionality assessment and it could conceivably be used to examine a multitude of other educational measurement problems. For example, it is quite possible that the model could be applied to test the equality of factor loadings across different subgroups as a means of determining if a set of items function differentially. These types of issues could be addressed with the nonlinear factor analytic approach proposed by McDonald.

In conclusion, the approximate χ^2 statistics and difference tests investigated in this study appear to be promising not only from an empirical standpoint (i.e. results obtained in this study) but also due to the strong theoretical foundation on which they rest. In the next section, the implications of the results obtained in this study to the practitioner will be discussed.

5.6. Implications of results to the practitioner

Based on the results obtained in this study, what suggestions could be put forth to the practitioner in search of a procedure that would enable him or her to determine if the assumption of unidimensionality has been compromised for a given set of item responses?

It is important to underline, a priori, that not one dimensionality assessment method that has been proposed either in this study or in the past literature is without shortcomings. Hence, it would be ill-advised to rely solely on one statistic when examining the problem of dimensionality. In fact, the practitioner should strive to amass information from several

procedures in the hopes that results will converge towards a similar solution. Having said this, what can be recommended with regards to the statistics examined in this study?

It is clear that all procedures examined in this study possess certain weaknesses. Indeed, logit-linear analysis results show that several factors seem to affect the accuracy with which the various procedures were able to either correctly accept or reject the assumption of unidimensionality. Except for Stout's T statistic with unidimensional data sets, each factor examined (test length, sample size, dimension strength and dimension dominance) negatively affected the performance of the procedures.

Having said this, it appears that the two most promising procedures examined in this study were the approximate χ^2 statistics, that is, χ^2_1 and χ^2_2 for the following four reasons:

- The two statistics had acceptable Type I error rates with unidimensional data sets and high rejection rates with two-dimensional data sets even when the second latent trait was defined by as little as six items.
- McDonald's nonlinear factor analytic model and accompanying computer program (NOHARMII) allow the researcher to not only compute the approximate χ^2 statistics but also provide both unidimensional as well as multidimensional item parameter estimates.
- NOHARMII enables the researcher to assess not only unidimensionality or departure from that assumption but also the fit of a two-, three-, four-, etc. dimensional model.

- Finally, NOHARMII allows the researcher to assess the actual fit of a given IRT model, for example, a two- versus three-parameter logistic model.

However, it also important to stress that preliminary results show that the two approximate χ^2 statistics should be interpreted cautiously with certain test structures containing 15 items and 500 examinees (most notably those that might include several items with low discrimination parameter values). Also, difference tests based on the latter two approximate χ^2 statistics are not recommended due to their inflated Type I error rates.

Though Stout's T statistic performed very well with unidimensional data sets, its usefulness was questionable with certain 15-item tests. However, rejection rates improved considerably with data sets simulated to clearly require two abilities (i.e. dimension dominance B) and those containing 30 or more items. Hence, the procedure could be used confidently by the practitioner interested in testing the hypothesis of unidimensionality for certain test structures containing as few as 15 items but more assuredly for data sets containing 30 or more items and 1000 or more cases.

However, given some of the advantages associated with McDonald's model that were previously enumerated, it would seem to be reasonable to compute the two approximate χ^2 statistics when assessing the dimensionality of a set of item responses and to, whenever possible, compare results to those obtained with

other procedures in order to arrive at a more informed decision.

CHAPTER 6

SUMMARY AND CONCLUSION

In this final chapter of the dissertation, a summary of the research and conclusions are presented. Specifically, the main findings will be outlined as well as certain limitations associated with the study and suggestions for future research in this area.

6.1. SUMMARY

As was previously mentioned, IRT models have been used extensively in the past decade not only in the development and analysis of educational test items but also in a host of other situations such as the equating of alternate test forms and the detection of differentially functioning items, to name a few. Indeed, the many advantages of IRT models, namely invariance of item and ability parameter estimates, have contributed to their ever increasing use by psychometricians in order to resolve a multitude of measurement-related problems. However, the majority of these models can be used legitimately only when a set of particular assumptions have been met, one of which is unidimensionality of the latent space. Indeed, most IRT models assume that only one latent trait underlies a set of item responses. This latent trait is usually interpreted as being an ability that is required on the part of the examinee to correctly answer a set of test items. For example, a general verbal ability is hypothesized to underlie the responses of examinees on items

comprising the GRE verbal subtest. However, the assumption of unidimensionality is rarely met with actual achievement test data given the multitude of factors that can come into play when answering an item. For example, it is quite conceivable that a mathematics item might require mathematical, verbal as well as reasoning abilities.

Given the apparent unrealistic nature of the assumption of unidimensionality, several researchers sought to examine the robustness of unidimensional item and ability parameter estimates when obtained from multidimensional data. Results showed that these unidimensional estimates were often biased in that they poorly recovered multidimensional ("true") parameter values. Thus, it became clear that attempting to fit (unidimensional) IRT models to sets of item responses that obviously conformed to a multidimensional structure could produce meaningless parameter estimates. Research was therefore geared towards the development of indices and/or statistics that would help the practitioner determine if the assumption of unidimensionality had been met or not for a given data set. Although indices based on a variety of models were proposed such as those derived from principal component analysis, linear factor analysis, multidimensional scaling, etc., these often proved to be very inaccurate and consequently of little use. However, results obtained with regards to two dimensionality assessment procedures, that is, Stout's T statistic as well as indices based on McDonald's nonlinear factor analytic model, did show promise. In addition,

contrary to the majority of techniques proposed, each procedure was based upon a sound theoretical framework. Indeed, Stout's T statistic was based on his concept of *essential dimensionality* whereas nonlinear factor analytic indices stemmed from the "weak" principle of local independence. Specifically, the T statistic tests the assumption that *essentially* one latent trait underlies item responses. In essence, the procedure is not very sensitive to minor dimensions which are often present in actual achievement test data. Therefore, the T statistic tests the null hypothesis that conditional item covariances tend towards zero as the number of items reaches infinity. Results obtained with the T statistic show that it can often correctly determine if *essentially* one or more than one latent trait underlies a set of item responses. However, its performance is questionable with certain data sets containing few items (less than 25) and small sample sizes (less than 750 examinees). On the other hand, nonlinear factor analytic indices, such as the mean absolute residual covariance and the sum of squares of the residual covariances are based on the "weak" principle of local independence, that is, conditional item covariances equal to zero after fitting a one-factor model to a set of item responses. Given that the relationship between observed item responses and the underlying latent trait is often assumed to contain a nonlinear component, McDonald has suggested fitting a nonlinear model (specifically, a polynomial approximation to a normal ogive model) to the matrix of item covariances. Results obtained with several indices based on this

principle seem to confirm the usefulness of McDonald's model with regards to the problem of determining if one or more than one latent trait underlies a set of item responses. However, indices that have been proposed and examined so far are descriptive in nature and thus offer no firm criterion that would help the researcher determine if the "weak" principle of local independence has been met after fitting a one-factor model, that is, if the assumption of unidimensionality holds for a particular data set.

The purpose of this research was therefore to investigate two approximate χ^2 statistics as well as χ^2 difference tests that are based on the "weak" principle of local independence and thus might possibly be used to help the practitioner determine if one or more than one latent trait is required to correctly answer a set of test items. In addition, results obtained with these procedures were compared to Stout's T statistic. Both approximate χ^2 statistics test the null hypothesis that the off-diagonal elements of a residual correlation matrix are equal to zero after fitting a specific factor model (for example, a unidimensional model). Specifically, we were interested in evaluating the performance of these two statistics with data sets simulated to be unidimensional and multidimensional in nature. In the unidimensional study, empirical Type I error probabilities were compared to nominal α levels in several conditions that varied according to test length, sample size as well as dimension strength. Generally, the approximate χ^2 statistics results showed

that the actual Type I error rates tended to be closer to the nominal ones as test length increased. Also, Type I error rates obtained with the χ^2 difference tests were very large. Finally, actual and expected Type I error probabilities obtained with Stout's T statistic were very similar across different conditions.

The second part of the research entailed examining the power of χ^2_1, χ^2_2, difference tests based on the approximate χ^2 statistics as well as Stout's T statistic with various multidimensional data sets. Specifically, rejection rates of the assumption of unidimensionality based on these statistics were computed for data sets that varied according to test length, sample size, dimension strength as well as dimension dominance. Results showed that all procedures, with the exception of χ^2_1 DF and χ^2_2 DF, performed very well with regards to correctly identifying the multidimensional nature of the data sets. The performance of the χ^2 difference tests, however, was questionable in that the very high rejection rates noted were probably attributable to the inflated Type I error rates obtained with unidimensional data sets. Finally, logit-linear analyses were carried out in order to assess how the frequency of rejections of the assumption of unidimensionality was affected by test length, sample size, dimension strength as well as dimension dominance. Separate analyses were carried out for χ^2_1, χ^2_2 and the T statistic. Results related to the two approximate χ^2 statistics generally indicated that the proportion of rejection rates was

higher for longer tests, larger samples, data sets containing mainly (high) discriminating items and simulated according to the dimension dominance B condition. However, the logit-linear model fitted to Stout's T rejection rates was considerably less parsimonious. The results, as was previously suggested, could be attributable to the particular multidimensional test structures that were examined in this study. Some of the limitations associated with this research will be discussed in the next section of the chapter.

6.2. LIMITATIONS OF THE RESEARCH

The main limitations of this study are those often encountered with Monte Carlo studies and deal chiefly with the conditions examined. A few of these shortcomings are outlined below.

First, the data sets were generated according to a two-parameter logistic model. Hence, the lower asymptote parameter was not estimated and set at zero. This might be unrealistic in some testing situations where we expect some low ability examinees to correctly answer a set of items due to guessing or other factors.

Second, the number of replications performed in each condition (100) might be considered less than optimal by some researchers. However, the number of replications performed is sufficient to investigate the general performance of the statistics. In addition, the number of replications done in this

study is far greater than that often found in dimensionality assessment studies, for example those of Berger & Knol (1990) and Knol and Berger (1991) where only 15 replications were performed.

Third, the item parameter means and variances that were chosen to simulate the unidimensional as well as multidimensional data sets were somewhat arbitrary. Nonetheless, these values enabled us to investigate the functioning of the approximate χ^2 and T statistics in three distinct conditions (that is, tests containing mainly low, moderate and high discriminating items).

Fourth, the multidimensional test structures (i.e. dimension dominance A and B) examined might not be representative of the majority of actual examinee item responses. However, given that this was an initial investigation of the approximate χ^2 statistics, it was felt that it would be more prudent to examine their usefulness in conditions whereby each latent trait was defined by a unique set of items, that is, each factor loaded on distinct items. Therefore "mixed" items (defining more than one latent trait) were not included in any of the simulated data sets. Hopefully, results obtained in this study will foster future research that may assess the effectiveness of the approximate χ^2 statistics in a wider range of conditions.

Fifth, it would seem reasonable to believe that the performance of the two approximate χ^2 statistics might be affected by larger samples. One of the problems typically encountered with χ^2 distributed statistics is their inflated Type I error rates with large sample sizes (Siegel & Castellan, 1988).

Hence, it is possible that the empirical Type I error rates obtained with both approximate χ^2 statistics might be affected by sample sizes exceeding those examined in this study (i.e. 1000 examinees).

Sixth, it must also be stressed that in the multidimensional study, abilities were simulated to be uncorrelated. In other words, the performance of the various statistics was assessed in a somewhat idealistic set of conditions. However, as was pointed out earlier, it was felt that it was necessary to control this factor given that this was an initial investigation of the procedures. It is important to first find out if the procedures function well in a set of restricted conditions before examining the effect of more complex factors. However, given that data sets tend to be more unidimensional as the correlation between abilities increases, it is plausible to hypothesize that the approximate χ^2 statistics would be less accurate in correctly identifying the multidimensional nature of data sets with correlated factors.

Finally, it must also be remembered that any resultant χ^2 is dependent upon the theoretical model and type of estimation used. An important issue then, is the degree of accuracy among estimates obtained with NOHARM as compared to those with other alternatives such as LISCOMP (Muthen, 1985) and TESTFACT (Wilson, Wood, & Gibbons, 1987). Both NOHARM and LISCOMP use "limited information" (i.e., pairwise information in the items) while TESTFACT is more theoretically sound in its use of "full-

information" methods. The validity of the approximate χ^2 depends in part on the degree to which full-information methods will yield more accurate parameter estimates as compared to models using limited information. It has been suggested that the loss of information should not be great when one does not use higher-order marginals (McDonald, 1991; Muthen, 1978). As well, Knol and Berger (1991) in their simulation study found little difference between the results of NOHARM and TESTFACT in recovering factor analytic parameters. However, their results must be interpreted cautiously due to the small number of replications (15) in each condition. The large amount of computing resources needed to run TESTFACT also limits its practical use to relatively few items and dimensions. Although more research is needed, it would seem that, from a practical perspective, there is not much to be gained in using full-information methods.

6.3. SUGGESTIONS FOR FUTURE RESEARCH

In light of results obtained in this research and limitations associated with the unidimensional and multidimensional studies, the following suggestions are offered regarding future investigations that might be undertaken in this area.

First, it would be important to assess the performance of the approximate χ^2 statistics with data sets simulated according to a three-parameter logistic model. Indeed, this would allow researchers to determine how the latter statistics are affected

by the lower asymptote parameter.

Second, the accuracy with which the two statistics reject the assumption of unidimensionality should be examined with a larger number of test structures. For example, in order to emulate actual test data, conditions containing mixed items and varying proportions of items defining each latent trait should be simulated. In addition, item parameters should be simulated according to different means and variances.

Third, future research should investigate the robustness of the indices to violations of underlying assumptions. For example, the effect of non-normality of the underlying latent variable(s) should be assessed. Browne (1986), in looking at this issue in the general factor analytic framework, has shown that both estimation and significance tests are quite insensitive to non-normality of the latent variables.

Fourth, it would seem possible that some indices of fit that are popular in the general factor analytic or structural equation modelling literature might be appropriate in the assessment of the dimensionality of a set of binary items. A review of these indices is found in McDonald and Marsh (1990).

Fifth, the effect of larger sample sizes on the performance of the approximate χ^2 statistics should be further investigated. Indeed, it would seem important to determine if the empirical Type I error probabilities of the latter two procedures are inflated by sample sizes exceeding 1000 which is often the case with these types of statistics.

Finally, the effect of correlated abilities on the two approximate χ^2 statistics should also be assessed. One would expect that the performance of the latter two procedures would worsen as the correlation between abilities increased (i.e. as the data sets become more "unidimensional"). Therefore, the correlation between the abilities should be varied in order to determine if the statistics are adversely affected in any way.

Despite the limitations associated with this research, preliminary results were very encouraging and suggest that the two approximate χ^2 statistics might be helpful to test developers who are interested in determining if the assumption of unidimensionality, crucial to IRT item analyses, has been met. At the very least, findings indicate that the two statistics are worthy of further research and attention.

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.

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

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

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

Bartholomew, D.J. (1983). Latent variable models for ordered categorical data. Journal of Econometrics, 22, 229-243.

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

Bejar, I.I. (1980). A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. Journal of Educational Measurement, 17, 283-296.

Bejar, I.I. (1988). An approach to assessing unidimensionality revisited. Applied Psychological Measurement, 12, 377-379.

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

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.

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

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord and M.R. Novick, Statistical theories of mental test scores. Reading, MA: Addison Wesley.

Bishop, Y.M.M., Fienberg, S.E., & Holland, P.W. (1975). Discrete multivariate analysis: Theory and practice. Cambridge, MA: MIT Press.

Bock, D.R., & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. Psychometrika, 4, 443-459.

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

Bock, R.D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. Psychometrika, 35, 179-197.

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

Browne, M.W. (1986). Robustness of statistical inference in factor analysis and related models (Research Report 86-1). Pretoria: Department of Statistics, University of South Africa.

Carlson, J.E. (1987). Multidimensional item response theory estimation: A computer program (Research Report No. ONR87-2). Iowa City, IA: The American College Testing Program.

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

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

Collins, L.M., Cliff, N., McCormick, D.J., & Zatkin, J.L. (1986). Factor recovery in binary data sets: A simulation. Multivariate Behavioral Research, 21, 377-391.

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

De Ayala, R.J., & Hertzog, M.A. (1989, March). A comparison of methods for assessing dimensionality for use in Item Response Theory. Paper presented at the meeting of the National Council on Measurement in Education, San Francisco, CA.

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

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 39, 1-38.

Doody, E.N. (1985). Examining the effects of multidimensional data on ability and item parameter estimation using the three-parameter logistic model (Report No. TM 850 360). Monterey, CA: CTB/McGraw-Hill. (ERIC Document Reproduction Service No. ED 258 992).

Doody-Bogan, E.N., & Yen, W.M. (1983, April). Detecting multidimensionality and examining its effects on vertical equating with the three-parameter logistic model. Paper presented at the meeting of the American Educational Research Association, Montreal, PQ.

Dorans, N.J., & Kingston, N.M. (1985). The effects of violations of unidimensionality on the estimation of item and ability parameters and on Item Response Theory equating of the GRE verbal scale. Journal of Educational Measurement, 22, 249-262.

Dorans, N.J., & Lawrence, I.M. (1988, April). An item parcel approach to assessing the dimensionality of test data. Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.

Drasgow, F., & Lissak, R.I. (1983). Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. Journal of Applied Psychology, 68, 363-373.

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

Eddins, J.M. (1984). Dimensionality, scoring and related problems in adaptive testing (Report No. 83-5-R). Urbana, Il: University of Illinois, Computer-Based Education Research laboratory. (ERIC Document Reproduction Service No. ED 243 929).

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

Goldstein, H. (1980). Dimensionality, bias, independence and measurement scale problems in latent trait test score models. British Journal of Mathematical and Statistical Psychology, 33, 234-246.

Goldstein, H., & Wood, R. (1989). Five decades of item response modelling. British Journal of Mathematical and Statistical Psychology, 42, 139-167.

Gourlay, N. Difficulty factors arising from the use of the tetrachoric correlations in factor analysis. British Journal of Statistical Psychology, 4, 65-72.

Green, S.B. (1983). Identifiability of spurious factors using linear factor analysis with binary items. Applied Psychological Measurement, 7, 139-147.

Hambleton, R.K., & Cook, L.L. (1977). Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 14, 75-96.

Hambleton, R.K., & Rogers, J.H. (1986). Promising directions for assessing item response model fit to test data (Report No. TM 860 372). Amherst, MA: University of Massachussets. (ERIC Document Reproduction Service No. ED 270 489).

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

Hambleton, R.K., & Swaminathan, H. (1985). Item response theory: Principles and applications, Boston, MA: Kluwer-Nyjhoff.

Harrison, D.A. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. Journal of Educational Statistics, 11, 91-115.

Hattie, J. (1982). Decision criteria for determining dimensionality. Dissertation Abstracts International, 42, 4415-A.

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.

Hayduk, L.A. (1987). Structural equation modelling with Lisrel: Essentials and advances. Baltimore, MD: The Johns Hopkins University Press.

Holland, P.W. (1981). When are item response models consistent with observed data? Psychometrika, 46, 79-92.

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

Horn, J.L. (1965). A rationale and test for the number of factors in factor analysis. Psychometrika, 30, 179-185.

Hsu, T.C., & Yu, L. (1989). Using computers to analyze item response data. Educational Measurement: Issues and Practice, 8, 21-28.

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

Jones, P.B. (1988, April). Assessment of dimensionality in dichotomously-scored data using multidimensional scaling: Analysis of HSMB data. Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.

Jones, P.B., Sabers, D.L., & Trosset, M. (1987). Dimensionality assessment for dichotomously scored items using multidimensional scaling (Report No. TM 870 416). Tucson, AZ: University of Arizona. (ERIC Document Reproduction Service No. ED 283 877).

Junker, B. (1988). User's guide to computer programs for Stout's unidimensionality statistic. Pittsburgh, PA: Carnegie-Mellon University.

Junker, B.W. (1991, April). Structural robustness and ability estimation in item response theory: A survey. Paper presented at the meeting of the American Educational Research Association, Chicago, IL.

Junker, B.W., & Stout, W.F. (1991, September). Robustness of ability estimation when multiple traits are present with one trait dominant. Paper presented at the international Symposium on Modern Theories in Measurement: Problems and Issues, Montebello, PQ.

Kingsbury, G.G. (1985). A comparison of item response theory procedures for assessing response dimensionality (Report No. TM 850 477). Portland, OR: Portland Public Schools. (ERIC Document Reproduction service No. ED 261 075).

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

Kingston, N.M., & McKinley, R.L. (1988, April). Assessing the structure of the GRE general test using confirmatory multidimensional Item Theory. Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.

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

Koch, W.R. (1983). The analysis of dichotomous test data using nonmetric multidimensional scaling (Report No. TM 830 617). Austin, TX: The University of Texas at Austin. (ERIC Document Reproduction service No. ED 235 204).

Kruskal, J.L., & Wish, M. (1978). Multidimensional scaling. Beverly Hills, CA: Sage.

Lazarfeld, P.F. (1950). The logical and mathematical foundation of latent structure analysis. In S.A. Stouffer et al. (Eds.), Measurement and Prediction. Princeton, NJ: Princeton University Press.

Liou, M. (1988). Unidimensionality versus statistical accuracy: A note on Bejar's method for detecting dimensionality of achievement tests. Applied Psychological Measurement, 12, 381-386.

Loehlin, J.C. (1987). Latent variable models: An introduction to factor, path and structural analysis. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F.M. A theory of test scores. Psychometric Monograph, 1952, No. 7.

Lord, F.M. (1953). The relationship of the test score to the trait underlying the test. Educational and Psychological Measurement, 13, 517-548.

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

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

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

McDonald, R.P. (1979). The simultaneous estimation of factor loadings and scores. British Journal of Mathematical and Statistical Psychology, 32, 212-228.

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

McDonald, R.P. (1982a). Some alternative approaches to the improvement of measurement in education and psychology: Fitting latent trait models. In D. Spearitt (Ed.), The improvement of measurement in education and psychology (pp. 213-233). Hawthorn, VI: Australian Council for Educational Research.

McDonald, R.P. (1982b). 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. (1989a). Future directions for item response theory. International Journal of Educational Research, 13, 205-220.

McDonald, R.P. (1989b). An index of goodness-of-fit based on noncentrality. Journal of Classification, 6, 97-103.

McDonald, R.P. (1991, November). Testing for approximate dimensionality. Paper presented at the International Symposium on Modern Theories in Measurement: Problems and Issues, Montebello, PQ.

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

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

McKinley, R.L. (1983, April). A multidimensional extension of the two-parameter logistic latent trait model. Paper presented at the meeting of the National Council on Measurement in Education, Montreal, PQ.

McKinley, R.L. (1988, April). Assessing dimensionality using confirmatory multidimensional IRT. Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.

McKinley, R.L., & Reckase, M.D. (1983). An application of a multidimensional extension of the two-parameter logistic latent trait model. Iowa City, IA: The American College Testing Program. (ERIC Document Reproduction Service No. ED 240 168).

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

Marsh, H.W., Balla, J.R., & McDonald, R.P. (1988). Goodness-of-fit in confirmatory factor analysis: The effect of sample size. Psychological Bulletin, 103, 391-410.

Miller, T. & Hirsch, T. (1991, April). Applying multidimensional item response theory (MIRT) and cluster analysis to the study of dimensional parallelism of alternate test forms. Paper presented at the meeting of the American Educational Research Association, Chicago, Il.

Mislevy, R.J. (1986). Recent developments in the factor analysis of categorical variables. Journal of Educational Statistics, 11, 3-31.

Morgan, R. (1989, March). An examination of the dimensional structure of the ATP biology achievement test. Paper presented at the meeting of the National Council on Measurement in Education, San Francisco, CA.

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

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

Muthen, B. (1983). Latent variable structural equation modelling with categorical data. Journal of Econometrics, 22, 43-65.

Muthen, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. Psychometrika, 49, 115-132.

Muthen, B. (1985). LISCOMP. Mooresville, IN: Scientific Software.

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

Nandakumar, R. (1988, April). Modification of Stout's procedure for assessing latent trait unidimensionality. Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.

Nandakumar, R. (1989, March). Traditional dimensionality vs essential dimensionality. Paper presented at the meeting of the American Educational Research Association, San Francisco, CA.

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

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

Norusis, M.J. (1988). SPSS-X advanced statistics guide. Chicago, IL: SPSS Inc.

Pandey, T.N., & Carlson, D. Application of item response models to reporting assessment data. In R.K. Hambleton (Ed.), Applications of Item Response Theory, Vancouver, BC: Educational Research Institute of British Columbia.

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. (1981). Guessing and dimensionality: The search for a unidimensional latent space (Report No. TM 810 389). Columbia, MI: University of Missouri. (ERIC Document Reproduction Service No. ED 204 394).

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

Reckase, M.D. (1986a, April). The discriminating power of items that measure more than dimension. Paper presented at the meeting of the American Educational Research Association, San Francisco, CA.

Reckase, M.D. (1986b, June). The interpretation of unidimensional IRT parameters when estimated from multidimensional data. Paper presented at the meeting of the Psychometric Society, Toronto, ON.

Reckase, M.D., Carlson, J.E., Ackerman, T.A., & Spray, J.A. (1986, June). The interpretation of unidimensional IRT parameters when estimated from multidimensional data. Paper presented at the meeting of the Psychometric Society, Toronto, Ont.

Reynolds, T.J. (1976). The analysis of dominance matrices: Extraction of unidimensional orders within a multidimensional context (Technical Report No. 3). Los Angeles, CA: University of Southern California, Department of Psychology.

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

Roznowski, M., Tucker, L.R., & Humphreys, L.G. (1991). Three approaches to determining the dimensionality of binary items. Applied Psychological Measurement, 15, 109-127.

Samejima, F. (1974). Normal ogive model on the continuous response level in the multidimensional latent space. Psychometrika, 39, 111-121.

Siegel, S., & Castellan, N.J. (1988). Nonparametric statistics for the behavioral sciences. New York, NY: McGraw-Hill Book Company.

Steiger, J.H. (1980a). Tests for comparing elements of a correlation matrix. Psychological Bulletin, 87, 245-251.

Steiger, J.H. (1980b). Testing pattern hypotheses on correlation matrices: Alternative statistics and some empirical results. Multivariate Behavioral Research, 15, 335-352.

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.

Swaminathan, H., & Gifford, J.A. (1983). Estimation of parameters in the three-parameter latent trait model. In D. Weiss (Ed.), New horizons in testing. New York, NY: Academic Press.

Sympson, J.B. (1978). A model for testing with multidimensional items. In D.J. Weiss (Ed.), Proceedings of the 1977 computerized adaptive testing conference. Minneapolis, MN: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

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

Traub, R.E. (1983). A priori considerations in choosing an item response model. In Hambleton, R.K. (Ed.) Applications of item response theory. Vancouver, BC: Educational Research Institute of British Columbia.

Traub, R.E., & Lam, Y.R. (1985). Latent structure and item sampling models for testing. Annual Review of Psychology, 36, 19-48.

Tucker, L.R., Humphreys, L.G., & Roznowski, M. (1986). Comparative accuracy of five indices of dimensionality of binary items (Technical Report No. 1). Champaign, IL: University of Illinois. (ERIC Document Reproduction Service No. ED 271 515).

Wang, M.M. (1988, April). Measurement bias in the application of a unidimensional model to multidimensional Item Response data. Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.

Warm, T.A. (1978). A primer of Item Response Theory (Tech. Rep. No. OG-941278) Oklahoma City, OK: U.S. Coast Guard Institute.

Way, W.D., Ansley, T.N., & Forsyth, R.A. (1988). The comparative effects of compensatory and noncompensatory two-dimensional data on unidimensional IRT estimates. Applied Psychological Measurement, 12, 239-252.

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

Wise, S.L. (1981). Some comparisons of four order-analytic methods and factor analysis for assessing dimensionality (Report No. 81-2). Urbana, IL: University of Illinois, Computer-Based Education Research Laboratory. (ERIC Document Reproduction Service No. ED 208 027).

Yen, W.M. (1983). Use of the three-parameter model in the development of a standardized achievement test. In R.K. Hambleton (Ed.), Applications of item response theory, Vancouver, BC: Educational Research Institute of British Columbia.

Zegers, F.E., & Ten Berge, J.M.F. (1983). A fast and simple computational method of minimum residual factor analysis. Multivariate Behavioral Research, 18, 331-340.

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

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

ABSTRACT

Assessing Test Dimensionality Using Two Approximate Chi-Square Statistics

The application of Item Response Theory (IRT) relies on the strong assumptions underlying the models, one of which is unidimensionality of the latent space. This assumption entails that item response probabilities are a function of a single underlying ability that would entirely explain item covariances. This has lead to the development of a multitude of statistical indices to assess the dimensionality of a set of test items. Indices have been based on linear factor analysis / principle component analysis, the Holland-Rosenbaum procedure, the assessment of *essential dimensionality*, and those based on the nonlinear factor analysis of a matrix of pairwise joint-proportions of item responses.

The latter two areas appear to be the most theoretically sound and promising. Stout's T statistic, testing for *essential dimensionality*, has been shown to be quite accurate over a wide variety of conditions; however, its usefulness is questionable with short test lengths (less than 25 items) and small sample sizes (less than 750 examinees) (Nandakumar, 1991a; 1991b; Stout, 1987; 1991). Indices based on nonlinear factor analysis, most notably the mean absolute residual covariance (Berger & Knol, 1990) and the incremental fit index (De Champlain & Gessaroli, 1991) also seem to be useful in the assessment of test dimensionality. However, they are purely descriptive and hence offer little criteria on which to assess the level of dimensionality.

Two procedures for assessing the dimensionality of a set of test items that are based on the computation of approximate χ^2 statistics are investigated in this study. The two statistics are based on McDonald's (1981) "weak" principle of local independence in that they test the null hypothesis that a matrix of residual correlations is equal to zero after fitting a unidimensional (i.e. one-factor) model. In addition, χ^2 difference tests (for a one- versus two-factor solution) were computed for two-dimensional data sets. Finally, results were compared to those obtained with Stout's T statistic. The main advantage of the two approximate χ^2 statistics and difference tests is that they involve actual hypothesis testing and are not merely descriptive indices. Hence, they may offer values on which decisions regarding the dimensionality of a set of test items may be based.

The purposes of this study were therefore to examine the extent to which the approximate χ^2 statistics, χ^2 difference tests and Stout's T statistic were able to correctly identify the number of dimensions underlying both unidimensional and multidimensional (simulated) data sets. Also, for the approximate χ^2 and Stout's T statistics, logit-linear analyses were performed in order to assess how rejection decisions obtained with each procedure were affected by test length, sample size, dimension strength as well as dimension dominance.

In order to examine these research problems, unidimensional and two-dimensional data sets were generated which varied according to test length, sample size as well as dimension strength. In addition, two-dimensional data sets varied according to dimension dominance. Abilities were randomly generated from a N(0,1) distribution. Finally, there were 100 replications in each condition.

With regards to the unidimensional study, results show that test length significantly affected the accuracy with which the two approximate χ^2 statistics could identify the unidimensional nature of the data sets, that is; empirical Type I error probabilities tended to be lower than expected (nominal) α values for shorter test lengths (15-item data sets) and close to expected values for the longer (45-item) data sets whereas they were inflated with the difference tests in almost all conditions. Results obtained with Stout's T statistic, on the other hand, indicate that the empirical α values were very close to nominal values, regardless of any particular condition.

Multidimensional results suggest that the rejection rates obtained with all statistics examined were affected by predictors and in that sense, none proved to be optimal. However, the two approximate χ^2 statistics did enable the rejection of the assumption of unidimensionality in a large number of conditions, including those that contained as few as 15 items and they were less affected by predictors than Stout's T statistic. Rejection rates obtained with Stout's T statistic were also quite high with longer tests and larger sample sizes.

In conclusion, results obtained in this study seem to suggest that the approximate χ^2 statistics can be used as means of assessing the dimensionality of short or long tests whereas Stout's T statistic can be helpful in the analysis of data sets containing 30 or more items and 1000 or more examinees.

Appendix A

Takane & De Leeuw's proofs (1987) that IRT and NLFA models are equivalent

The following set of proofs will show that the two-parameter normal ogive model, popular in IRT modelling, is mathematically equivalent to a nonlinear factor analytic model, in this case Christoffersson's factor analysis of dichotomized variables (Christoffersson, 1975).

Let $\underline{\mathbf{x}}'$ $(\underline{x}_1 \ldots, \underline{x}_n)$ be a random vector of response patterns to n binary items on a test. Each $\underline{x}_i$ is assigned a value of 1, if the examinee correctly answers the item, or 0, if there is an incorrect response. Let $\underline{\mathbf{u}}$ be an m-component random vector of abilities ($m \leq n$) with its density function denoted by $g(\mathbf{u})$. $\underline{\mathbf{u}}$ is unobservable directly, but is assumed to follow a multivariate normal distribution with mean $\mathbf{0}$ and covariance I (identity matrix); that is $\underline{\mathbf{u}} \sim N(0,(I))$. The domain of $\underline{\mathbf{u}}$ (denoted by U) is the multidimensional region defined by the direct product of $(-\alpha,\alpha)$.

In IRT, the two-parameter normal ogive model specifies the marginal probability that $\underline{\mathbf{x}} = \mathbf{x}$ (Bock & Aitkin, 1981; Bock & Lieberman, 1970) as,

$$Pr(\underline{X}=x) = \int_U Pr(\underline{\mathbf{x}}=x|\mathbf{u})\, g(\mathbf{u})\, d\mathbf{u} \qquad (1)$$

where $Pr(\underline{\mathbf{x}}=\mathbf{x}|\mathbf{u})$ is the conditional probability of observing response pattern $\mathbf{x}$ given $\underline{\mathbf{u}} = \mathbf{u}$. Also, it is assumed that,

$$Pr(\underline{\mathbf{x}}=\mathbf{x}|\mathbf{u}) = \prod_i^n (p_i(\mathbf{u}))^{x_i} (1-p_i(\mathbf{u}))^{1-x_i} \qquad (2)$$

$$p_i(\mathbf{u}) = \int_{-\infty}^{\mathbf{a}'\mathbf{u}+b} \phi(z)\, dz = \Phi(\mathbf{a}'\mathbf{u}+b) \qquad (3)$$

(that is, local independence) with,

where ϕ is the density function of the standard normal distribution and Φ, the normal ogive function (i.e., the cumulative distribution function of the standard normal distribution).

On the other hand, in the factor analytic model proposed by Christoffersson (1975), the marginal probability of response pattern $\mathbf{x}$ is specified as,

$$Pr(\mathbf{x}=\mathbf{x}) = \int_R h(\mathbf{y})\, d\mathbf{y} \qquad (4)$$

where R is the multidimensional region of integration and

$$\mathbf{y} = \mathbf{C}\mathbf{u} + \mathbf{e} \qquad (5)$$

Equation 5 corresponds to the common factor analytic model with $\mathbf{C}$ being the matrix of factor loadings, $\mathbf{u}$, the vector of factor scores (abilities in an IRT framework) and $\mathbf{e}$, the random vector of uniqueness components distributed as $N(\mathbf{0}, \mathbf{Q}^2)$ where $\mathbf{Q}^2$ is further assumed to be diagonal (linear local independence), and $\mathbf{u}$ and $\mathbf{e}$ are independent of each other. It follows that,

$$\mathbf{y} \sim N(\mathbf{0}, \mathbf{C}\mathbf{C}' + \mathbf{Q}^2) \qquad (6)$$

(marginal distribution of $\mathbf{y}$) and

$$\underline{\mathbf{y}}|\mathbf{u} \sim N(C\mathbf{u}, Q^2) \tag{7}$$

(conditional distribution of $\underline{\mathbf{y}}$ given $\underline{\mathbf{u}} = \mathbf{u}$). The continuous random variables, $\underline{\mathbf{y}}$ are dichotomized by $\underline{x}_i = 1$, if $\underline{y}_i \geq r_i$ -or- o, if $\underline{y}_i < r_i$ for i=1,...n, where r_i is the threshold parameter for variable i. Therefore, R, the region of integration above, is the multidimensional parallelepiped defined by the direct product of intervals, $R_i = (r_i, \alpha)$ if $\underline{\mathbf{x}}=1$ and $R_i=(-\alpha, r_i)$ if $\underline{\mathbf{x}} = 0$. Now (1) including (2) and (3) is equivalent to (4) with $\underline{\mathbf{y}}$ defined in (5). We first prove that (4) → (1). From (4) we have

$$Pr(\underline{\mathbf{x}}=\mathbf{x}) = \int_R h(\mathbf{y})\, d\mathbf{y} \tag{8}$$

$$= \int_R \left(\int_U f(\mathbf{y}|\mathbf{u})\, g(\mathbf{u})\, d\mathbf{u} \right) d\mathbf{y} \tag{9}$$

$$= \int_U g(\mathbf{u}) \left(\int_R f(\underline{\mathbf{y}}|\mathbf{u})\, d\mathbf{y} \right) d\mathbf{u}, \tag{10}$$

where $f(\underline{\mathbf{y}}|\mathbf{u})$ is the conditional density of $\underline{\mathbf{y}}$ given $\underline{\mathbf{u}}=\mathbf{u}$. But because of (7) we have

$$\int_R f(\mathbf{y}|\mathbf{u})\,d\mathbf{y}=\prod_i \int_{R_i} f_i(y_i|\mathbf{u})\,dy_i \qquad (11)$$

$$=\prod_i \left(\int_{r_i}^{\alpha} f_i(y_i|\mathbf{u})\,dy_i\right)^{x_i} 1-\left(\int_{r_i}^{\alpha} f_i(y_i|\mathbf{u})\,dy_i\right)^{1-x_i} \qquad (12)$$

where

$$\int_{r_i}^{\alpha} f_i(y_i|\mathbf{u})\,dy_i=\Phi\left(\frac{c_i\mathbf{u}-r_i}{q_i}\right) \qquad (13)$$

for $i = 1,\ldots,n$. In this instance q^2 is the i-th diagonal element of Q^2. Equation (12) is thus equivalent to (3) by setting

$$a_i=\frac{c_i}{q_i} \qquad (14)$$

and

$$b_i=-\frac{r_i}{q_i} \qquad (15)$$

for $i = 1,\ldots,n$.

At first glance, it appears as though factor analysis with c_i, r_i and $q_i(i=1,...,n)$ has more parameters than IRT with only a_i and $b_i(i=1,...n)$. However, according to the authors, when the data are dichotomous, the variance of $\mathbf{y}$ cannot be estimated due to the lack of relevant information in the data, and thus, q_i can be set to an arbitrary value. Hence, effective number of parameters is identical in both models. In conclusion, the authors mention that the equivalence of marginal probabilities in IRT and FA models holds approximately with logistic (IRT) models also, as long the logistic distribution provides a good approximation of the normal distribution (i.e. normal ogive).

Appendix B

Stout's T statistic

Stout's procedure involves the following seven steps:

1. Dividing the items into three subtests

Firstly, the total number of items must be divided into two short assessment subtests (AT1 and AT2) of length M as well as a longer partitioning subtest of length n referred to as the partitioning test (PT) (Stout, 1987).

A linear factor analysis of a tetrachoric matrix or expert judgment (content analysis) is utilized to determine which items are to be included in AT1. Items are selected so that they are as unidimensional as possible, i.e., items with the highest loadings of the same sign on the second extracted factor are retained for AT1. Indeed, Nandakumar (1987) states that, in the unidimensional case, second factor loadings should ideally be random with values near zero whereas in the multidimensional case, the second factor should exhibit a distinct pattern of loadings, that is, all items of the same ability will have loadings of the same sign (positive or negative). Thus, items with the highest second factor loadings of the same sign are chosen for AT1 and deemed to be unidimensional items, both when d=1 and d>1.

Items in AT2 are chosen so that they have the same difficulty distribution as those found in AT1 and they are selected to be representative of the remaining items on the test. In other words, if very difficult items are retained in AT1, similarly difficult items will be assigned to AT2. The objective of AT2 is to correct for the pre-asymptotic statistical bias in Stout's statistic.

Finally, the remaining items are included in the PT which enables the researcher to group examinees into subgroups of approximately equal ability and size.

2. Assign examinees to subgroups

Based on their score on the partitioning test (PT), assign examinees to different subgroups. Examinees who correctly or incorrectly answer all of the items on PT are excluded from the remaining analyses. For example, if n=20, there will be 19 subgroups. Finally, all subgroups containing less than 20 examinees are usually eliminated.

3. Compute the "usual" variance estimate for the *k*-th subgroup for AT1

The AT1 score of the j-th examinee from subgroup k is equal to,

$$Y_j^{(k)} = \sum_{i=1}^{M} U_{ijk}/M$$

where $Y_j^{(k)}$ = the proportion correct score for examinee j from subgroup k.

U_{ijk} = the response of the jth examinee to the ith item from subgroup k.

The mean examinee AT1 score for subgroup k is given by,

$$\overline{Y}^{(k)} = \sum_{j=1}^{J_k} Y_j^{(k)}/J_k \qquad (2)$$

and the variance estimate of examinee AT1 scores from subgroup k is

equal to,

$$\hat{\sigma}_k^2 = \sum_{j=1}^{J_k} (Y_j^{(k)} - \overline{Y}^{(k)})^2 / J_k \qquad (3)$$

4. Compute the unidimensional variance estimate for the *k*th subgroup for AT1.

Let,

$$\hat{P}_i^{(k)} = \sum_{j=1}^{J_k} U_{ijk} / J_k \qquad (4)$$

and,

$$\hat{\sigma}_{U,k}^2 = \sum_{i=1}^{M} \hat{P}_i^{(k)} (1 - \hat{P}_i^{(k)}) / M^2$$

be the unidimensional variance estimate for subgroup *k*.

5. Normalize and combine the different subgroup variance estimates to form the statistic T_L for AT1.

Let,

$$\hat{\mu}_{4,k} = \sum_{j=1}^{J_k} (Y_j^{(k)} - \overline{Y}^{(k)})^4 / J_k \qquad (6)$$

and,

$$\hat{\delta}_{4,k} = \sum_{i=1}^{M} \hat{p}_i^{(k)} (1 - \hat{p}_i^{(k)}) (1 - 2\hat{p}_i^{(k)})^2 \qquad (7)$$

also,

$$S_k^2=[(\hat{\mu}_{4,k}-\hat{\sigma}_k^4)+\delta_{4,k}/M^4+2\sqrt{(\hat{\mu}_{4,k}-\hat{\sigma}_k^4)\,\delta_{4,k}/M^4}]/J_k \quad (8)$$

Now, let

$$T_L=\frac{1}{K}\sum_{k=1}^{K}[\hat{\sigma}_k^2-\frac{\hat{\sigma}_{U,k}^2}{S_k}] \quad (9)$$

where L stands for long test and K corresponds to the number of subgroups remaining after those with too few examinees have been discarded in step 2.

6. Compute the statistic T_B on AT2 items

Go through steps 1 to 5 for items included in AT2 and compute T_B (correction for bias) according to equation 36.

7. Compute Stout's T statistic for *essential* dimensionality

Stout's T-statistic for the assessment of *essential* dimensionality is given by

$$T=\frac{T_L-T_B}{\sqrt{2}} \quad (10)$$

The significance level associated with the statistic is obtained by referring to the upper tail of the standard normal distribution. Hence, large *p*-values would be indicative a multidimensional test whereas low *p*-values would suggest the test is *essentially* unidimensional.

8. Modification of Stout's T statistic

Nandakumar (1987) has also proposed a correction procedure for Stout's T statistic which reduces bias when the test contains only items with high discrimination values. Indeed, this researcher has shown that the Type I error rate associated with the T statistic tends to be inflated when the discrimination indices of items are too high. Specifically, a spurious factor, which is composed of easy items with high factor loadings, is identified by the statistic (Nandakumar, 1987). Hence, the easy items are included in the first assessment test (AT1) whereas the partitioning test (PT) is left with mainly difficult items. Consequently, the various subgroups (except for the high ability group) tend to be comprised of examinees who vary highly in ability which results in large within-group variability and violation of the assumption of local independence. This will lead to the (incorrect) rejection of the assumption of *essential* unidimensionality. The procedure involves computing a Wilcoxon rank sum test and can be presented as follows:

(i) Rank the n items in a test from most difficult (1) to least difficult (N).

(ii) Ascertain the ranks of M items in AT1 and compute the sum of these ranks (W_s).

(iii) Calculate the mean $E(W_s)$ and the standard deviation $SD(W_s)$ of the sum W_s under the assumption of randomly distributed ranks,

$$E(W_s) = .5M(N+1) \qquad \textbf{(11)}$$

$$SD(W_s) = \frac{1}{12} M(N-M)\ (N+1))^{1/2} \qquad (12)$$

(iv) Compute the critical value C given by,

$$C = \mathbf{E}(W_s) + Z_\alpha (SD(W_s)) \qquad (13)$$

where Z_α = the 100(1-α)-th percentile of a $N(0,1)$ distribution, α corresponding to the user-specified level of significance.

(v) Finally, if W_s > C, conclude that M items in AT1 are all too easy.

If the items are deemed to be too easy, items with the highest factor loadings of the opposite (negative, for example) are selected for AT1. If the items are regarded as being sufficiently heterogeneous with regards to difficulty, they are retained for AT1.

Appendix C

Examples of data sets conforming to dimension dominance conditions A and B

Hypothetical two-dimensional 15 item ACT-E data set

Dimension dominance A

ITEM NUMBER	ITEM DESCRIPTION	a_1	a_2
1	Pure item of θ_1	0.85	0.00
2	Pure item of θ_1	0.65	0.00
3	Pure item of θ_1	0.98	0.00
4	Pure item of θ_1	0.73	0.00
5	Pure item of θ_1	1.05	0.00
6	Pure item of θ_1	0.96	0.00
7	Pure item of θ_1	0.54	0.00
8	Pure item of θ_1	1.00	0.00
9	Pure item of θ_2	1.09	0.00
10	Pure item of θ_2	0.45	0.00
11	Pure item of θ_2	0.79	0.00
12	Pure item of θ_2	0.97	0.00
13	Pure item of θ_2	0.00	0.86
14	Pure item of θ_2	0.00	1.01
15	Pure item of θ_2	0.00	1.12

Hypothetical two-dimensional 15 item ACT-E data set

Dimension dominance B

ITEM NUMBER	ITEM DESCRIPTION	a_1	a_2
1	Pure item of θ_1	0.85	0.00
2	Pure item of θ_1	0.65	0.00
3	Pure item of θ_1	0.98	0.00
4	Pure item of θ_1	0.73	0.00
5	Pure item of θ_1	1.05	0.00
6	Pure item of θ_1	0.96	0.00
7	Pure item of θ_1	0.54	0.00
8	Pure item of θ_1	1.12	0.00
9	Pure item of θ_2	0.00	0.97
10	Pure item of θ_2	0.00	1.00
11	Pure item of θ_2	0.00	0.39
12	Pure item of θ_2	0.00	0.74
13	Pure item of θ_2	0.00	0.86
14	Pure item of θ_2	0.00	1.01
15	Pure item of θ_2	0.00	1.12

Appendix D

Descriptive statistics for parameters generated according to three test structures: "45-item, 1000 examinee" data sets

Mean, standard deviations, skewness and kurtosis of difficulty as well as discrimination parameters for 100 "45-item, 1000 examinee" data sets.

Test structure	ACT-E		SAT-V		ASVAB-AR	
	Mean	SD	Mean	SD	Mean	SD
Mean a_1	.702	.049	1.046	.080	1.427	.110
Var. a_1	.067	.018	.167	.049	.275	.086
Skew. a_1	-.025	.380	.006	.372	.035	.396
Kurt. a_1	-.473	.563	-.464	.541	-.438	.588
Mean a_2	.719	.047	1.073	.092	1.465	.121
Var. a_2	.064	.017	.157	.046	.256	.066
Skew. a_2	-.050	.495	.039	.433	.020	.444
Kurt. a_2	-.216	.958	-.327	.700	-.321	.630
Mean d	-.011	.150	-.002	.115	-.009	.111
Var. d	.911	.185	.568	.125	.554	.116
Skew. d	.008	.391	.037	.371	-.009	.298
Kurt. d	.136	.731	.067	.622	-.070	.609