DECISION CRITERIA FOR DETERMINING UNIDIMENSIONALITY

by

Ⓒ John Allan Hattie

A thesis submitted in conformity with the requirements
for the Degree of Doctor of Philosophy in the
University of Toronto

UNIVERSITY OF TORONTO

SCHOOL OF GRADUATE STUDIES

PROGRAM OF THE FINAL ORAL EXAMINATION

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

OF

JOHN ALLAN HATTIE

2:00 p.m., Tuesday, April 21, 1981

Room 111, 63 St. George Street

DECISION CRITERIA FOR DETERMINING UNIDIMENSIONALITY

Committee in Charge:

Professor R. Stren, Chairman
Professor R. Hambleton, External Examiner
Professor C. MacLeod
Professor R. McDonald, Supervisor
Professor S. Nishisato
Professor R. Traub, Internal Appraiser
Professor M. Wahlstrom

John Allan Hattie

Biography

1950	Born, New Zealand
1971	B.A., University of Otago
1974	M.A., University of Otago
1974-Present	Doctoral Studies, University of Toronto

Graduate Studies

Major: Measurement & Evaluation

Intermediate Statistics & Research Design	Professor D. Burrill
Elements of Factor Analysis & Related Techniques	Professor R. McDonald
Elements of Scaling	Professor S. Nishisato
Advanced Test Theory	Dr. R. Traub
Personality Measurement	Dr. A. Even
Advanced Seminar in Measurement & Experimental Design	Professor S. Nishisato

First Minor: Applied Statistics

Design of Experiments	Professor J. Ogilvie
Applied Regression Analysis	Professor A. Paull

Second Minor: Computer Applications

Credit granted on the basis of the M.A.

Decision criteria for determining unidimensionality

John Allan Hattie

One of the fundamental assumptions of measurement theory is that a set of items forming a test is unidimensional. The purposes of this dissertation were (1) to review various methods for determining unidimensionality and to assess the rationale of these methods; (2) to attempt to clarify the term unidimensionality, and to show how it differs from other terms often used interchangeably; and (3) to assess the effectiveness of various indices proposed to determine unidimensionality.

Indices based on answer patterns, reliability, component and factor analysis, and latent traits were reviewed and it was shown that many of these lacked a rationale, that for many the sampling distributions were not known, and that many were adjustments to an established index to take into account some criticisms of it. Altogether 87 indices were reviewed.

It was demonstrated that unidimensionality often is used interchangeably with reliability, internal consistency, and homogeneity. Reliability was defined as the ratio of true score variance to observed score variance. Internal consistency has been used often as a synonym for unidimensionality, and it also denotes a group of methods that are intended to estimate reliability. Internal consistency methods are based on the variances and covariances of test-items, and depend on only one administration of a test.

Homogeneity seems to refer more specifically to the similarity of the item correlations, but the term is often used as a synonym for unidimensionality. Unidimensionality was defined as the existence of one latent trait underlying the data. The usefulness of the terms internal consistency and homogeneity was questioned.

A Monte Carlo simulation was conducted to assess the 87 indices under known conditions. A three-parameter, multivariate, logistic latent-trait model was used to generate item responses. Difficulty, guessing, discrimination, and the number of factors underlying the data were varied.

Many of the indices were highly intercorrelated, some resulted in estimates outside their theoretical bounds, and most were particularly sensitive to the intercorrelations between the factors. Indices based on answer patterns, reliability, component analysis, linear factor analysis, and on the one-parameter latent trait model were ineffective. The sums of absolute residuals from a nonlinear factor analysis (specifying one factor with cubic terms) and from two-parameter latent trait models (Christoffersson, 1975; McDonald, 1980; Muthen, 1978) were able to discriminate between cases with one latent trait and cases with more than one latent trait.

Acknowledgements

It is obvious from the many references in this thesis that R.P. McDonald, the chairman of my thesis committee, has been an influential contributor and an original pioneer to the topic of unidimensionality. More than that, he has encouraged, criticised, inspired, and patiently guided me through this thesis. I thank you for contributing so much as a teacher, patron, and friend.

I am indebted to the other members of my committee, Ross Traub, Shizuhiko Nishisato, and Merlin Wahlstrom for their assistance, encouragement, and contributions to this dissertation.

The environment of MECA, OISE is an excellent place to complete a doctorate. In particular, thank you Colin Fraser, Jamshid Etezadi, Richard Wolfe and Don Burrill. Yet there are so many others: the Burnett's, Hundleby's, Thomasson's, Parish's, and Fitzgerald's. Thank you MECA for Monday to Thursdays, the GROUP for Fridays, and those others for the weekends. You have one thing in common yet are beautifully multidimensional.

Contents

List of Tables

List of Figures

Chapter I

Introduction

That a set of items forming an instrument all measure just one thing in common is a most critical and basic assumption of measurement theory. Besides being the basis of most mathematical measurement models there are many reasons why unidimensionality is so crucial. We must agree with Lumsden (1976) that the beginning of measurement is the conception of the measureable attribute. How can we make any claims to measure if our measuring instrument has a number of different types of items based presumably on different attribute conceptions? Indeed the assumptions which lead to a unit of measurement implicitly require the use of a unidimensional set of items. To make psychological sense when relating variables, ordering persons on some attribute, forming groups on the basis of some variable, or making comments about individual differences, it is required that the variable be unidimensional -- that is, that the various items measure the same ability, achievement, attitude, etc. McNemar (1946) put the various arguments most clearly

> Measurement implies that one characteristic at a time is being quantified. The scores on an attitude scale are most meaningful when it is known that only one continuum is involved. Only then can it be claimed that two individuals with the same score or rank can be quantitatively and, within limits, qualitatively similar in their attitude towards a given issue. As an example suppose a test of liberalism consists of two general sorts of items, one concerned with economic and the other with religious issues. Two individuals could thus arrive at the same numerical score by quite different routes. Now it may be true that economic and religious liberalism are correlated but unless highly correlated the meaning of scores based on such a composite is questionable. (p.268)

Despite this importance, there is not an accepted and effective index of the unidimensionality of a set of items. Lord (1980) contends that there is a great need for such an index and Hambleton, Swaminathan, Cook, Eignor, and Gifford (1978) argued that "testing the assumption of unidimensionality takes precedence over other goodness of fit tests of a latent trait model sinoe, if the assumption of unidimensionality is untenable, the results of the other tests are more difficult to interpret" (p.487).

Yet it may be unrealistic to search for indices of unidimensionality or sets of unidimensional items. Some have argued that we should ignore unidimensionality and look for other desirable qualities of a set of items (Lord, 1974), such as whether the estimates are consistent, or whether the estimates provide a useful and effective summary of the data. Others have claimed that a set of items will not be unidimensional except for the most simple variables, yet a unidimensional test can be factorially complex. It seems meaningful to quest after an index if we rephrase the question and ask not "Is a test unidimensional or not?", but rather "Are there decision criteria that determine how close a set of items is to being a unidimensional set?" In this thesis various indices are reviewed and a simulation is described to evaluate whether the indices are adequate decision criteria for determining unidimensionality.

It may be argued that the first task should be to define what is meant by unidimensionality. In a sense this was done in the opening sentence: a set of items is unidimensional if the items measure just one thing in common. The problem, however, is to know whether a set of items measure just one thing in common. A more precise explication of the concept is needed. The aim is not to make the concept of unidimensionality clear and precise, as this presumes that initially someone set down a clear and precise meaning and that a test has unidimensionality as Joan has chickenpox and a fern has an aroma. Rather we will trace, in the spirit of Wittgenstein (1958), how the word is, and has been, used.

It seems that unidimensionality was initially identified as a desirable property of tests in the 1940's and 1950's. (Lumsden, 1976, found the earliest mention was by Vernon, 1938.) It was used in the same way as homogeneity and internal consistency and it was not until the recent increase of interest in latent trait models that the term has been defined more clearly and precisely. Still it is possible to find some authors interchanging the terms reliability, internal consistency, homogeneity, and unidimensionality, while other authors have more specific uses for them. Appendix A demonstrates how these terms have often been used interchangeably. Part of the aim of Chapter II is to detail the various uses to which these terms have been put, and make the boundaries of the terms less fuzzy. Definitions are proposed for these terms at the end of Chapter II that seem consistent with how many authors have used the terms (or at least, what they appear to have intended when they used them), and that set clearer boundaries between the terms.

One criticism of a search for decision criteria to determine unidimensionality is that an act of judgement and not an index is required. Kelley (1942) argued that embodied in such concepts is a belief or point of view of the investigator and an act of judgement is demanded when a researcher asserts that items measure the same function. Certainly, it may be possible to recognize by inspection whether one test appears more unidimensional than another. Also even if we have an index we must use judgement when interpreting it, particularly as the sampling distribution for most indices is not known. An index must therefore be seen as only part, but probably a very important part, of the evidence used to determine whether a test is unidimensional.

In all, over 87 indices have been identified that someone at some time has suggested could be a good index of unidimensionality. Most proposers do not offer a rationale for their choice of index, even fewer assess its performance relative to other indices, and hardly anyone has tested out the indices using data of known dimensionality. The aim of this thesis is to review the literature and assess whether there is a rationale for each index. A simulation is conducted to study the behaviour of the indices under known conditions. Chapter II presents the review, Chapter III the design of the Monte Carlo simulation, Chapter IV the results and discussion of the simulation, and Chapter V an integration of the previous chapters.

Chapter II

Indices to determine unidimensionality

The first four sections of this chapter detail various indices that have been proposed to index unidimensionality. The sections relate to methods based on answer patterns, reliability, factor analysis, and latent trait theory. The final section reviews specific attempts to study unidimensionality. Given that there are many indices it is not possible to discuss each fully or to cite the researcher(s) who has (have) used each index. Table 1 lists the indices in the order that they are discussed in this Chapter, gives references to the writers who have used or recommended each index, and states whether a rationale is presented (a subjective judgement, recorded either yes, part, low, or no). (All 87 indices are listed in Appendix B.)

Indices based on answer patterns

Guttman and Loevinger have been primarily responsible for developing indices of unidimensionality based on answer patterns. Both researchers have used the amount by which a set of item responses deviates from the ideal scale pattern as the basis for an index. The two methods differ only in the procedure for counting the deviations.

Guttman (1944, 1950) developed scalogram analysis and the reproducibility coefficient to provide "a simple method for testing a series of qualitative items for unidimensionality" (p. 46). Guttman's work is based on Walker's (1931, 1936, 1940) notions of the "ideal answer pattern". Walker had developed a coefficient of the "goodness

Table 1

Indices, rationale, and researchers who have recommended each index

Index	Rationale	Researchers who have recommended the index
Green RepB	Part	Green (1956)
		White & Saltz (1957)
Consistency	Yes	Green (1956)
		White & Saltz (1957)
Loevinger H_t	Yes	Loevinger (1944)
		Gage & Damrin (1950)
		Lumsden (1959, 1961)
		Hoffman (1975)
Alpha	No	Gage & Damrin (1950)
		Cronbach (1951)
		Lumsden (1959)
		Freeman (1962)
		Guilford (1965)
		Horrocks & Schoonover (1968)
		Payne (1968)
		Ryan (1979)
Horst's alpha	No	Horst (1953)
		Lumsden (1959)
Raju's alpha	Yes	Terwilleger & Lele (1979)
		Raju (1980)
Mean alpha	Yes	Cronbach (1951)
		Humphreys (1956)
		Cattell & Tsujioka (1964)
		Cattell (1978)
Mean correlation	Low	Cronbach (1951)
		Kaiser (1968)
		Silverstein (1980)
No. zero correlations	Low	Mosier (1936)
		Humphreys (1952)
		Armor (1974)
Item-test correlation	Low	Mosier (1940)
		Kelley (1942)
		Nunnally (1970)
KR-21	No	Kuder & Richardson (1936)
		Gage & Damrin (1950)
		Raju (1980)
Percent variance	Part	Cattell & Tsujioka (1964)
		Hambleton & Traub (1973)
		Carmines & Zeller (1979)
		Hutten (1979)
		Reckase (1979)
No. eigenvalues > 1	Low	Laforge (1965)
		Armor (1974)
		Koch & Reckase (1979)

Eigenvalue 1/ Eigenvalue 2	Low	Lumsden (1957, 1961) Bentler (1972) Hambleton (1980) Hutten (1980) Lord (1980) Smith (1980)
(Eigen 1 - Eigen 2)/ (Eigen 2 - Eigen 3)	No	Divgi (1980)
Eigen/Variance	Low	Kelley (1942)
Sum residuals	Part	Lumsden (1959) McDonald (1981)
No. residuals > .01	Part	Lumsden (1959)
Correlation raw & factor scores	No.	Dubois (1970)
Chi-square (1 factor)	Part	Bock & Lieberman (1970)
Chi-square (2 factor - 1 factor)	Part	Joreskog (1978)
Tucker & Lewis	Low	Tucker & Lewis (1973)
Theta	Low	Bentler (1972) Armor (1974) Carmines & Zeller (1979) Greene & Carmines (1980)
Omega	Low	Heise & Bornstedt (1970) Armor (1974) Smith (1974a, 1974b) Carmines & Zeller (1979) Greene & Carmines (1980)
Green et al.	Part	Green, Lissitz & Mulaik (1977) Hattie & Hansford (1979) Watkins & Hattie (1980)
Nonlinear factor analysis	Yes	McDonald (1967b, 1980a, 1981) McDonald & Ahlawat (1974) Lam (1980) Etezadi (1981)
Christoffersson-Muthen	Yes	Christoffersson (1975) Muthen (1978, 1981) Lord (1980) Muthen & Christoffersson (1981)
Two-parameter latent trait model	Yes	McDonald (1981) McDonald & Fraser (in prep.)
One-parameter latent trait model	Part	Wright (1977) Reckase (1979) Rentz & Rentz (1979) Wright, Mead & Bell (1979) Wright & Stone (1980)
Lord's chi-square	Part	Lord (1953) Hambleton et al. (1978)

of fit" (which he called a "hig") of a series of item responses to a perfect answer pattern. A perfect answer pattern occurs when a total test score equal to n, is composed of correct answers to the n easiest questions, and therefore of correct answers to no other questions.

Guttman (1971) defined a scale as a "one-dimensional structure", yet he was aware that perfect scales are not to be expected in practice and he developed a coefficient of reproducibility which was an index of the amount by which a scale deviates from the perfect scale. Any test that had a reproducibility coefficient of at least .90 could be considered, according to Guttman (1945), an acceptable approximation to a perfect scale. In a 1944 article Guttman recommended .85 as the cutoff point. Neither of these values takes into account the effect of the number of categories for each item, or the difficulty of the items.

The index of reproducibility is a function of the number of errors. These are defined as unity entries that are below a cutoff point and zero entries that are above it. Various authors have suggested rules to determine this cutoff point. Guttman (1950) determined his coefficient of reproducibility by counting the number of responses that could have been predicted wrongly for each person on the basis of his/her total score, dividing these errors by the total number of responses and subtracting the resulting fraction from 1. White and Saltz (1957) and Nishisato (1975) detail the method and provide examples. Guttman was aware that his coefficient was affected by the range of difficulties of the items. The coefficient has a lower bound that is a function of the item difficulties and the coefficient can be well above .5 when the item difficulties depart from .5.

Jackson (1949) proposed a Plus Percentage Ratio (PPR) coefficient that was free from the effect of difficulty values. His method is very cumbersome and time consuming, and has many of the problems of the coefficient of reproducibility (e.g., the problem of determining the cutoff). To overcome these deficiencies, Green (1956) proposed approximation formulae for dichotomous items. Green reported an average discrepancy between his and Guttman's formula of .002. (See Nishisato, 1975, pp. 19-23 for details.)

Green presented two coefficients, RepA and RepB , but noted that the bounds of RepA and RepB are usually not 0 and 1. Green proposed an index that is zero when RepA (or RepB) is equal to the reproducibility coefficient expected by chance (RepI), and is unity when RepA is unity. This criterion he called an index of consistency (CONSIS)

$$\text{CONSIS} = (\text{Rep} - \text{RepI})/(1 - \text{RepI}), \qquad (1)$$

where Rep is the obtained reproducibility of the test using RepA or RepB , RepI is the reproducibility that would be obtained with the same set of item difficulties and complete independence between items, and 1 is perfect reproducibility. (Incidentally, Green's method of identifying errors is the same as Loevinger's but Green sums over those item pairs whose members are adjacent in difficulty level and not over all pairs of items as does Loevinger.)

Loevinger (1944, 1947, 1948) claimed that underlying tests of ability are two assumptions. First, scores at different points of the scale reflect different levels of the same ability, and second, for any two items in the same test, the possession of abilities required to complete one item may help or may not help but they will not hinder or make less likely adequate performance on the other items. These assumptions led Loevinger to define a unidimensional test as one such that, if A's score is greater than B's score, then A has more of some ability than B, and it is the same ability for all individuals A and B who may be selected.

Loevinger developed what she termed an index of homogeneity. When all the test items are arranged in order of increasing difficulty the proportion of subjects who have passed both items, P_{ij} is calculated for all pairs of items. From this the theoretical proportion, $(P_i P_j)$, who would have passed both items had they been independent is subtracted. These differences are summed over the n(n-1) pairs of items

$$S = \Sigma\Sigma (P_{ij} - P_i P_j) \qquad (2)$$

A test made up of completely independent items would have a value for S of 0, but S does not have an upper limit of unity when the test is perfectly homogeneous. The upper limit is fixed by the proportion of examinees passing the more difficult item in each pair (P_j)

$$S_{max} = \Sigma\Sigma(P_j - P_iP_j) \qquad (3)$$

The index of homogeneity thus proposed by Loevinger was the ratio of the quantities (2) and (3) ($H_t = S/S_{max}$). The coefficient is one for a perfectly homogeneous test and zero for a perfectly non-homogeneous test. Yet for some sets of items, such as those that allow guessing, the lower bound may not necessarily be zero.

Comments

There have been many criticisms of indices of unidimensionality based on answer patterns. The most serious objection is that these methods can only achieve their upper bounds if the strong assumption of scalability is made. Another major criticism is that there is nothing in the methods that enables one to distinguish a test of just one trait from a test of a constantly weighted composite of abilities. Loevinger, who recognised this objection, suggested that in such cases methods of factor analysis were available that could help in determining whether there were many abilities measured or only one. On the other hand, Guilford (1965) argued that it is possible to say that a test that measures, say, two abilities is a homogeneous test provided that each and every item measures both abilities. An example is an arithmetic-reasoning test or a figure-analogies test. Following from this objection it is possible that answer-based methods can be of little use when there is a conjunctive model, in the Coombs (1964) sense.

Coombs (1964) attempted to extend the Guttman scalogram technique to the multidimensional case. He postulated a unidimensional continuum, along which each item and each subject could be represented. One of Coombs' models is the conjunctive model which assumes that an excess of ability in one dimension does not compensate for the lack in another. Thus, successful performance on a task requires a certain minimum on each of the several relevant dimensions. An item failed by an individual must be higher on at least one dimension than an item passed. To pass such an item an individual must exceed it on all relevant dimensions.. An example cited by Coombs (1964, p. 246) is that of an individual taking a history test in French. He has to know enough French to be able to understand the questions, but no matter how much more French he knows, it will not help answer the questions; and he has to know enough history to answer the questions, but no matter how much history he knows, it will not compensate for not knowing enough French to understand the questions.

A further objection is that it is possible to construct a set of items (not based on the conjunctive model) that form a perfect scale yet would appear not to be unidimensional. For example, consider the following set of items:

1. Jump over the 2 centimeter high jump
2. Spell cat
3. Multiply 9 x 16
4. What is Magna Carta?

These items conceivably could form a perfect scale and yield a reproducibility coefficient and a Loevinger H_t of 1, yet many authors (e.g., Lumsden, 1959) have claimed that such an example cannot be considered unidimensional. A similar example is a set of 10 items testing different abilities, one item at a level of difficulty appropriate to each grade from one through ten. The test is given to a group of 10 children, each of which is an average student at each grade level from one to ten. A perfect scale very probably could result. It might seem that using these examples, as have Loevinger (1947), Humphreys (1949), and Lumsden (1959) amongst others, is confusing the method of assessing dimensionality with the identification of the dimension(s) measured. It is possible to say that both sets measure one characteristic, which we could label development. Saying a test is unidimensional does not identify that dimension, in the same way that saying a test is reliable does not determine what it is reliably measuring.

Indices based on reliability.

Loevinger (1948) claimed that reproducibility coefficients and indices of homogeneity were similar to split-half reliabilities. The similarities have led researchers such as White and Saltz (1957) to posit that the difference between certain reliability coefficients and homogeneity or reproducibility is "more apparent than real" (p. 96). Gage and Damrin (1950) empirically compared Loevinger's homogeneity index and various reliability coefficients including alpha. They found that Loevinger's H_t was much lower than the reliability coefficients (average H_t =.28, average alpha = .91), and found it difficult to interpret H_t because there was no standard error formula for H_t that would give a basis for interpreting H_t. Gage and Damrin did note that, unlike alpha, H_t did not increase as the number of items increased.

Alpha

Cronbach (1951) systematically presented a general formula that was previously well known, and he discussed its interpretations. He believed that the previous name for this general formula, Kuder-Richardson formula 20 (KR-20), was an "awkward handle for a tool that we expect to become increasingly prominent in the test literature" (p. 299). Instead Cronbach renamed it coefficient alpha (α). Unfortunately the introduction of yet another name has led to more confusion as many authors still seem unaware that given dichotomously scored items, alpha is equivalent to KR-20 (Kuder & Richardson, 1937), Guttman's (1945) lambda-3, Hoyt's (1941) index of reliability, and Jackson and Ferguson's (1941) formula of "rational equivalence".

Cronbach proved: 1) that alpha is the mean of all possible split-half coefficients; 2) that alpha is the value expected when two random samples of items from a pool like those in the given test are correlated; and 3) that alpha is a lower bound to the proportion of test variance attributable to common factors among the items. Cronbach claimed that alpha is an upper bound to the proportion of the test variance due to the first factor among the items, and in tests where all items measure the same thing it was essential that a large proportion of the test variance be attributable to the principal factor running through the test.

Novick and Lewis (1967) established that alpha is less than or equal to the reliability of the test with equality if and only if the items are essentially tau-equivalent. This is the same as saying that the off-diagonal elements of the variance-covariance matrix of observed scores are all equal. This implies the weaker property of unit rank. (But we do not have the converse, that is, unit rank does not imply that the off-diagonals are all equal.) Thus, there exists a diagonal matrix that, on subtracting from the variance-covariance matrix reduces it to unit rank. Unfortunately, there is no systematic relationship between the rank of a set of variables and how far alpha is below the true reliability. As a consequence, it is not reasonable to claim that the higher alpha becomes the more likely it is that a set of items has unit rank.

Cronbach was aware of criticisms made by Loevinger (1948) amongst others, that alpha was a poor index of unidimensionality because, as it is based on product-moment correlations, alpha cannot attain unity unless the items are all of equal difficulty. This criticism has led other authors to modify alpha in various ways, and obtain other indices of unidimensionality. Generally these modifications are variations on dividing the index by its maximum value (index/index-max).

Index/index-max formulae

Loevinger (1954) argued that her H_t was more desirable than alpha because it took varying item difficulties into account and H_t could be 1 for a perfectly homogeneous test. Her index could attain 1 because H_t is of the index/index-max type. In fact, Loevinger defined her coefficient as a weighted average of probabilities for each item, adjusted so that the coefficients would equal zero for a perfectly heterogeneous test and unity for a perfectly homogeneous test.

Horst (1953) argued that instead of using Loevinger's method of estimating average item intercorrelation corrected for dispersion of item difficulties, it is more "realistic" to estimate average item reliability corrected for dispersion of item difficulties. The problem in such an index is to obtain plausible estimates of item reliability. Horst used KR-18 as an estimate of item reliability, and derived an index that he claimed was a more realistic estimate of the reliability coefficient than alpha. It yields a higher estimate if the items are of unequal difficulty and will have unity as an upper limit irrespective of the dispersion of item difficulties.

Both Loevinger and Horst define the maximum test variance given that the item difficulties remain the same, or alternatively that the item covariances be maximum when the item difficulties are fixed. Terwilliger and Lele (1979) and Raju (1980) instead maximize the item covariances when test variances are fixed but the item difficulties are allowed to vary.

Although his work predated these criticisms, Cronbach was aware of the problems of alpha, and predicted many of the adjustments that Loevinger, Horst, and Raju were to make. He claimed that many of the possible modifications to alpha had no practical effect, that some of the adjusted indices were also affected by item dispersion, and that it could even be a virtue that alpha does not reach unity for items of unequal difficulty. To demonstrate the lack of practical effect Cronbach first held constant the relation between the "underlying factors" by fixing the tetrachoric correlations at .30. He allowed p_i to vary between .10 and .90 and held p_j constant at .50. Under these conditions, phi ranged from .14 to .19. Cronbach concluded that as tetrachorics were usually less than .30 and thence the effect of differing difficulties on phi was even less, then phi is very nearly constant over a wide range of item difficulties. In another example Cronbach constructed four hypothetical tests with a normal, peaked or rectangular distribution of difficulties, and phi's of various ranges (see his Table 8). The effects on the average phi and alpha were negligible and Cronbach concluded that neither phi nor alpha was affected in any practically important way.

Cronbach demonstrated that Loevinger's H_t was markedly affected by variation in item difficulty. In the first example cited above where phi's ranged from .14 to .19, Loevinger's H_t for the same items ranged from .19 to .42. (See Carroll, 1950, for another example of the marked effects of variation in item difficulty on Loevinger's coefficient.)

In another example given by Cronbach, he let five items have perfect tetrachoric intercorrelations with p_i's of .40, .45, .50, .55, and .60. Such a test is perfectly homogeneous or reproducible in the Loevinger and Guttman sense. He then described a 10-item test composed of the previous five items plus five others whose p_i's are .25, .30, .35, .65 and .70. While both tests, the 5- and the 10- item tests, have H_t's of 1, the second makes a greater number of discriminations. Even though the second test has a lower average phi, there is less redundancy. Cronbach claimed that the phi coefficient reported whether a second item gave new information that the first did not. Certainly redundancy can be desirable when the accuracy of a single item is low. Yet for Cronbach, the point was that the phi coefficient that tells when items do and do not duplicate each other was a better index just because it does not reach unity for items of unequal difficulty.

Cronbach summarized his defence of alpha based on his analysis of phi's by commenting that efforts to correct indices as a means of controlling for the dispersion of item difficulties are mistaken in that these indices lose information, and the difference between alpha and these corrected indices would be negligible. To see the similarities and also to summarize the indices, Table 2 presents Loevinger's H_t, alpha, Horst and Raju's modifications, and also an index used by DuBois, Loevinger and Gleser (1952) to find the most homogeneous subset of items and thus, they claim, maximize the homogeneity of the total test (see also Gupta and Burnett, 1972). This latter index, called a saturation coefficient, was developed as a result of Cronbach's criticism of Loevinger's work and is merely a function of alpha. It is very clear from Table 2 that there are many similarities between the measures. Also note that alpha is dependent on the length of the test and this leads to a new problem — and more indices.

Table 2

Loevinger's H_t, alpha, Horst's alpha, Raju's alpha,

and the saturation index of Dubois et al.

Loevinger's H_t =
$$\frac{\sigma^2_t - \Sigma\sigma^2_i}{\sigma^2_{hom} - \Sigma\sigma^2_i}$$

Alpha =
$$\frac{n}{n - 1} \quad \frac{\sigma^2_t - \Sigma\sigma^2_i}{\sigma^2_t}$$

Horst's alpha =
$$\frac{\sigma^2_t - \Sigma\sigma^2_i}{\sigma^2_t} \quad \frac{\sigma^2_{hom}}{\sigma^2_t}$$

Raju's alpha =
$$\frac{\sigma^2_t - \Sigma\sigma^2_i}{\sigma^2_t - \sigma^2_m}$$

DuBois et al.'s saturation =
$$\frac{\sigma^2_t - \Sigma\sigma^2_i}{\sigma^2_t}$$

Note. n = number of items

σ^2_t = total test variance

σ^2_i = item variance

σ^2_m = maximum variance assuming the items form a perfect Guttman scale

σ^2_{hom} = maximum variance of a test with the same distribution of item difficulties as the test under consideration.

Indices based on correcting for the number of items

Conceptually the homogeneity of a test should be independent of its length. To use Cronbach's (1951) analogy: A gallon of homogenized milk is no more homogeneous than a quart. Yet alpha increases as the test is lengthened and so Cronbach proposed that an indication of inter-item consistency could be obtained by applying the Spearman-Brown formula to the alpha for the total test, thereby estimating the mean correlation between items. The derivation of mean-alpha is based on the case in which the items have equal variances and equal covariances, and then applied more generally. Cronbach recommended the formula as an overall index of internal consistency. If mean-alpha is high, alpha is high; but alpha may be high even when items have small intercorrelations -- it depends on the spread of the item intercorrelations and on the number of items. Cronbach pointed out that a low mean-alpha could be indicative of a non-homogeneous test and recommended that when mean-alpha is low, only a study of correlations among items would indicate whether a test could be broken into more homogeneous subtests.

Mean inter-item correlation

Cattell (1978), Cattell and Tsujioka (1964), and Humphreys (1956) have called the mean inter-item coefficient an index of homogeneity. Kaiser (1968) argued that a better estimate of average intercorrelation is based on the largest eigenvalue of the matrix of correlations between items and derived a measure he called gamma.

Nature of the inter-item correlations

Cronbach was aware that the magnitude of the inter-item correlations should be high and/or there should be a large number of items, and he recommended close inspection of the correlations. Armor (1974) argued that inspecting the inter-item correlations for patterns of low or negative correlations was usually not done, and was probably the most important step of all as it contains all information needed to decide upon the dimensionality. Armor claimed that by assessing the number of intercorrelations close to zero it was possible to avoid a major pitfall in establishing unidimensionality. That is, it becomes possible to assess whether more than one independent dimension is measured.

Mosier (1936) also placed an emphasis on the number of inter-item correlations close to zero. In an analysis of 39 items from the Thurstone Personality Schedule (Thurstone and Thurstone, 1930) Mosier found 390 of the 741 inter-item correlations were not significantly different from zero, and 77 were negative. Mosier concluded that

> with so many zero and negative coefficients among the intercorrelations it is apparent that these items are not measuring the same trait. It would seem that the nature of the 'single trait' will continue to be obscured so long as we continue to rely solely upon the criterion of internal consistency in any of its variant forms, without investigating also the item intercorrelations upon which Richardson has shown it is based. (pp. 279-280)

Item-test indices

Mosier (1940) also pointed out that if the distributions of the correlations are skewed then it is possible for a test to have a high average inter-item correlation and yet have a modal inter-item correlation of zero. Using the same argument based on the skewness of the correlations, Mosier was critical of Richardson's claim that "the item-test coefficient gives an indication of the extent to which the item measures what the test as a whole measures". While this is correct if the test is measuring a single trait, if the test is measuring a composite of two or more traits there are obvious difficulties when using item-test correlations. (Lumsden, 1961 has also made this point.) Loevinger (1947) argued that the point-biserial coefficient (correlation of item with total test score) is not desirable as it does not have a maximum value of one and it varies with the distribution of item difficulties.

Kuder-Richardson formula 21

Because KR-21 is easy to calculate it has been used as an index of homogeneity. KR-21 differs from KR-20 in that for the derivation of the former all the items are assumed to be equal in difficulty. KR-21 and KR-20 are related by

$$\text{KR-20} = \text{KR-21} + (n^2/(n-1))(\sigma^2_p/\sigma^2_t) \qquad (4)$$

where σ^2_p is the variance of the item difficulties, and σ^2_t is the total variance. Hence, KR-20 equals KR-21 if and only if $\sigma^2_p = 0$. KR-20 is a minimum when all items are equally difficult and this minimum value is given by KR-21.

Sampling variability and interpretation

Unfortunately for many of the indices suggested the sampling variation is not known. Gage and Damrin (1950) despaired that there was no kind of 'norm', as they called it, that gave a basis for interpreting values of Loevinger's H_t. But this lack applies also to many of the reliability-based indices.

There have been attempts to determine the sampling error of alpha and related formulae (Aoyama, 1957; Baker, 1962; Cleary & Linn, 1968; Feldt, 1965; van der Kamp, 1976; Kristof, 1969). Cleary and Linn were careful to point out that the sampling errors should not be used for confidence limits as the distribution of the sample reliability coefficient is skewed and it is a biased estimate of the population value.

In addition to the difficulties of computing the sampling errors there are problems of interpretation. That is, depending on the purpose of administering a test the reliability coefficient may be interpreted in various ways. Take, for example, McNemar's (1946) criticism of Conrad and Sanford's (1944) study that reported average intercorrelations among three sets of attitude items of .08, .12, and .07. McNemar concluded that because these were so low the meaning of total scores on each of these aspects "is enigmatic" (McNemar, 1946, p. 363). Yet depending on the number of items in the test this 'low' inter-item correlation could become an alpha (for 50 items) of .81, .87 and .79, respectively -- which may seem less enigmatic!

Hence, it may be misleading merely to interpret the magnitude of reliability coefficients and further the lack of confidence intervals limits comparisons between the coefficients.

Green, Lissitz and Mulaik

Green, Lissitz and Mulaik (1977) observed that while high "internal consistency" as indexed by a high alpha results when a general factor runs through the items, this does not rule out obtaining high alpha when there is no general factor running through the test items. As an example, they used a 10-item test that occupied a five-dimensional common factor space. Furthermore, they used orthogonal factors and had each item loading equally (.45) on two factors in such a way that no two items loaded on the same pair of common factors. This factor pattern matrix satisfies Thurstone's (1935) requirements for simple structure since no item requires all common factors to account for its common factor variance. The factors are also well determined with four items having high loadings on each factor. Each item has a communality of .90. They calculated alpha to be .81, and pointed out that "commonly accepted criteria" would lead to the conclusion that theirs was a unidimensional test. But this example is far from unidimensional. On another criterion, already listed above (p. 21), it can be determined that 15 of the 45 distinct inter-item correlations are zero. This should be a cause for concern.

Green et al. assume that they have a non-unidimensional case in their example because they have five orthogonal factors. These five factors however, can be rotated in many ways to get other than a simple structure. For example, using a principal axis solution 40% of the total variance can be accounted for by the first component. Using Green et al.'s notion of "commonly accepted criteria" this might be taken to indicate a unidimensional solution — a conclusion not inconsistent with the interpretation of an alpha of .811.

In a Monte Carlo study, Green et al. found: 1) that alpha increases as n increases; 2) that alpha increases rapidly as the number of parallel repetitions of each type of item increases; 3) that alpha increases as the number of factors pertaining to each item increases; 4) that alpha readily approaches and exceeds .80 when the number of factors pertaining to each item is two or greater and n is moderately large (approximately equal to 45); and 5) that alpha decreases moderately as the item communalities decrease. They then concluded that the chief defect of alpha as an index of dimensionality is its tendency to increase as the number of items increase.

They were careful to note that Cronbach realized the effect of the number of items and that Cronbach recommended the average inter-item correlation. So Green et al. included this index in their simulation and found that the average inter-item correlation is unduly influenced by the communalities of the items and by negative inter-item correlations. Looking closely at their average inter-item correlation results, we observe an interaction between the number of factors and the size of the communalities. As the number of factors goes up, the inter-item correlation goes down, but comparisons within the same number of factors indicate that as the communalities go up so too does the average inter-item correlation.

Clearly, the use of alpha as an index of unidimensionality is extremely suspect.

Indices based on factor analysis

Cronbach (1951) claimed that for a test to be interpretable, it is essential that all items be factorially similar. That is, a large proportion of the test variance should be attributable to the principal factor running through the test. He then stated that alpha estimates the proportion of the test variance due to all common factors among the items, and it indicates how much the test score depends upon the general and group, rather than on the item specific factors. Cronbach claimed that this was true provided that the inter-item correlation matrix was of unit rank, otherwise alpha is an underestimate of the common factor variance. Cronbach stated that this underestimation is not serious unless the test contains distinct clusters. These statements led Cronbach, and subsequently many other researchers, to make the further claim that a high alpha indicates a "high first factor saturation" (p. 330) or that alpha is an index of "common factor concentration" (p. 331) and the implication is that alpha is related to dimensionality. Cronbach's inferences have been questioned by McDonald (1970, 1978, 1981) who demonstrated that alpha is only a lower bound to the proportion of variance due to all the common factors.

Underlying the intention to use alpha is the belief that it is somehow related to the rank of a matrix of item intercorrelations. Lumsden (1957) claimed that a necessary condition for a unidimensional test (that is, "a test in which all items measure the same thing", p. 106) is that the matrix of inter-item correlations is of unit rank (see also Lumsden, 1959, p. 89). Obviously he means that the matrix fits the Spearman case of the common factor model. That a unidimensional test has been defined in terms of unit rank leads to certain problems, the most obvious of which is how to determine statistically when a sample matrix of inter-item correlations has unit rank. Some of the estimation issues relate to whether component or factor analysis should be used; how to determine the number of factors; the problem of communalities; the role of eigenvalues; the choice of correlations; and the occurrence of "difficulty" factors. These issues will now be discussed and various decision criteria enumerated.

<u>Indices based on principal components</u>

As the first principal component explains the maximum variance then this variance, usually expressed as percent of total variance, has been used as an index of unidimensionality. The implication is that the larger the amount of variance explained by the first component the closer is the set of items to being unidimensional. An obvious problem is how "high" does this variance need to be before we can conclude we have a unidimensional test. Carmines and Zeller (1979), without any rationale, contended that we should expect at least 40 percent of the total variance to be accounted for by the first component before we can say a set of items is measuring a single phenomenon. Reckase (1979) recommended that the first component should account for at least 20 percent of the variance. However, it is not difficult to invent examples in which a multidimensional set of items has higher variance on the first component than a unidimensional test. (See, for example, the results from the simulation described in the next Chapter.)

As many of the components probably have much error variance and/or are not interpretable, there have been many attempts to determine how many components should be "retained". Most common is to retain only those components with eigenvalues greater than one (see Kaiser, 1970, for a justification, although there are many critics of his argument, e.g., Gorsuch, 1974; Horn, 1965, 1969; Linn, 1968; Tucker, Koopman & Linn, 1969).

Lumsden (1957, 1961), without giving reasons, suggested that the ratio of the first and second eigenvalues would give a reasonable index of unidimensionality although he realised that besides having no maximum value, little is known about the extent to which such an index may be affected by errors of sampling or measurement. Lumsden (1957) calculated the ratio on four subsets of number-series items and concluded that the three with ratios of 20:1, 38:1, and 35:1 were unidimensional whereas the ratio of the fourth set of 15:1 was sufficient, he claimed, to make the hypothesis of unidimensionality untenable. Lord (1980) argued that a rough procedure for determining unidimensionality was the ratio of first to second eigenvalues and an inspection as to whether the second root is not much larger than any of the others. Lord and Novick (1968) reported a study by Indow and Samejima (1966) where a thirty-item test of non-verbal reasoning was administered to 883 students. The first nine eigenvalues were approximately 16.0, 1:3, 1.0, .9, .9, .8, .5, .3, and .2 (which represent the following percentages of variance: 53.33, 4.33, 3.33, 3.00, 3.00, 2.67, 1.67, 1.00, .67). As the first eigenvalue is so large, and as the second is almost as small as the later ones, Lord and Novick concluded that there seems good reason to treat their data as arising from a one-dimensional space. In another example (Lord, 1980), the first 12 eigenvalues from a 50 item word relation test (read approximately from a graph) were 10.2, 2.2, 1.3, 1.0, 1.0, .9, .9, .8, .8, .7, .7, and .7 (this represents the following percentages of variance: 20.4, 4.4, 2.6, 2.0, 2.0, 1.8, 1.8, 1.6, 1.6, 1.4, 1.4, and.1.4). Using his criterion, Lord concluded that the items were "reasonably uni-dimensional" (p. 21).

A possible index to operationalise Lord's criteria could be the difference between the first and second eigenvalues divided by the difference between the second and third eigenvalues. Divgi (1980) argued that this index seemed reasonable because if the first eigenvalue is relatively larger than the next two largest eigenvalues this index will be large, whereas if the second eigenvalue is not small relative to the third, then regardless of the variance explained by the first component and despite the number of eigenvalues greater than or equal to one, this index will be small. It is not difficult, however, to construct cases where this index must fail. For example given four common factors, if the second and third eigenvalues are nearly equal then the index could be high. But in a three factor case, if the difference between the second and third eigenvalues is large, then the index would be low. Consequently the index would mark the four factor case as unidimensional, but not the three factor case!

Another index similar to the first eigenvalue or percent of variance is due to Kelley (1942). Kelley introduced a coefficient of coherence that he claimed was a measure of the unity, coherence, or singleness of purpose of a test. Essentially the coefficient is a ratio of the sum of the item variances to the variance of the first centroid factor. Given that the centroid is an approximation to the first principal component, Kelley's index approximates the reciprocal of the usual first eigenvalue.

The sum of squared residuals, or sum of the absolute values of the residuals after removing one component has been used as an index of unidimensionality. Like many other indices there is no established criterion for how small the residuals should be before concluding that there is a unidimensional test. There have been suggestions that the absolute size of the largest residual or the average squared-residual are useful indices of the fit. Thurstone (1935), Kelley (1935), and Harman (1979) argued that all residuals should be less than $(N-1)^{-.5}$, where N is the sample size (i.e., the standard error of a series of residuals).

Indices based on factor analysis

Factor analysis differs from component analysis primarily in that it estimates uniquenesses for each item given a hypothesis as to the number of factors. There are other differences between the two methods and Hattie (1979, 1980a) has clearly demonstrated that contrary conclusions can result from using the two methods. When using the maximum likelihood estimation method, assuming normality, the hypothesis of one common factor can be tested in large samples by a chi-square test. Using binary data, it clearly cannot be assumed that there is multivariate normality. Fuller and Hemmerle (1966) assessed the effects on chi-square when the assumption of normality was violated. They used uniform, normal, truncated normal, Student's t, triangular and bimodal distributions in a Monte Carlo study. They concluded that the chi-square was relatively insensitive to departures from normality. Their study was confined to sample sizes of 200, five items, and two factors; and it is not clear what effect departures from normality have on the chi-square for other sets of parameters, particularly when binary items are used.

It should be noted, however, that the chi-square from the maximum likelihood method is proportional to the negative logarithm of the determinant of the residual covariance matrix, and is therefore one reasonable measure of the nearness of the residual covariance matrix to a diagonal matrix. This property is independent of distributional assumptions and justifies the use of the chi-square for the present purpose applied to the matrix of tetrachorics from binary data. (But this does not justify use of the probability table for chi-square.)

Further, it is possible to investigate whether two factors provide better fit than one factor by the difference in chi-squares. Joreskog (1978) conjectured that the chi-squares from each hypothesis (for one factor and for two factors) are independently distributed as a chi-square with ($df_2 - df_1$) degrees of freedom. Given that many of the chi-square values typically reported are in the tails of the chi-square distribution, it is not clear what effect this has on testing the difference between two chi-square values.

Instead of using chi-squares, McDonald (1980a) has recommended that the residual covariance matrix supplies a nonstatistical but very reasonable basis for judging the extent of the misfit of the model of one factor to the data. Further, McDonald argued that in practice the residuals may be more important than the test of significance for the hypothesis that one factor fits the data, since the hypothesis, "like all restrictive hypotheses must be false and will be proved so by the chi-square test on a sufficiently large sample. If the residuals are small the fit of the hypothesis can still be judged to be 'satisfactory'" (p.8).

Tucker and Lewis (1973) provided what they called a "goodness of fit" test based on the ratio of the amount of variance associated with one factor to total test variance. The suggestion is that for the one factor case this "reliability" coefficient may be interpreted as indicating how well a factor model with one common factor represents the covariances among the attributes for a population of objects. Lack of fit would indicate that the relations among the attributes are more complex than can be represented by one common factor. The sampling distribution of the Tucker-Lewis coefficient is not known, and there is no value suggested above which it could be concluded that a set of items is unidimensional. It is also not clear why the authors do not condone using the statistic to indicate how well the hypothesised number of factors explain the interrelationships, yet they do claim that it summarizes the quality of interrelationships (see Tucker and Lewis, 1973, p.9).

Dubois (1970) argued that it should be possible for a unidimensional test to attain a correlation of .95 between the obtained raw score and the factor that is common to the items. Dubois believed that such a coefficient could be a reportable statistic for any test claiming unidimensionality.

Maximizing Reliability

One index that has been rediscovered often is maximized-alpha (Lord, 1958; Armor, 1974):

$$\text{maximized-alpha} = (n/(n-1))(1 - 1/\lambda_1), \tag{5}$$

where λ_1 is the largest eigenvalue. Armor compared maximized-alpha and alpha for the single factor case and concluded that maximized-alpha and alpha differ in a substantial way only when some items have consistently lower correlations with all the remaining items in a set. This led Armor to propose what he termed "factor scaling" which involves dropping items that have lower factor loadings. He suggested dropping any item with factor loadings less than .3. Then as alpha and maximized-alpha do not differ by much when only high loading items are retained, Armor proposed that rather than bothering to get the set of weights that would lead to maximized-alpha, the test developer could use unit weights (which leads to alpha). Armor also suggested that maximized-alpha could be used to discover multidimensionality, although he did not specify how this was done. The obvious interpretation of Armor's remarks is that a large maximized-alpha is an indication of a unidimensional test, whereas a low value indicates a multidimensional test. Unfortunately

Armor only used four examples in his comparison of maximized-alpha and alpha and the differences between these two coefficients were .05, .02, .002, and .0, none of which give validity to his claim of "substantial" differences. A more extensive Monte Carlo study is necessary to assess differences between the methods.

Along a similar line, Nishisato (1980) discussed calculating the standard deviation of each item-score obtained by using option weights that maximize alpha, and retaining those items which have relatively large standard deviations. He stated that this procedure guarantees selecting a set of most internally consistent items in the sense that alpha is maximized. Nishisato (1979, 1980) also presented an example that demonstrated the effect that maximized-alpha has on test length and information content compared with other weighting schemes.

Omega

McDonald (1970) and Heise and Bohrnstedt (1970) independently introduced another coefficient that has been used by other researchers as an index of unidimensionality. This index, called theta by McDonald and omega by Heise and Bohrnstedt, is based on the factor analysis model and is a lower bound to reliability, with equality only when the specificities are all zero. Omega is a function of the item uniquenesses and these can be satisfactorily estimated only by fitting the common factor model efficiently. Approximations from the component analysis would generally yield underestimates of the uniquenesses and thus lead to spuriously high omegas.

Should factor analysis be used on binary items?

It has often been claimed that because of "difficulty factors", the factor analytic model is not appropriate for binary data (e.g., Cotton, Campbell & Malone, 1957; Kim & Rabjohn, 1980). Consideration of the issues related to the propriety of using factor analysis on dichotomous data leads to a discussion of two methods that have features particularly appropriate to analysing binary data: nonlinear factor analysis (McDonald, 1967a) and the method of Christoffersson (1975) and Muthen (1978) for factor analysis of dichotomous data.

Difficulty factors.

A problem with using binary data is the presence of so-called difficulty factors. This problem has a long history dating back to Spearman (1927) and Hertzman (1936), and often is cited as a reason against performing factor analysis on dichotomous data (Gorsuch, 1974).

Guilford (1941) in a factor analysis of the Seashore Test of Pitch Discrimination obtained a factor that was related to the difficulty of the items. That is, the factor loadings showed a tendency to change as a linear function of item difficulty. Three possibilities were suggested to account for the presence of the difficulty factor. It may have something (unspecified) to do with the choice of correlation coefficient; it may be that difficulty factors are related to distinct human abilities; or it may have something to do with chance or guessing.

Wherry and Gaylord (1941) argued that difficulty factors are obtained because phi and not tetrachoric correlations are used. This was, they argued, because tetrachorics will be one in cases of items measuring the same ability regardless of differences in difficulty, whereas the sizes of phi's are contingent upon difficulty. They contended that if difficulty factors were found even when tetrachorics were used (as in Guilford's case) then this must be considered disproof of the unidimensional claim.

Tetrachoric correlations

In deriving the tetrachoric correlation it is supposed that the item scores are estimates of continuous, normally distributed and linearly related variables. There have been many formulae suggested for calculating tetrachorics. Elderton (1906) recommended a method based on an infinite series and argued that the first seven terms should be calculated. McNemar (1955) used the first four, the IBM Scientific Subroutines package (1970) the first six, the IMSL package (1980) the first seven, and Christoffersson (1975) used the first 10 terms. Variants of Elderton's method were used by Castellan (1966). Kirk (1973) argued that his 8-point Gaussian integration supplemented by the Newton-Raphson method was more accurate than Saunders' method (detailed in Froemel, 1971), and Divgi (1979) demonstrated that a similar method to Kirk's was even faster than Kirk's method! The point is that a decision as to the best method is yet to be made although some commonly used methods are known to be poor, e.g., the cos-pi formula (Brown & Benedetti, 1979). Many of the methods are based on Elderton's infinite series and discussion of these relates merely to the recommended length of the series approximation. A close inspection of simulation studies comparing the various methods indicates that (as might be expected) it is the extreme values that are most often poorly determined.

Consider the following example where extreme values cause problems. Guilford (1965) presented a case where the cells (a,b,c,d) were (1,9,9,81) and (0,10,10,80), which have tetrachorics of 0.0 and -1, respectively. Here there are only two differences in the allocation of subjects yet a wide discrepancy in the correlation. This example may not be atypical as it is easy to imagine a person guessing the correct answer to one item, which could thus account for the small difference in observations. Using a different line of argument, Carroll (1945) also demonstrated that tetrachorics can be affected by guessing and that the values of tetrachorics tend to decrease as the items become less similar in difficulty. Lord (1980) was emphatic that tetrachorics should not be used when there is guessing.

There have been various suggestions that tetrachorics should be interpreted with caution if the resulting values are very low or very high. Pearson (1901) recommended caution when the tetrachorics are less than .16 or above .84. Yet tetrachorics have continued to be used for assessing dimensionality primarily because, it is claimed, they help prevent difficulty factors from emerging. But empirically, the sample matrix of tetrachorics is often not positive-definite (i.e., non-Gramian). This may be a problem when (perhaps without justification) using maximum likelihood computer programmes as opposed to least squares methods. Further, Lord and Novick (1968) have contended that tetrachorics cannot be expected to have just one common factor except under certain normality assumptions, whereas such distributional considerations are irrelevant for dimensionality defined in terms of latent traits.

Gourlay (1951) was aware of the effect of guessing on tetrachorics yet argued that a more fundamental problem was that the test items often violated the assumption of normality. This hint was to lead to an important discovery by Gibson (1959, 1960) who recognized that difficulty factors can be considered as being caused by nonlinear regressions of tests on factors. However a general treatment of nonlinearity was not given until McDonald (1962a, 1965a, 1965b, 1967a, 1967b, 1967c, 1976a).

Before discussing nonlinear methods it should be noted that there have been empirical comparisons between phi's and tetrachorics (see Comrey and Levonian, 1958; Dingman, 1958; and Guilford, 1965). Generally the conclusion was that phi coefficents were not as ill-behaved as many would have us believe.

Nonlinear factor analysis

McDonald (1965a, 1967a) used a form of nonlinear factor analysis to derive a theory for difficulty factors. Using a factor analysis of subtests of Raven's Progressive Matrices, McDonald demonstrated that a difficulty factor emerged, in that the loadings of the subtests were highly correlated with the subtest means. He showed that this factor corresponded to the quadratic term in a single factor polynomial model and hence argued that the difficulty factor corresponded to variations in the curvatures of the regressions of the subtests on a single factor. McDonald and Ahlawat (1974) in a Monte Carlo study convincingly demonstrated that if the regressions of the items on the latent traits are linear, there are no spurious factors; that a factor whose loadings are correlated with difficulty need not be spurious; that binary variables whose regressions on the latent trait are quadratic curves may yield 'curvature' factors but there is no necessary connection between such 'curvature' effects and item difficulty; and that binary variables that conform to the normal ogive model yield, in principle, a series of factors due to departure from a linear model. The conclusion was clear: the notion of 'factors due to difficulty' should be dropped altogether and could reasonably be replaced by 'factors due to non-linearity'.

McDonald (1979) described a method of nonlinear factor analysis using a fixed factor score model which, unlike the earlier random model (McDonald, 1967a), obtains estimates of the parameters of the model by minimizing a loss function based either on a likelihood ratio or on a least squares function. Etezadi (1981) has investigated numerical methods for the multifactor cubic case of this method with first-order interactions. Using empirical and simulated data, Etezadi demonstrated little difference between the least squares and the maximum likelihood results. The least squares method was computationally much faster. Although it would appear from recent work (Etezadi, 1981; McDonald, 1980a) that a cubic model is desirable to give a good approximation to a normal ogive, McDonald (1967a) has reported cases where a quadratic model supplies good fit; good fit was assessed by the size of the residuals.

While nonlinear factor analysis is conceptually very appealing for attempting to determine dimensionality it has been used, unfortunately, relatively infrequently. Other than work done by McDonald, McDonald and Ahlawat, and Etezadi, it was possible to find only one use of the method. Lam (1980) found that one factor with a linear and a quadratic term provided much better fit to Raven's Progressive Matrices than a one or two factor linear model. If nonlinearities are common when dealing with binary data then a nonlinear factor analysis seems necessary. Moreover, if a one factor cubic provides good fit to a set of data then the rank of the inter-item correlation matrix is three. Hence, the claim that unit rank is a necessary condition for unidimensionality is incorrect.

As in the linear case two indices that can be used to assess the fit of one latent trait to the data are the absolute sum of squares of residuals and the number of residuals greater than a specified value, for example, greater (in absolute value) than 0.01.

Corballis (1968) claimed without proof that nonlinear factor analysis cannot be used on binary data in spite of McDonald's mathematical and numerical demonstrations to the contrary. Yet Corballis was much more critical of a method due to Horst that aimed at eliminating difficulty factors.

Horst's method

Horst (1964) contended (prior to McDonald's evidence and theory) that the dimensionality of a set of items will be accounted for in part by the dispersion of item preferences and that it would be useful to devise procedures for segregating this part. Following Guttman (1954) he called that part which is due to variations in item difficulties the simplex part. After this simplex is removed, the residual covariance matrix is factored, presumably by linear common factor analysis, to "yield the dimensionality of a set of items after they have been freed from the effects of the simplex phenomenon" (Horst, 1964, p. 134).

Corballis (1968) mounted a strong attack on Horst's method. His most effective argument is that if we have a perfectly scaled test then when we remove the simplex we get zero dimensionality. Thus the method removes much (or, in this case, all) of what we are looking for.

It seems that Horst's method is of limited use or of no use for determining the dimensionality of sets of items. To further confirm this conclusion, Hoffman and Gray (1978) used simulated data to investigate what happened when Horst's method was used. Without exception they found that there was no recovery of factors when using the Horst procedure, factor solutions resulted in bipolar and split factors, and two of their four simulated matrices were singular. The splitting usually resulted in high difficulty items loading on one factor and low difficulty items loading on a second factor, the precise effect Horst's method was intended to avoid. They concluded that Horst's method was of no benefit.

"Difficulty factors" summarized

It has been demonstrated that the spurious factors that many researchers used as reasons against the use of the common factor model were due not to the choice of correlation, but to nonlinearities. There has been a recent surge of interest in using tetrachorics as the basis for factor analysis as evidenced by the work of Bock and Lieberman (1970), Christoffersson (1975), and Muthen (1978).

Christoffersson and Muthen's method

Bock and Lieberman (1970) distinguished between two possible approaches to estimating parameters in latent trait theory. In "unconditional" estimation of item parameters the data are regarded as arising from a sample from a specified population and integrated over the distribution of latent ability in order to estimate the item parameters. The "conditional" method treats subjects as arbitrarily given and estimates item and ability parameters simultaneously. (Note that Bock et al. are not following accepted usage of the terms "conditional" and "unconditional".) The unconditional approach has been used by Lawley (1944) and Lord (1952) in what Bock and Lieberman call the heuristic method. This involves estimating item difficulty by using the normal deviate corresponding to percent correct in the sample and the item discrimination power is estimated by the loading of the item in a one-factor linear common factor analysis of the matrix of sample tetrachoric correlations.

Bock and Lieberman detail an unconditional maximum likelihood method that overcomes the difficulty posed by possible non-positive definiteness of the tetrachoric correlation matrices. The basic data used are the pattern of a subject's item responses across all n items in the test. The probability of a given response pattern, p_i, is used in the likelihood equation

$$L = (n!/\prod_i r!) \prod^{2n} p_i^{r_i} \qquad (6)$$

where r_i is the number of subjects observed to have a particular response pattern. The likelihood equation specifies the probability of joint occurrence of all possible patterns of item responses. Then standard maximum likelihood procedures are used to obtain the normal equations to be solved for the estimates of the item parameters. Bock and Lieberman compared these item-parameter estimates, transformed into a standard item-difficulty index and a correlational item-discrimination index that they termed a reliability index (an estimate of the point-biserial correlation between item score and the latent trait), with those from the unconditional heuristic approach for two sets of data. The item difficulty estimates agreed to the third decimal place. The item discrimination estimates agreed to a difference of unity in the second decimal place. Bock and Lieberman concluded that "the excellent agreement of the two methods depends upon the use of exact tetrachoric correlation coefficients in the factor analysis" (p. 193) when using the heuristic method. Despite their similar results from the two methods they state that they used a poor method for estimating tetrachorics and recommended Froemel's method (which is based on Elderton's series method).

They reported one important difference between the two methods and this relates to dimensionality. Bock and Lieberman argued that for the heuristic method, goodness of fit is usually checked by analysing the matrix of tetrachorics specifying a one factor solution and then inspecting the eigenvalues of the matrix of correlations rescaled by the inverses of the unique loadings. For their data they reported a sharp break between the size of the first root and that of the remaining roots, with the latter clustering around one and they concluded that "these results would ordinarily be taken to support the assumption of unidimensionality" (p. 191). The likelihood ratio tests of one factor from a maximum likelihood factor analysis gave chi-squares of 25.88 and 53.74, each on five degrees of freedom for the two data sets employed. Clearly in both cases a one-factor solution must be rejected. Bock and Lieberman concluded that the 'teet of unidimensionality' provided by maximum likelihood factor analysis cannot be relied upon when tetrachoric correlations are used in place of the estimates which their procedure assumes.

A major obstacle to using the Bock and Lieberman method is a practical one. The method requires the use of 2^n possible response patterns across all items which, for example, would require 32768 possible patterns for a 15 item test, and more than 107 million response patterns for a 30 item test. Further, a 2n by 2n matrix must be inverted and few computers have so much storage.

Nishisato (1966, 1970a, 1970b, 1971; see also Lim, 1974; Svoboda, 1972) and Christoffersson (1975) demonstrated that very little efficiency in estimation is lost using only information from the first and second order joint probabilities of binary scored items compared to using all possible 2^n proportions, as in the Bock and Lieberman approach. Muthen (1978) developed an estimation method that was computationally faster than Christoffersson's. There seems to be little difference between results obtained by using the Christoffersson-Muthen methods and those from maximum likelihood factor analysis based on the tetrachoric correlations. To illustrate the similarities, the various methods have been applied using one of the data sets provided in Bock and Lieberman (Table 3). Clearly the differences ~~are~~ very small, at most .02 for the difficulty values and at most .05 for the reliability indices. The methods of Bock and Lieberman, Christoffersson, and Muthen may be preferred because there is no possibility of non-positive definite matrices, but some users will have problems of access to the programmes, and of machine incompatabilities when not using IBM machines. There are also severe limits on the number of items that can be analysed, and the methods can yield Heywood variables.

Bock and Aitken (1981) also pointed out the practical limitations of the Bock and Lieberman work, but stated that the Christoffersson and Muthen method, from a statistical point of view "is also objectionable because it assumes that the form of the distribution of ability effectively sampled is known in advance. Since item calibration studies are typically carried out on arbitrarily selected samples, it is difficult to specify _a priori_ the distribution of ability in the population effectively sampled" (p. 3). Instead, Bock and Aitken proposed to estimate the item parameters by integrating over the empirical distribution.

Table 3

Comparison among the Bock and Lieberman (B&L), heuristic, Christoffersson (CH), and Muthen (MU) methods to calculate item difficulty and item reliability using 5 items from the Law School Admission test

Item No.	B&L	Heuristic	CH GLS	CH TLS	MU GLS	MU ULS
			Item difficulty			
1	.946	.946	-.964	-.948	-.946	-.946
2	.407	.407	-.409	-.405	-.407	-.407
3	.745	.745	-.761	-.755	-.746	-.745
4	.268	.268	-.274	-.270	-.269	-.268
5	1.007	1.007	-1.020	-1.007	-1.007	-1.007
			Reliability index			
1	.489	.483	.492	.507	.519	.492
2	.544	.543	.557	.555	.513	.536
3	.702	.694	.692	.692	.678	.712
4	.420	.429	.427	.433	.433	.418
5	.381	.385	.369	.382	.395	.377

Note. GLS = Generalized least squares

TLS = Two stage least squares

ULS = Unweighted least squares

The Bock and Aitken method was applied to the same data as was the Bock and Lieberman, Christoffersson, and Muthen methods (see Bock & Aitken, 1981). The differences in the estimates are very small. It is also debatable whether the observed distribution of ability, with the usual sampling errors and errors of measurement, is the best distribution to work from.

Green, Lissitz and Mulaik revisited

Before concluding this section on indices based on factor analysis we return to Green, Lissitz and Mulaik (1977) who have suggested two further indices. The first they called u , given by

$$u = \sum_{i\neq j}\sum |\overline{r}_{ij}| / \sum_{i\neq j}\sum (h^2{}_i h^2{}_j)^{.5} \qquad (7)$$

where h^2 is a communality from a principal component analysis. Unlike many other proposers of indices, Green et al. did offer a rationale for their index. When there is a single common factor among the items, the loadings on this factor equal the square root of their respective communalities. Also the correlation between any two items equals the product of their respective factor loadings or the square root of the product of the respective communalities. For any particular pair of items i and j, Green et al. suggest that the inequality $|\overline{r}_{ij}| \leq (h^2{}_i h^2{}_j)^{.5}$ holds. Equality is attained for items occupying a single common factor space and the inequality is strict for items occupying more than one dimension. When there is one factor, u equals 1; when there are more factors, u takes on values less than 1 and has a lower limit somewhere above 0.

Green et al. calculated the value of u in their Monte Carlo study (described above, pp. 24–25) and found u to be relatively independent of both the number of items and the communality of the items. While it did increase as the number of factors loading on an item increased, they claimed it did not increase as much as alpha. Table 4 presents some summary data taken from the Green et al. paper and it appears that contrary to their claim, u is affected as much as alpha, although the values are much lower than alpha. Further, u does seem to distinguish between a one-factor solution and a more-than-one-factor solution but it does not relate in any systematic way to the number of extra factors required.

A serious problem with u is that it requires knowledge of the communalities of the items, which depends upon knowledge of the correct dimensionality. In Green et al.'s simulation they provided the communalities, but in practice these are not known and u can be (and nearly always is) larger than 1, as $|\overline{r}_{ij}| \geq (h^2_i h^2_j)^{.5}$ (because the communalities are usually underestimated). The usefulness of u is therefore most questionable.

Green et al. suggested a further index. Given that $|\overline{r}_{ij}|$ is affected by the communalities of the items, one aim would be to counter this effect. If the correlations are first corrected for communality by dividing them by the product of the square roots of their respective communalities in the same manner as one would calculate a correction for attenuation, and the resulting values averaged, then one gets an index not affected by the communalities. Green et al. contended that such an index also takes on values 0 to 1,

Table 4

Green, Lissitz and Mulaik's indices broken into subsets according to the number of common factors among the items, and the number of factors relating to a single variable

	No. of factors relating to a single item				No. of common factors among the items					
	1	2	3	4	1	2	3	4	5	6
$\bar{r}_{ij}$	.24	.24	.33	.40	.60	.30	.19	.22	.23	.24
alpha	.85	.91	.93	.96	.96	.93	.86	.89	.86	.89
u	.39	.40	.54	.66	1.00	.49	.32	.36	.38	.40
r	.43	.46	.63	.75	1.05	.52	.35	.41	.45	.47

with 1 indicating unidimensionality. This index also depends upon knowledge of the communalities and is affected by the number of factors determining a variable (see Table 4).

Second-order factor analysis

Lumsden (1961) claimed that it is possible that variance in important group factors may be obscured if the group factors are each measured by only a single item in the set that was tested for unidimensionality. He gave as an example four items whose ideal factor constitutions were

$G + V + S_V + E$	(e.g., verbal analogy = V)
$G + N + S_N + E$	(e.g., number series = N)
$G + K + S_K + E$	(e.g., matrix completion = K)
$G + M + S_M + E$	(e.g., mechanical problem = M)

where G is a general factor, S a specific factor, and E is error. The correlation between these items will be determined by the product of the G loadings, and Lumsden stated that the remainder of the variance would be treated as error. He then wrote that the "matrix of item intercorrelations for these four items will be of unit rank. A verbal analogy, a number series, a matrix completion and a mechanical problem are not, however, measuring the same thing and unit rank is not, therefore, a sufficient condition of unidimensionality" (pp. 107-108). Obviously, we can argue (and I would say correctly) that these items do indeed measure a unidimensional trait (e.g., intelligence). It is quite reasonable to find a second-order factor underlying a set of correlations between first-order factors and then make claims regarding unidimensionality. Hattie (1981), for example, used a second-order unrestricted maximum likelihood factor analysis to investigate the correlations between four primary factors he identified (and cross validated) on the Personal Orientation Inventory (Shostrom, 1966, 1972). The hypothesis of one second-order factor could not be rejected in nine of the eleven data sets and Hattie concluded that "it thus seems that there is a unidimensional construct underlying the major factors of the POI" (p.4), which could be identified as self-actualization. As already argued, any decision criterion for assessing dimensionality does not necessarily identify the nature of the unidimensional set. A judgement and investigation of the validity of the data set is still required by the researcher.

Indices based on latent trait models

A theory of latent traits is based on the notion that responses to items can be accounted for, to a substantial degree, by defining k characteristics of the examinees called latent traits, which we will denote by $\underset{\sim}{\theta} = (\theta_1,\theta_2 \ldots,\theta_k)$. The vector $\underset{\sim}{\theta}$ is a k-tuple that is interpreted geometrically as a point in k-dimensional space. The dimensionality of the latent space is one in the special case of a unidimensional test. The regression of item score on theta is called the item characteristic function. For a dichotomous item, the item characteristic function is the probability $P(\theta)$, of a correct response to the item. For a one dimensional case a common assumption is that this probability can be represented by a three-parameter logistic function

$$P(\Theta) = c + (1 - c)(1 + e^{-da(\Theta-b)})^{-1} \quad (8)$$

or alternatively by a normal ogive function

$$P(\Theta) = c + (1 - c) \int_{-\infty}^{a(\Theta-b)} (1/\sqrt{2\pi}) e^{-z^2/2} dz \quad (9)$$

The difference between (8) and (9) is less than .01 for every set of parameter values when d is chosen as 1.7 (Haley, 1952). Lord (1980) notes that in principle, examinees at high ability levels should virtually never answer an item incorrectly whereas, in practice, such an examinee occasionally makes a careless mistake. Since the logistic function approaches its asymptotes less rapidly than the normal ogive, such mistakes will have less effect for the logistic than for the normal ogive model. Hence Lord suggests that the logistic model is probably preferable in practical work (and therefore will be used in this dissertation).

In the following sections the fundamental assumption of the model is stated, the parameters of the model defined, and various (sub) models and estimation procedures are outlined. As with the description of the factor analysis model, only theory relevant to the present thesis topic and for the discussion of various suggested indices is presented.

A fundamental assumption

The most critical and fundamental assumption is that of local independence (Anderson, 1959; McDonald, 1962b). Lord (1953a) has claimed that this is almost indispensable for any theory of measurement. The principle of local independence requires that any two items be uncorrelated when theta is fixed and does not require that items be uncorrelated over groups in which theta varies. Lord and Novick (1968) give the definition of local independence more concrete meaning by saying that

> an individual's performance depends on a single underlying trait if, given his value on that trait, nothing further can be learned from him that can contribute to the explanation of his performance. The proposition is that the latent trait is the only important factor and, once a person's value on the trait is determined, the behaviour is random, in the sense of statistical independence. (p. 538)

Unfortunately, there has been much misunderstanding about the principle of local independence, particularly the relationship between the method of factor analysis and the principle (e.g., Goldstein, 1981). Many of the earlier attempts at clarifying the relationship confused the problems of estimating the latent traits with the theory and definition of latent traits; for example, see Lazarsfeld (1959). McDonald (1967a, 1981) has provided the clearest statement of the relationship.

McDonald argues that the statement of the principle of local independence contains the mathematical definition of latent traits. That is, $\theta_1, \ldots, \theta_k$ are latent traits, if and only if they are quantities characterizing examinees such that, in a subpopulation in which they are fixed, the scores of the examinees are mutually statistically independent. Thus a latent trait can be interpreted as a quantity that the items measure in common, since it serves to explain all mutual statistical dependencies among the items. Since it is possible for two items to be uncorrelated and yet not be entirely statistically independent, the principle is more stringent than the factor analytic principle that their residuals be uncorrelated. If we reject the principle of local independence in favour of some less restrictive principle then we cannot retain the definition of latent traits, since it is by that principle that latent traits are defined. McDonald points out that we can, however, reject or modify assumptions as to the number and distribution of the latent traits, and the form of the regression function (e.g., make it nonlinear instead of linear), without changing the definition of latent traits.

It is not correct to claim that the principle of local independence is the same as the assumption of unidimensionality (e.g., see Hambleton, et al., 1978, p.487; Gustafsson, 1980, p.218). The principle of local independence holds (by virtue of the above argument), for 1, 2, ..., k latent traits. It even holds for the case of zero latent traits, that is, for a set of n items that are (unconditionally) mutually statistically independent. Unidimensionality is defined here as the existence of one latent trait underlying the set of items.

McDonald (1981) outlined ways in which it is possible to weaken the strong principle, in his terminology, of local independence. The strong principle implies that not only are the partial correlations of the test items zero when the latent traits (which are the same as factor scores) are partialled out, but also the distinct items are then mutually statistically independent, and their higher joint moments are products of their univariate moments. A weaker form, commonly used, is to ignore moments beyond the second order and test the dimensionality of test scores by assessing whether the residual covariances are zero (see also Lord & Novick, 1968, pp. 544-545, 225). Under the assumption of multivariate normality the weaker form of the principle implies the strong form, as well as conversely. McDonald (1979, 1981) argued that this weakening of the principle does not create any important change in anything we would wish to say about the latent traits, though strictly it weakens their definition. Further, McDonald (1979) suggested that it may be reasonable to abandon the principle of local independence in favour of conditional patterning, under which each member of a population of examinees is characterized by one or more quantities such that the covariance matrix of the residuals about the regressions of $y_1, \ldots, y_n$ on some $x_1, \ldots, x_k$ has a prescribed pattern. That is, certain elements of the residual matrix are constrained to be zero. Certainly if McDonald's weak form of the principle is adopted, it is not appropriate to call the regressor variables $x_1, \ldots, x_k$ latent traits since they no longer account for all the relations between the items, and if by the dimensionality of a set of items we mean the number of latent traits then it is clear that it is not meaningful to abandon the principle of local independence while retaining the concept of dimensionality.

Lord (1953a) in an early specification of the assumption (his restriction IV) suggested a heuristic for assessing whether the assumption of local independence is met in a set of data. Take all examinees at a given score level and apply a chi-square test (or an exact test) to determine if their responses to any two items are independent. Then, because the distribution of combined chi-squares is ordinarily also a chi-square (with $df = df_1 + df_2 + \ldots + df_n$

degrees of freedom) the resulting combined chi-square may be tested for significance. If the combined chi-square is significant then Lord claimed it must be considered that the test is not unidimensional. When assessing how this statistic behaved (for subsequent inclusion in this thesis) it soon became obvious that it does not matter whether a chi-square or an exact test is used, as in both cases the probability is nearly always close to one.

Parameters of the latent trait model

Discrimination

The parameter a in (8) or (9) represents the discriminating power of the item, the degree to which the item response varies with ability level. It is proportional to the slope of $P(\theta)$ at the point $\theta = b$ which there takes its maximum value of $.425a(1-c)$ for the logistic model (Lord, 1980). Typically empirical values fall between .5 and 2.0. Ree (1979) claimed that values typically range from .5 to 2.5 with a mean of 1, that items with discrimination less than .5 are insufficiently discriminating and that values greater than 2.0 are infrequently found. Ree used values between .65 and 1.61 ($\overline{X}$ = .95, s.d. = .28) in a simulation study. Ross (1966) used two twenty-item multiple-choice tests (N=1000) and reported ranges of .47 to 1.99 and .30 to 1.97, with means of 1.28 and 1.06 for the two tests, respectively. In an empirical study of 80 SAT verbal items (N=2995) Lord (1968) reported a range from .4 to 1.7 ($\overline{X}$ = 1.07, s.d. = .40).

Hambleton and Traub (1971) in a simulation study chose a mean value of .59. This value is low for typical tests but Hambleton and Traub chose their value because when related to the usual biserial (and assuming theta is normally distributed and there is no guessing) it represents a biserial of .50. This is because $a = r_b/(1-r_b)^{.5}$. (See Lord, 1980, and Urry, 1974, for comparisons between latent trait and conventional item analysis statistics.)

Lord (1968, 1975) found that when a is being estimated, it may increase without bound, although typical values are less than +2. Lord fixed an arbitrary upper bound of 1.7 in his computer programme but Wright (1977) has used the need for this bound as evidence that discrimination parameters are not estimable.

Difficulty

Parameter b in equations (8) and (9) is a location parameter and determines the position of the item characteristic curve along the ability scale. Typically referred to as item difficulty it represents the inflexion point which is $\theta = b$ in the logistic model. For simulation studies, values are commonly rectangularly distributed between -2 and +2 (Hambleton & Traub, 1971; Kingsbury & Weiss, 1979; Panchapakesan, 1969; Urry, 1970, 1971, 1974). Ahlawat (1976) chose values of b between -1 and +1, reflecting a narrower range of difficulty. Dinero and Haertel (1977) in a 30 item simulation reported values between -1.53 and 2.07 ($\overline{X}$ = .11, s.d. = .94). Ree (1979) claimed values typically fell between ± 3, yet set a range of

±2.5 as bounds in his simulation and reported a range between -1.65 to 1.97. Empirically, Lord (1968) found values between -1.5 and 2.5 ($\overline{X}$ = .58, s.d. = .87) and Ross (1966) had ranges from -3.14 to .49 and -2.07 to 1.42.

Guessing

The parameter c in equations (8) and (9) is the height of the lower asymptote, the probability that a person completely lacking in ability ($\theta = -\infty$) will answer the item correctly. It is called the guessing or pseudo-chance level. If an item cannot be answered correctly by guessing, then c = 0. For a 5-option multiple-choice item, for example, the theoretically expected value of c is 1/5 = .20; yet in practice values less than .20 are typically found as test writers usually attempt to provide attractive incorrect alternatives.

Lord (1968) found a mean of .16 (s.d. = .01) and a range between .04 and (his imposed upper limit of) .20 for the 80 SAT five-option verbal items. In simulation studies using 5-option items, Hambleton and Traub (1971) chose values of c between 0.0 and 0.2, Kingsbury and Weiss (1979) chose constant values of .10 and Ree (1979) used values between .09 and .35 ($\overline{X}$ = .20, s.d. = .05).

Ability

The latent trait θ provides the scale on which all item characteristic curves have some specified mathematical form, for example the logistic or normal ogive. A joint underidentifiability of

θ, a, b, and c, is removed by choosing an origin and unit for θ.

Latent trait models

Using a combination of these parameters, several latent trait models have been suggested. All models assume the principle of local independence.

Three-parameter model

The mathematical form of the three-parameter logistic curve has been given in (8). Hambleton et al. (1978) have provided an excellent overview of the procedures used in estimation and have detailed various numerical methods commonly used. Ree (1979) compared three programmes for estimating the three-parameter models and found that no one programme was optimal over all conditions. For data-sets based on rectangular distributions of ability, ETS's LOGIST (Wood, Wingersky and Lord, 1976) was superior, while Urry's OGIVIA (Urry, 1977, 1978) provided better estimates for normally distributed data, yet both OGIVIA and ANCILLES (Urry, 1977, 1978) could not always estimate the parameters of all items. Hutten (1980), using LOGIST on 1328 items, reported that only one third of guessing parameters were estimable for 1000 subjects in each of 25 tests.

The programming of these three-parameter methods usually involves various restrictions, for example, fixing bounds on the discrimination, or, more often, making stipulations about the guessing parameter. The c value can be estimated with maximum likelihood methods but Lord (1968) has indicated that this does not work well.

Convergence is slow and absurd values are sometimes obtained (e.g., negative values). Jensema (1976) suggested that a simple way to estimate c is to 'eyeball' the lower tail of a plot of the proportion of examinees responding correctly to an item at each test score. It is typical to chose some fixed value for c, usually zero or the reciprocal of the number of options.

Two-parameter model

If c is fixed (perhaps at zero) then we have two parameters to estimate and this is statistically more tractable. Lord (1953b) and Birnbaum (1968) have obtained maximum likelihood estimates of the parameters in the two-parameter normal ogive and the logistic models, respectively. Also there have been proposed heuristic estimation procedures (e.g., Urry, 1974; Jensema, 1976, and Schmidt, 1977), and Bayesian methods (Birnbaum, 1969; Owen, 1975). The methods of Bock and Aitken (1980), Bock and Lieberman (1970), Christoffersson (1975), and Muthen (1978), described above as methods for factor analysing tetrachoric correlations can immediately be recognized as equivalent, in the one-factor case, to fitting the two-parameter normal ogive model.

McDonald (1980a) developed a method to fit a latent trait model by the analysis of covariance structures. Using a harmonic analysis of the normal ogive model in terms of orthogonal polynomials (McDonald, 1967a), he fitted the coefficients of the first few terms of the orthogonal polynomial series as common factor loadings of item covariances. This method allows estimation of the difficulty and discrimination parameters when guessing is presumed to be zero or some known value. When McDonald and Fraser (in prep.) were testing the robustness of this method using a bimodal distribution obtained by mixing two normal distributions, they observed that very poor results were occasionally obtained in subsamples with seemingly representative true parameter values. It was noted that a method introduced by Fraser into the computer programme gave much better estimates of these "bad" cases, and did not give appreciably worse estimates in any cases. McDonald and Fraser called this method the "normal ogive by harmonic analysis robust method" (NOHARM). This is a very fast method and can easily handle large data sets. Even with estimates of guessing read into the programme, McDonald presented results that showed that the NOHARM method was extremely robust against departures from normality.

One-parameter model

The most used and the most researched model involves only the estimation of the difficulty parameters. It is often called the Rasch model after the pioneering work of Rasch (1960, 1961, 1966a, 1966b, 1968, 1977). It is assumed that there is no guessing, that the discrimination parameter a is constant (not necessarily equal to one as many have written: e.g., Baker, 1977; Reckase, 1979), and, of course, that the principle of local independence applies. There are various methods for estimating the parameters in the one-parameter model (e.g., Brooks, 1965; Cohen, 1976; Gustafsson, 1980; Rasch, 1960; Wright & Douglas, 1977; Wright & Pachapakesan, 1969; Wright &

Stone, 1979).

That there are many indices of how adequately the data "fits" the model is cited often as one of the major advantages of the Rasch model. In one of the earliest statements, Wright and Panchapakesan (1969) claimed that "if a given set of items fit the (Rasch) model this is evidence that they refer to a unidimensional ability, that they form a conformable set. Fit to the model also implies that item discriminations are uniform and substantial, that there are no errors in item scoring and that guessing has had a negligible effect" (p. 252). Omitting the phrase "and substantial", these sentiments have often been requoted and an earnest effort made to find and delete misfitting items and people. Rentz and Rentz (1979) write that "the most direct test of the unidimensionality assumption is the test of fit to the model that is part of the calibration process" (p. 5). To illustrate the various "fit statistics", consider the following indices that are typically used.

First, the ability of persons with all or no responses correct cannot be estimated as, it is argued, we do not have information as to just how much more or less able these people are compared to the rest of the sample (Wright & Stone, 1979, p. 32). Similarly, items that are answered incorrectly or correctly by all persons are omitted from the analysis. Then the item and person parameters are estimated and the "most natural" fit statistic, the between-fit t is calculated (Wright, Mead & Bell, 1979). The reason it is the "most natural" is that it "is derived directly from the 'sample-free' requirements of the model" (Wright, Mead & Bell, 1979, p. 10). (In fact, however, the statistic is sample-dependent.) The sample is divided into subgroups based on score level according to estimated ability, and then the observed proportion of successes on each item in each ability subgroup is compared with that predicted from the estimates of the item difficulties given by the total sample. Wright et al. derive a standardized mean-square statistic that has an expected value of 0 and variance of one. A more general statistic, a total-fit t, evaluates the general agreement between the variable defined by the item and the variable defined by all other items over the whole sample. Again the expected value is 0 and variance is 1.

Thus the total-fit t summarizes overall item fit from person to person and the between-fit t focuses on variations in item responses between the various subgroups. Wright et al. note that any t value larger than 1.5 ought to be examined for response irregularities and values greater than 2.0 "are noteworthy" (p. 13). They also recommended that a "within-group" mean-square be calculated that summarizes the degree of misfit remaining within ability groups after the between group misfit has been removed from the total. This has an expected value of 1 (see Wright & Stone, 1979, p. 53).

Another fit statistic is error impact, which is the proportional inflation of standard errors of calibration or measurement that could be caused by misfit of the item (see Wright, 1977; Wright, Mead & Bell, 1979, p. 14; Wright & Stone, 1979, pp. 135-136). Its average value over all items is the reciprocal of average test information (see Birnbaum, 1968, pp. 460-468). Thus the greater the information provided by a test the smaller is the error impact and thus the smaller the standard error of measurement and the greater the precision.

Wright recommends the use of a discrimination index which he describes as an index of the linear trend of residual departures from model expectation across ability groups, expressed relative to a modal value of one. When the discrimination index is near 1 then the observed and expected item characteristic curves are close together. When the index is substantially less than 1, the observed item characteristic curve is flatter than expected, hence that item is failing to differentiate among abilities as well as other items. Wright and Stone (1979, p. 53) contend that the discrimination index is related to the point-biserial, but is less influenced by the dispersion in ability of the sample. Wright et al's programme also prints a "person separability index" which is identical to alpha.

One of the advantages often claimed by advocates of the Rasch model is that it is possible to find persons who do not fit the model. These can be rejected when calibrating items or they can be looked at separately and questions can be asked as to why the test seems to be giving poor estimates for these people. References to person variability can be traced back to Mosier (1940, 1942) who argued that a total test score is an imperfect representation of an individual's test score and depends on the individual's variability. Efforts by Thouless (1936) and Anderson (1958) regarding function fluctuation have been subsequently ignored. Levine (1977), Levine and Rubin (1976), Lumsden (1977, 1978), Mead (1975), Vale and Weiss (1975), Waller (1974), Weiss (1973), and Wright (1977) have resurrected the issue, apparently independently of the early work.

Generally, the argument for deleting persons is that these persons have "aberrant" scores because of factors unrelated to the dimension that is assessed by the test. Such factors may include guessing, test-wiseness, fatigue, nervousness, cheating, boredom and cultural bias. Including such persons in a calibration may distort resulting estimates. Garrison and White (1979) found a correlation between true score and observed score of .90 in a group of 200 examinees, but a correlation of only .18 for 25 (known) misfitting people in the group. After these 25 persons were omitted the correlation was .99. Ryan, Garcia-Quintanta and Hamm (1980) found that between two and three percent of subjects were misfitting and that "in the vast majority of cases person misfit occurred when low ability students correctly responded to relatively difficult items" (p. 10). There is no reason why the person-fit statistics cannot be obtained when using the two- and three-parameter models.

From the one-parameter model there has been a vast array of indices suggested, but they appear to lack theoretical bases, little is known of their sampling distributions, and they seem to be ad hoc attempts to help a test developer understand the behaviour of items. Their developers have paid insufficient attention to helping psychometricians understand the assumptions, methods of calculation, and justification for them. One of the obvious problems of many of the tests of fit is that with large samples, a chi-square is almost certain to be significant. For this reason and also because the distribution of the indices is only an approximation to chi-square, it is not clear how valid the methods are for evaluating items. (See George, 1979, for a critical evaluation of this approximation.) There have been critics of claims made by the Rasch proponents, yet any study that assesses unidimensionality must include various "tests of fit" based on their claims. The tests are primarily motivated by a desire to assess the various assumptions, particularly the unidimensionality assumption. Gustafsson and Lindblad (1978) found that tests of fit, like those used by Wright et al., did not lead to rejection of the one-parameter model even for data generated according to an orthogonal two factor model. As an alternative Gustafsson (1980) proposed tests of fit based on the person characteristic curve rather than the item characteristic curve, but he pointed out that prior to using these tests "it does seem necessary to use factor analysis to obtain information about the dimensionality of the observations" (p.217).

So far only the one-, two-, and three-parameter models, which all assume a single latent trait, have been considered. But there have been attempts to derive multidimensional models.

Multidimensional models

Problems relating to generality, estimation and computer resources have usually been given as the reason why there have been so few attempts at developing multidimensional latent trait models. When discussing multidimensional models, first it must be established how the different dimensions relate to each other. Coombs (1964; Coombs & Kao, 1954, 1955) has detailed three methods: the conjunctive model (that is, excess of one ability, no matter how large, does not compensate for the lower ability in other dimensions); the disjunctive model (that is, an examinee will pass an item if he/she is dominant over the item in any one dimension and will fail only if the item dominates him/her in all dimensions); and the compensatory model which assumes that an examinee's reponse to an item is a function of a weighted sum of underlying attributes. If this sum exceeds a particular value the examinee passes the item. If not, he/she fails. It is the compensatory model that is primarily investigated, as there are serious problems, theoretically and computationally, with trying to generalize the latent trait model to take account of conjunctive or disjunctive assumptions (see Coughlan, 1974).

Sympson (1977) objected to the use of the compensatory model on psychological grounds, arguing that it is unreasonable for a person low in one ability to still pass an item if he/she is high in another. Instead Sympson proposed a partially compensatory model where a decrease in one ability could only be offset by a large increase in the other ability. Outside of a relatively narrow ability range, the probability of passing reduces to zero regardless of the value of the stronger ability. The problem, however, was to operationalize the model and provide estimation routines. Sympson did propose a three-stage method based on regression, component analysis and iterative least squares. Sympson claimed that he had solved the programming for the first two stages and was working on the third stage. Some of the problems of Sympson's method are that he estimates the responses to a particular item from the responses to all other items using a multiple linear regression, whereas the regression of items on a latent trait is certainly nonlinear; he sometimes obtains negative estimates for the guessing parameter; and Lord (1977) has questioned Sympson's approximation methods when maximum likelihood methods can be developed and used. Sympson does provide an example of his method, but unfortunately it is for the unidimensional case (which cannot be an example of a partially compensatory model).

To develop a compensatory model consider the linear combination rule (see McDonald, 1979), according to which a dependent variable y has a nonlinear regression on a weighted linear combination of the independent variables $x_1, \ldots, x_n$. Thus for many latent traits, $\Theta_1, \ldots, \Theta_k$, we may write

$$P_j(\Theta_1,\Theta_2, \ldots,\Theta_{ik}) = f(\Sigma a_{jl}\Theta_l - b_j) \qquad (10)$$

where $j = 1, \ldots, n$ and $l = 1, \ldots, k$. Hence, a multivariate logistic model can be written as

$$P_j(\Theta_{i1},\Theta_{i2},\ldots,\Theta_{ik}) = c+(1-c)(1 + e^{-d(\Sigma a_{jl}\Theta_l - b_j)})^{-1} \qquad (11)$$

Note that the model contains $k + 2$ parameters for each item, namely the k discrimination parameters for each trait, plus one difficulty and one guessing parameter. While it is difficult to programme estimation procedures for this model (however, see Coughlan, 1974), it can be used to generate multidimensional data. The parameters of the model can be chosen so that the separate abilities (Θ_k) are correlated to any degree the researcher desires. This is done by multiplying the matrix of chosen a's by a triangular factor of the matrix of desired intercorrelations. Thence the abilities are transformed to possess the desired intercorrelations. Such a method for generating data is employed in the simulation described in the next Chapter.

Previous attempts to assess unidimensionality

Hutten

Hutten (1980) used 25 verbal and quantitative tests from well known aptitude and achievement batteries. Random samples of 1000 were used and the tests ranged from 16 to 85 items. Unidimensionality was assessed on the basis of the ratio of the first and second largest eigenvalues of matrices of tetrachoric correlations. (In an earlier paper Hutten, 1979, used only the percentage of variance accounted for by the first principal component. This was not used in the 1980 study.) Hutten wrote, without citing evidence, that the ratio criterion was "a procedure which has been used extensively for this purpose" and that "high values of the ratio indicate unidimensional tests. Low values suggest multidimensionality" (p. 15).

She reported eigenvalue ratios between 3.33 and 14.96. There were significant negative correlations between the ratio and her fit statistics of both the one- and three-parameter models (r = -.47 and -.54, respectively, p < .05). Hutten reported that the correlations for model-fit with indicators of item-discrimination spread and of guessing were not significant. When calculating relationships with guessing, Hutten included only "those tests for which there were minimally four hard items" (p. 24). She does not explain why "minimally four hard items" were sufficient for the test to be included nor is it clear which tests were excluded. Spearman rank correlations of .00 and -.12 between guessing and the one- and three-parameter fit, respectively, are reported. If all tests minus the four she reports as having no guessing are included, then the correlations are -.45 (p = .02) and -.45 (p = .04). Also, there is an indication that the number of items is negatively related to the index of unidimensionality (r = -.32, p = .06). Further, the correlation between the fit for the one- and three-parameter models is .98, and Hutten reports that 9 of the 25 tests fit the one-parameter model better than the three-parameter model. Surprisingly, no tests fit the three-parameter model better than the one-parameter model.

Ryan

Ryan (1979) claimed that a unidimensional latent trait defined a single variable that all test items reflected and on which all subjects could be ordered. "All items inherently possess the single trait and the continuous latent trait is defined by the collection of items" (p. 1). He then states that a unidimensional trait "simply means that all items measure the same variable. The claim of unidimensionality is therefore quite similar to the classical test theory assumption of parallel or homogeneous test items" (p. 2). Ryan uses alpha as his index of dimensionality. He also used this index in previous work (Helsley, Suber & Ryan, 1977; Ryan & Hamm, 1976), wherein he argued that the Rasch model was robust with respect to violations of the unidimensionality assumption. In the 1979 study, Ryan generated three-parameter data for 300 subjects on a 30-item test. Guessing (0, .25, .5), discrimination means (1, 2, 3), and discrimination standard deviations (.25, .5, .75) were controlled and there were 100 replications for each condition. Using a three-way analysis of variance design, he found that alpha increased as the mean discrimination increased, alpha decreased as the discrimination became more variable, and a major decrease in alpha occurred as guessing was introduced. Despite the questionable use of alpha for indexing unidimensionality, the study highlights the importance of guessing.

Hambleton and Traub

Hambleton (1969) included five items in a 15, 30, and 45 item simulation that were constructed to measure a second ability orthogonal to the first ability. (He does not specify how these items were generated.) In the simulations the items were constrained to have equal discriminations and no guessing. The main aim was to investigate whether the Rasch model was robust with respect to this kind of violation of its assumptions. Hambleton found that the model did not provide a good fit and further, that the fit for the other items was also affected. He wrote that

> in the simulation where the proportion of items measuring a second ability was 33% (5 out of 15), the attempt to fit the data failed so completely that nearly every item was rejected by the model. Since, 67% of the simulated items could be regarded as having been simulated to satisfy the assumptions of the model, this result suggests that rejecting items from a test on the basis of the chi-square test of the goodness of fit of the items to the model is, by itself, a very hazardous way to proceed. (p.101)

Hambleton and Traub (1973) used empirical data to investigate fit to the one- and two-parameter models. These included a verbal test (N = 1319, n = 45), a mathematics test (N = 1319, n = 20), and a verbal test (N = 1208, n = 80). To assess dimensionality they looked at the amount of variation explained by the first factor in a principal axes common factor analysis on the matrix of tetrachoric item correlations. Obtained percentages of 22.1, 31.7, and 20.5 led Hambleton and Traub to comment that there was evidence of a dominant first factor, and to proceed as if each test was unifactorial. They were careful to note that more factors would be needed to generate an acceptable factor solution for each test. They concluded that the more parameters in the test model, the better the fit between the observed and expected distributions. Further, the best fit was obtained for the test that came closest to satisfying their unidimensional assumption (i.e., the Mathematics test with 31.7 percent of variance explained by the first factor).

Lumsden

Lumsden (1957) defined a unidimensional test as "a test in which all items are measuring the same thing" (p. 4) and he claimed that answer pattern methods were the "fundamental definition to which all approaches to the construction of unidimensional tests are at some stage related" (p. 7). Lumsden reviewed many methods and chose to use methods based on factor analysis in his empirical studies.

He divided a group of sixth graders into an item selection group (N = 404) and a validation sample (N = 190). Then four sets of number-series items were sorted from a larger set using rational sorting, centroid component analysis and an investigation of the residuals. Four sets were identified (of 25, 19, 13, and 39 items). By using the amount of variance accounted by the first component (63.5, 76.2, 72.0, and 65.6, respectively), the ratio of first to second eigenvalues (20:1, 38:1, 35:1, 15:1), coefficient alpha (.90, .95, .93; .93), and Horst's modification of alpha (.91, .96, .94, .95), Lumsden concluded that the first three sets seemed to be measuring the same thing, but the fourth set needed more than one factor.

Lumsden concluded that the factor analysis procedure is both theoretically and practically adequate for the construction of unidimensional tests. He reiterated his faith in the use of factor analytic methods to construct unidimensional tests in a recent review of test theory (Lumsden, 1976).

Reckase

Reckase (1972) attempted to derive a multidimensional counterpart of the Rasch model but with simulated multivariate and empirical data, he found that the Rasch model yielded better fit to the data than did his multivariate model. It appears that Reckase abandoned the model.

Koch and Reckase (1979) used 11 separate 50-item multiple-choice tests on 16 samples (average N = 200) and calibrated the items using the one- and/or three-parameter model. They then asked 110 undergraduates to take a tailored test using these items as the item bank and also a conventional 50 item test. Only the conventional test was checked for dimensionality. As there were over 20 factors with eigenvalues greater than one (using a principal component analysis, probably of phi correlations) it was obvious to Koch and Reckase that "the unidimensional assumption of the latent trait models had been violated" (p. 32). They used this violation as the reason for obtaining low reliabilities even though they pointed out that their linking procedures may not have been the best.

Reckase (1979) completed a more systematic study of the effects of violating the one- and three-parameter model assumption of unidimensionality. Reckase used tests of verbal and quantitative ability, and three tests of measurement with sample sizes of 3126, 3126, 208, 176 and 181, respectively. He also simulated data on five other tests (N=1000 each). All tests had 50 items. To simulate the data Reckase used a method developed by Wherry, Naylor, Wherry and Fallis (1965). Starting with a factor structure and known uniquenesses (same value for every item -- hence he had a Rasch model), a set of z scores was randomly generated so as to provide each subject a set of scores for each independent content factor and for each trait uniqueness value. Then the matrix of factors by people matrix was multiplied by the z scores. The scores on the k "traits" for the N people were dichotomized to yield correct and incorrect responses as specified by supplied difficulty indices. Finally, the binary scores were correlated with each other thus producing the inter-item correlation matrix.

For the first data set, Reckase used one factor with loadings of .9 and a normal distribution of difficulties (mean unspecified). For the second set he used loadings of .9 randomly distributed on two factors and a normal distribution of difficulties. Then a nine-factor set was generated that included a dominant first factor with .7 loadings, items randomly distributed to the other eight factors with .6 loadings, and a normal distribution of difficulties. The fourth set also had nine factors but with loadings of .9 and 0 distributed randomly over the factors and a normal distribution of difficulties. The final set was the same as the previous set except that .3 loadings were used rather than .9.

Reckase did not say why he changed from a rectangular to normal distribution of item difficulties nor did he report the mean of the difficulties. He used classical difficulty values and not latent trait difficulty parameters and he did not give any information as to the value used for the uniquenesses. Further, the factor patterns were not checked for linear dependencies between the columns of factor loadings.

All of the data sets were subjected to three factor analyses: a principal component analysis of phi correlations, a principal factor analysis of phi correlations, and a principal component analysis of tetrachoric correlations. Reckase commented that the phi's were included as the tetrachorics did not yield Gramian matrices for the small data sets, even though data based on an analysis of the tetrachorics were provided. Both the principal component and principal factor analyses of phi's, and principal component analysis of tetrachorics overestimated the number of factors when it was known that there was a unidimensional set of items and the percentage of variance explained by the tetrachorics was much greater than for the phi's.

Reckase inspected the rotated and unrotated factor loadings and noticed a relationship between one of the factors and the discrimination values. This is not surprising since Henrysson (1962) had established just such a mathematical relationship must exist. When Reckase plotted the mean three-parameter discrimination estimates, the mean probability of fit from the one-parameter model, and the mean-square fit for each of the models against the first eigenvalues from the tetrachoric factor analysis, he found that as the first eigenvalue increased the average mean-square fit also increased. When the one-parameter chi-square statistic was plotted against the first eigenvalue, then, when a dominant first factor is present, fit is directly related to the size of the first eigenvalue, but there seems to be no relationship otherwise. In his conclusion, Reckase stated that his results demonstrated that if an average three-parameter discrimination estimate of .6 is desired, then the first factor should have an eigenvalue based on phi coefficients of 10 or greater for 50 item tests, or account for at least 20 percent of the total variance. It must be added, however, that in general a large eigenvalue does not necessarily result in high discriminations.

Using the three sets of measurement-test data, Reckase factor analysed the correlations between the various item statistics. The first factor had relatively high loadings on phi and tetrachoric factor loadings, point biserials, and three-parameter discrimination indices, which led Reckase to conclude that this factor is obviously a discrimination factor. The second factor had the various difficulty indices as well as the three-parameter discrimination which appeared to confirm Lord's (1975) finding that difficulty and discrimination may not be independent. The third factor has two high loadings, the three-parameter guessing and one-parameter fit, with opposite signs. This indicated that the goodness of fit to the one-parameter latent trait model may be closely related to variations in guessing. His simulation is restricted to a few special cases, and, as he pointed out, it is hardly realistic. This is so, particularly given that there there was no allowance for correlations between the factors. Swaminathan (1977) after criticising Reckase's study, concluded:

> Clearly, considerable further research is needed in order to establish relationships, if any, between factor models and latent trait models. One possible suggestion for further research is to study the lack of fit through carefully simulated data with varying numbers of factors and to correlate the fit statistic with an index of factorial simplicity (or complexity). This relationship should be studied for the various latent trait models. Another suggestion is to approach the problem of dimensionality from the direction opposite to that employed by Reckase. This would involve generating data to fit the latent trait models and carrying out a factor analysis on the data by varying the parameters involved to study the relationship between the two models. The theory developed by McDonald (1967a) can be employed for this purpose. (p. 389)

McDonald

Certainly McDonald has been a major contributor to the study of unidimensionality. We have already seen the influence of his writings when discussing the principle of local independence (McDonald, 1962b, 1981); his development of nonlinear factor analysis and its relation to binary items (McDonald, 1965a, McDonald & Ahlawat, 1974); the relation between factor analysis and latent trait models (McDonald, 1980b); the development of multidimensional latent trait models (McDonald, 1979); the meaning of coefficient alpha (McDonald, 1970, 1976b); and we note that McDonald has also supplied computer programmes to carry out the various analyses (McDonald, 1965b, 1967d; McDonald & Fraser, 1978; McDonald & Leong, 1974, 1976). His work, which clearly has been heavily relied upon in the above review, was brought together in a recent paper (McDonald, 1981) and serves as a link between previous studies and the simulation to be explained in the next chapter.

McDonald was critical of how the terms unidimensionality, internal consistency, homogeneity, and reliability have been used, particularly their interchangeability. He claimed that the terms homogeneity and internal consistency have been treated in the literature as near synonyms, both lack clear and universally accepted definitions, and too often have been interchanged with unidimensionality, which does have a clear meaning. McDonald argues that we should never say of two unidimensional tests that one is more unidimensional than the other, as the notion of unidimensionality implies that a set of items <u>is</u> or <u>is not</u> unidimensional. Replacing unidimensional with homogeneous or internally consistent, he continues, merely hides the contradiction. Yet the three terms do denote different (though related) concepts. Further, while it is semantically correct to say that a test is or is not unidimensional, it is reasonable to ask whether a set of items approximate a unidimensional set, and to seek an index to assess that closeness.

McDonald's account of unidimensionality begins with a discussion of tests in general (i.e., regardless of how they are scored). He contends that if there is a prevailing conceptualization of the notion that a set of n tests is 1-dimensional, 2-dimensional, ..., or k-dimensional, it is that 1, 2, ..., or k common factors explain the correlations of scores from distinct items or tests in the set, computed over a defined population of examinees. More precisely, a set of n items is considered to be of dimension k if the residuals of the n variables about their regressions on k further (hypothetical) variables, the common factors, are uncorrelated. In particular, he asserts that the Spearman case with just one common factor "gives us a reasonable definition of a <u>unidimensional</u> set of quantitative tests". That is, if the tests fit the single-factor model, we can say that the entire battery is unidimensional. McDonald then discusses some of the problems of using linear common factor analysis, arguing that the nonlinear model is almost certainly necessary when dealing with binary items, and concluding that "it is reasonable to assert that a set of n tests or a set of n binary items is unidimensional if and only if it fits a nonlinear factor model with one common factor" (McDonald, 1981, pp. 14-15). "In principle, therefore, nonlinear factor analysis supplies a general test of the unidimensionality or homogeneity of a set of binary items without the strong and false assumption of linear item characteristic curves that is implicit in the usual attempts to assess dimensionality prior to fitting a latent trait" (McDonald, 1981, p.41).

Commenting about indices of unidimensionality, McDonald writes that unlike factor analysis, latent trait theory has not yet developed to the point where it routinely supplies reasonable decisions as to the dimensionality of binary data sets. Yet, as reviewed above, there have been many indices proposed. McDonald (1981) refers to the use of indices such as percent of variance on the first component as "crude ancillary devices" (p. 16). He suggests that since the two-parameter latent trait model can be fitted by using the analysis of covariance structures then a "possibly crude but apparently satisfactory measure of misfit of the model (can be given) in the familiar form of the magnitudes of residuals" (pp. 16-17).

McDonald highlights a problem we have already referred to, namely that even when we can show (by whatever means) that we have a unidimensional set of items, this does not mean that we can "label" the set. Typically when we have a factor we look closely at the loadings and induce some common abstractive property and thus label the factor. McDonald calls this the empirical heuristic. "There is no logically necessary connection between the mathematical conception of a set of unidimensional variables and the substantive conception of a set of tests or items that measure in common just one property of the examinees in a given population" (McDonald, 1981, p. 24). Moreover, McDonald extends an earlier idea (in McDonald & Mulaik, 1979) that a unidimensional set of variables may not only cease to be unidimensional in the context of further variables, but they may not keep the same factor loadings on their original common factor. Clearly, there are problems in labelling factors.

In concluding, three quotations from McDonald's work seem adequately to sum up this section.

> Indeed, it is somewhat ironic that a prime reason for the invention of these (latent trait) models was a belief that common factor analysis could not be used to explore the dimensionality of binary items, and yet good methods for testing the prior assumption of unidimensionality have not been developed as part of, or ancillary to, these models. (McDonald, 1976a, pp. 226-227)

> Yet we still have to resort to a form of linear factor analysis for a crude test of unidimensionality, we still have reason to doubt the estimation procedures that have been proposed, and we still have no satisfactory statistical criterion for rejecting the model, and no satisfactory criterion for regarding its fit as adequate. (McDonald, 1980c, p. 9)

Regarding indices, there have been many criteria suggested

> as though it is easier to invent new procedures than to justify existing ones. (McDonald, 1967a, p. 11)

Conclusion

Although we have seen that the terms unidimensionality, reliability, internal consistency, and homogeneity have often been used interchangeably, it seems reasonable to recognize the distinctions that have been made between them. Reliability is classically defined as the ratio of true score variance to observed score variance. There are various methods for estimating reliability, such as test-retest, parallel forms, and split-half methods. The internal consistency notion always involves an internal analysis of the variances and covariances of the test items, and depends on only one test administration. Methods of internal consistency at least include split-half coefficients, alpha, and KR-21. It seems that internal consistency is defined primarily in terms of certain methods that have been used to index it. Homogeneity has been used in two major ways. Lord and Novick (1968) and McDonald (1981), for example, used homogeneity as a synonym for unidimensionality, whereas others have used it specifically to refer to the similarity of the item intercorrelations. That is, a perfectly homogeneous test is one in which all the items intercorrelate equally. Cattell (1964, 1978; Cattell and Tsujioka, 1964) has been a principal advocate against aiming for high homogeneity, in this latter sense. He has noted that many authors desire high homogeneity and he commented that aiming for high homogeneity leads to scales in which the same question is rephrased a dozen different ways. He claimed that a test that includes many items that are almost repetitions of each other can cause an essentially "narrow specific" to be blown up into a bloated specific or pseudo-general factor. In Cattell's colourful words:

> the bloated specific will then 'sit on top' of the true general personality factor as firmly as a barnacle on a rock and its false variance will be inextricably included in every attempted prediction from the general personality factor. Moreover, the 'crime' will be as hard to detect, without a skillful factor analysis, as it is insidious in its effects, for the intense pursuit of homogeneity has ended in a systematically biased measure. (Cattell & Tsujioka, 1964, p. 8)

Unidimensionality can be defined as the existence of one latent trait underlying the data. A unidimensional test is not necessarily reliable, internally consistent, or homogeneous.

Thus only reliability and unidimensionality have clear and distinct meanings. Whether internal consistency and homogeneity are meaningful terms in describing attributes of items and/or tests remains questionable.

There have been many indices proposed as decision criteria for determining unidimensionality. Some indices must fail, others seem more appropriate in specific conditions, while others look promising. None of the attempts to investigate unidimensionality has provided clear decision criteria for determining it. What seems needed is a Monte Carlo simulation to assess the various indices under known conditions. Such a simulation is outlined in the next chapter.

Chapter III

Method

In the previous chapter it was suggested that a Monte Carlo study be carried out to investigate the effectiveness of the many indices of unidimensionality. In a Monte Carlo procedure random numbers are used to construct artificial data with prescribed characteristics. For the purposes of this study there exists an infinite number of data sets that could be generated by selecting different levels of the parameters. Obviously, decisions needed to be made about the parameters to be used and this choice had to be realistic if the results were to be of practical relevance. Also computer availability, size, and time were limiting factors.

Methodology

The (k + 2)-parameter multivariate logistic latent trait model (11) was used to generate data (see pp. 67-68), as this model made it possible to control the critical variables. The univariate version of this model has often been used (Lord, 1980) and, of most importance, the model allowed control over the number of factors underlying the data.

Data were generated according to the specifications listed below and each index listed in Appendix B was calculated. Appendix B contains a listing of each index, the programme that calculated the index, its acronym, and comments on any calculation idiosyncracies.

Number of items

The number of items was 15. This is a short test compared to many nationally normed and teacher-made tests, but it is not uncommon for test developers to use approximately 15 items to measure a single ability and longer tests usually contain short subtests about this length. Also computer-adaptive tests rarely include more than 15 items. It would have been desirable to vary the length of the test and, in particular, to include longer tests (although see Hutten, 1980, p. 70 above). Because of computer resources this was not possible. It was judged more efficient not to have tests of varying length, but rather to assess the indices with 15 items. Those that seem reasonable could be included in later simulations to determine the effect of varying test length.

Number of factors and discrimination

A most critical test is whether the indices can distinguish between data that have one or two factors, and between one and many factors. Hence it was decided to use one, two, and five factors. The choice of five factors was arbitrary but it is not unusual with binary data to get factors numbering up to one-third the number of items. Certainly, any useful index should be able to differentiate correctly between one and five factors.

For the one factor case the choice of discrimination values was straightforward. Two sets of values were chosen. The first set were .6, 1.0, and 1.4 with five items at each value (see pp. 56-57 for justification of this choice). The second set used values all of 1, as equal discrimination is an assumption of the Rasch model.

There is a relationship between the factor loadings and discrimination values (see pp. 67-68), yet the choice of weights for the two and five factor cases is not obvious. First, a model is to be employed in which the latent traits are correlated, as it is not typical to get orthogonal factors in tests of ability. Further, a basic postulate is that ability tests are generally positively correlated. Thurstone (1934, 1935) argued that unitary abilities that are psychologically meaningful have (oblique) positive manifold. That is, not only is there simple structure, but negative cell entries in the factor pattern are excluded. Thurstone (1934) claimed that

> the reference abilities by which the negative factorial coordinates disappear and by which at the same time a large number of zero coordinates appear in the factorial matrix and by which the factorial coordinates become psychologically meaningful consitute unitary abilities. (p. 130)

We have seen also (p. 25) that by rotating the factors we can achieve different amounts of first factor variance. If principal axes are used, then the maximum variance is explained by the first factor. A configuration was decided upon that took the above considerations into account and seemed to reflect likely empirical cases.

Discrimination values of 1 were formed into a simple structure pattern and multiplied by a triangular matrix based on intercorrelations of .1 between all factors for one set of cases, and on intercorrelations of .5 for the other set. Thus we have simple structure, positive manifold, oblique factors, and there are no linear dependencies between the factors. Principal axes were calculated to check the amount of variance explained by the factors. (Discrimination values of one approximately represent biserials of .707, see Lord, 1980.) Table 5 presents the amount of variance explained by the principal components and Table 6 presents one example -- the five factor case with .5 intercorrelations. The initial factor loadings were post-multiplied by the triangular decomposition of factor correlations to give the actual factor loadings to be used in the simulation. Note that the first factor variance decreases as the number of factors increases, and as the intercorrelations decrease. Obviously many patterns could be chosen; these values seem reasonable.

Difficulty

Two sets of difficulty level values were chosen (see pp. 57-58). One set was (-2, -1, 0, 1, 2) and the other reflected a narrower range (-1, -.5, 0, .5, 1).

Guessing

Given five-option multiple-choice items, it is reasonable to assume that c is bounded on the interval (0, .2) (see p. 58). Three levels of guessing were chosen: all .0, reflecting no guessing, all .2, and a mixture of .0, .1, and .2.

Ability distribution and number of examinees

The ability parameters were drawn randomly from a unit normal distribution. A sample size of 500 subjects was chosen as a compromise between the desire for a larger sample and available

Table 5

The amount of variance explained by the components according to the true number of factors and the intercorrelation between the factors (or discrimination)

True No. of factors	Intercorrelation	Percentage of variance explained by each principal component					
		1	2	3	4	5	Remainder
5	.1	17	12	12	12	12	35
5	.5	27	10	10	9	6	38
2	.1	31	25	(3)	(3)	(3)	35
2	.5	39	14	(4)	(4)	(4)	35
1	(mixed)	51	(5)	(5)	(5)	(5)	29
1	(all 1's)	53	(3)	(3)	(3)	(3)	35

Table 6

An example of the initial discrimination values, the triangular decomposition of factor intercorrelations, and the actual discrimination values used in the simulation – The five factor model with .5 intercorrelations

	Initial Discrimination Values					Triangular Decomposition of factor correlations					Final Discrimination Values				
No.	1	2	3	4	5						1	2	3	4	5
1	1	0	0	0	0	1.00					1.00	.50	.50	.50	.50
1	1	0	0	0	0	.50	.87				1.00	.50	.50	.50	.50
3	1	0	0	0	0	.50	.29	.82			1.00	.50	.50	.50	.50
4	0	1	0	0	0	.50	.29	.20	.79		.50	.87	.29	.29	.29
5	0	1	0	0	0	.50	.29	.20	.16	.78	.50	.87	.29	.29	.29
6	0	1	0	0	0						.50	.87	.29	.29	.29
7	0	0	1	0	0						.50	.29	.82	.20	.20
8	0	0	1	0	0						.50	.29	.82	.20	.20
9	0	0	1	0	0						.50	.29	.82	.20	.20
10	0	0	0	1	0						.50	.29	.20	.79	.16
11	0	0	0	1	0						.50	.29	.20	.79	.16
12	0	0	0	1	0						.50	.29	.20	.79	.16
13	0	0	0	0	1						.50	.29	.20	.16	.77
14	0	0	0	0	1						.50	.29	.20	.16	.77
15	0	0	0	0	1						.50	.29	.20	.16	.77

computer resources. Experience with latent trait models indicated that parameter estimates are unstable with samples up to 300. At 500 the estimates seem to become reasonably stable (see Swaminathan & Gifford, 1979). To further ensure stable results each set of parameters was used to obtain 24 sets of data.

An alternative method for determining sample size would have been to use power analysis. Unfortunately because there were no previous studies of unidimensionality similar to this present one, there was no way of determining power. However, assuming univariate analysis, an approximation can be obtained. Given conventional power of .80 (Cohen, 1969, pp. 51-54), a large effect size (f = .40, Cohen, 1969, p. 279), and a significance level of .01, then using 24 replicates of 500 subjects is sufficiently powerful to detect significant differences (see Cohen, 1969, for details).

Models

There are 36 models (2 [difficulty] x 3 [guessing] x 6 [1 factor, discrimination mixed; 1 factor, discrimination all 1; 2 factor, intercorrelation .1; 2 factor, intercorrelation .5; 5 factor, intercorrelation .1; 5 factor, intercorrelation .5]).

The programmes

Six separate Fortran computer programmes were used. All programmes were run on the OISE DEC-10 computer. Appendix B details which programme was used to calculate each index. Two were written specifically (Hattie, 1980b, 1980c) and the other four were obtained from their authors (see below). The IMSL library (1980) was used extensively in UNIDIM and all indices (except those based on answer patterns) were checked at least twice with alternative programmes. Because there were no alternative programmes available, the answer pattern methods were checked by comparing results from the programme with examples given in the literature (e.g., Nishisato, 1975; Raju, 1980).

For the indices based on factor analysis, where appropriate, principal component, maximum likelihood, and least squares analyses were used on phi and tetrachoric correlation matrices. Because of machine incompatibilities and insufficient core it was not possible to get the maximum likelihood FADIV (Andersson, Christoffersson, & Muthen, 1974) programme working on the DEC-10, so their two-stage least squares method was used. This two-stage programme was checked using data in their manual and comparing an analysis of Rotter's (1966) data with output kindly supplied by Muthen (personal communication, 1980). Wright, Mead and Bell's (1979) BICAL programme was used for the Rasch estimates. No changes were made to the programme and data were checked using two versions of BICAL and cross-checking with the earlier MESAMAX programme. For the two-parameter latent trait model, Fraser's (1980) NOHARM programme was used. This programme has been extensively used at OISE and reports on its success in recovering known parameters are given in McDonald and Fraser (in prep.). Etezadi's (1981) NONLIN programme was used to perform the nonlinear factor analysis. Linear, quadratic, and cubic terms on one factor were specified. The least squares method was used and to hasten convergence the true ability estimates used to generate the data in the UNIDIM programme were used as initial estimates. For

Lord's index (see pp. 55-56) a special programme was written. This was checked using the SPSS and the IMSL subroutines.

The IMSL GGUBFS (uniform distribution) and GGNQF (normal distribution) were used to generate random numbers. These were successfully checked for normality (or rectangularity), sequential independence, and for runs and gaps (see Schrage, 1979, for further details). The seed was a lottery number bought at a local store (113167).

All programmes were run on the OISE DEC-10 computer. Computer processing time for each programme is presented in Chapter IV.

Method of analysis

Given that there are so many indices, a four-stage analysis was conducted. Only the indices that met each criterion were retained for the subsequent analyses.

First, the intercorrelations between subsets of the indices were calculated. If a set of indices intercorrelated greater than .85, only one was retained. The one retained was that which has the best rationale (in Table 1) or, if the rationales were similar, that which was easiest to calculate.

The second criterion was that the mean indices for the one factor data should be either all larger, or all smaller than the mean indices on the two and five factor data. This seems a minimum demand if the index is to serve as a decision criterion for determining unidimensionality.

Thirdly, a 3-way multivariate analysis of variance (MANOVA) was calculated. A multivariate rather than a univariate analysis takes into account the correlations among the variables. As there are ceiling effects, the data were transformed using a logarithmic transformation prior to analysis (see Mosteller & Tukey, 1977). Finn's (1978) MULTIVARIANCE programme was used.

In the MANOVA, the one factor (mixed discrimination and all discriminations equal to 1), the two, and the five factors (intercorrelations .1 and .5) were combined into one (MANOVA) factor with six levels, and the following post-hoc Scheffe contrasts were calculated: 1) That there are no differences between the means on the one factor (mixed) and the one factor (all 1); 2) That there are no differences between the means on the one factor (mixed + all 1) and the two factor (intercorrelation among factors of .1); 3) That there are no differences between the means on the one factor (mixed + all 1) and the two factor (,5); 4) That there are no differences between the means on the one factor (mixed + all 1) and the five factor (.1); and 5) That there are no differences between the means on the one factor (mixed + all 1) and the five factor (,5). Also the remaining indices should meet the second criterion within each of the guessing and difficulty levels. To set confidence bounds on the contrasts and to assess significance, a stringent level of .001 was used.

The final criterion relates to the number of times the values of the indices from the one factor data overlapped with values from the two and five factor data. The minimum and maximum values of the indices from the one factor simulations were determined and the number of occasions when data from the two factor (.1 and .5, separately) and five factor (.1 and .5, separately) simulations were outside these limits was recorded.

Chapter IV

Results and Discussion

The results are presented in five sections. After comments on general issues, there are four sections corresponding to the four criteria outlined in the previous chapter.

General Comments

In 290 out of the 864 simulations the matrix of tetrachorics was non-Gramian. This is a problem, though not an unexpected one, for matrices of tetrachoric correlations. Table 7 presents a cross-tabulation of the occasions by number of factors, discrimination (or intercorrelation), difficulty, and guessing. In all but one case where there was a wide range of difficulties and zero guessing, the matrix was non-Gramian. Most cases with wide difficulty and mixed guessing also led to non-positive definite matrices.

The other problem related to estimating the chi-square hypothesizing two factors (for CH2-CH1 — see Appendix B for a glossary of acronyms). On 105 occasions convergence did not occur after 500 iterations. Consequently missing values were assumed in any analysis involving this variable. There did not appear to be any systematic pattern as to the conditions when this lack of convergence occurred.

On a timesharing computer (such as a DEC-10) processing time (CPU), rather than elapsed time, is most meaningful. Table 8 presents the average time for the 24 replications over all 36 models. There

Table 7

The number of non-Gramian tetrachoric matrices cross-tabulated with number of factors, distribution of discrimination (or correlation), difficulty and guessing

		Number of factors					
		1		2		5	
Difficulty	Guessing	Mix	All 1	.1	.5	.1	.5
	.0	2	0	1	0	1	0
-1 1	.2	0	0	0	0	0	0
	Mix	0	0	0	0	0	0
	.0	24	24	24	24	23	24
-2 2	.2	2	0	4	1	5	0
	Mix	24	24	20	23	17	23

Table 8

Average CPU time (in minutes) and standard deviation for the 24 replications over the 36 models

Programme	Mean	s.d.
UNIDIM	94.69	48.27
NONLIN	65.84	20.85
BICAL	24.62	2.45
FADIV	49.74	36.66
NOHARM	23.58	3.62
LORD	22.83	3.73

was a negative skew in time for UNIDIM, NONLIN, and FADIV. In these runs the time increased with the number of factors, whereas run times for NOHARM, BICAL, and LORD did not vary markedly with respect to the number of factors.

The average test score was 8.24 (s.d. = 9.38). There were no statistically significant associations between differences in the means and differences in level of difficulty, or in the number of factors (6 levels including discrimination and intercorrelation). Also cell interactions among these factors and the factor of guessing were not statistically significant (Table 9). The mean for guessing fixed at .0 was lower than the means when guessing was mixed or all .2 guessing (7.48, 8.24, and 8.99, respectively), as was expected.

Correlations

The obvious conclusion that can be drawn from a study of the intercorrelations is that many indices provide much the same information.

Of the indices based on answer patterns, LOEV and CONSIS were highly correlated; thus only one needs to be retained. REPB seems to be measuring something different (Table 10).

Except for KR-21, the various alpha and intercorrelation indices all intercorrelated very highly (Table 11), so only mean-alpha and KR-21 were retained.

Table 9

Degrees of freedom, F's, and significance levels from a MANOVA of the test means

Source	df	F	Sig.
Difficulty	1	1.70	.192
Guessing	2	8895.46	<.001
No. Factors	5	3.93	.002
D x G	2	1.18	.307
D x F	5	1.27	.276
G x F	10	1.42	.166
D x G x F	10	.75	.673
Within	828	.02	

Table 10

Means, standard deviations and correlations of indices based on answer pattern methods

Index	1	2	3
1 LOEV	100		
2 REPB	62	100	
3 CONSIS	98	53	100
Mean	.27	.85	.22
s.d.	.12	.03	.09

Note. Decimal points are omitted in the correlation matrices presented in Tables 10-16.

Table 11

Means, standard deviations and correlations of

indices based on reliability and intercorrelations

Index	1	2	3	4	5	6	7	8	9	10	Mean	s.d.
1 ALPHA	100										.70	.12
2 MNALPHA	94	100									.16	.07
3 KR-21	93	88	100								.59	.19
4 HORST	100	95	94	100							.68	.12
5 MNPHI	94	100	89	95	100						.16	.07
6 PHI-05	95	92	89	95	92	100					74.42	28.19
7 PHI-01	95	95	89	95	98	99	100				65.53	30.44
8 MNTET	91	94	76	90	93	89	90	100			.30	.12
9 TET-05	95	92	87	95	92	100	98	91	100		75.21	28.54
10 TET-01	95	95	87	95	95	99	100	92	99	100	66.69	30.96
11 RAJU	99	91	88	98	91	94	92	92	94	93	.81	.11
12 KAISER-phi	94	100	88	95	100	93	96	94	92	95	.17	.07
13 KAISER-tet	90	92	73	89	91	99	89	100	90	92	.31	.12
14 KR-20B4	97	91	83	96	91	92	90	96	93	92	.64	.11
15 KR-20	96	91	82	95	90	91	90	96	92	91	.64	.11
16 PBIS	98	99	91	98	99	95	97	94	95	97	.33	.10
17 THETA-phi	99	94	92	99	94	96	95	91	96	95	.72	.11
18 THETA-tet	95	87	78	94	87	91	89	93	92	90	.85	.08
19 OMEGA-phi	98	90	91	98	91	93	92	88	93	92	.82	.09
20 OMEGA-tet	97	87	91	97	88	90	90	90	92	90	.90	.06
21 TETBND	-97	-90	-85	-94	-91	-99	-97	-89	-99	-98	29.49	27.29

	11	12	13	14	15	16	17	18	19	20	21
11 RAJU	100										
12 KAISER-phi	92	100									
13 KAISER-tet	91	92	100								
14 KR-20B4	98	92	95	100							
15 KR-20	98	91	95	100	100						
16 PBIS	95	99	93	95	94	100					
17 THETA-phi	99	95	91	97	96	97	100				
18 THETA-tet	97	88	93	97	97	91	96	100			
19 OMEGA-phi	97	91	87	95	94	95	98	94	100		
20 OMEGA-tet	99	89	90	96	96	93	98	99	96	100	
21 TETBND	-93	-92	-89	-91	-91	-92	-95	-92	-92	-92	100

For the Green, Lissitz and Mulaik indices the principal component indices of r and the maximum likelihood estimates of u are highly related (Table 12). Only GLM-r-pc-phi, GLM-r-ml-phi, GLM-r-ml-tet, and GLM-u-ml-phi were retained. As expected, most GLM indices were greater than unity.

The least squares estimates of the DUBOIS indices were highly related to the principal component indices, hence the least squares indices were discarded (Table 13).

The amount of variance explained by the first component and the amount explained by the first factor were highly intercorrelated (Table 14), although the means differ (from 16.82% to 35.74%). It seemed necessary to retain only one variance index (VARIANCE-pc-phi). Also retained were KELLEY-phi, both EDIFF's, EGIN>1-phi, and E1/E2-phi.

From the chi-squares and residuals, only CHI-phi, RES-pc-phi, RES-pc-tet, RES-ml-phi, RES-ls-tet, and T&L-phi had to be retained (Table 15). Again there were marked differences between the means of the sum of residuals (7.96 to 33.00). The chi-square means for the solutions based on the phi and tetrachoric correlations were markedly different (200.91 and 796.54, respectively), yet the values correlated .87, and the sum of residuals correlated .95. There were 90 degrees of freedom for both chi-squares. The two Tucker-Lewis indices intercorrelated highly, yet they had lower correlation with the chi-squares and sum of residuals. The indices retained from Table 15 were CHI-phi, RES-pc-phi, RES-pc-tet, RES-ls-tet, and T&L-phi.

Table 12

Means, standard deviations and correlations of indices suggested by Green et al.

Index	1	2	3	4	5	6
1 GLM-r-pc-phi	100					
2 GLM-r-pc-tet	99	100				
3 GLM-r-ml-phi	86	86	100			
4 GLM-r-ml-tet	62	63	79	100		
5 GLM-u-ml-phi	68	66	43	18	100	
6 GLM-u-ml-tet	71	71	47	22	97	100
Mean	1.96	1.67	1.39	1.27	1.28	1.10
s.d.	.26	.28	.28	.31	.31	.14

Table 13

Means, standard deviations and correlations of DuBois' indices

Index	1	2	3	4	5	6
1 DUBOIS-pc-phi	100					
2 DUBOIS-pc-tet	55	100				
3 DUBOIS-ml-phi	89	51	100			
4 DUBOIS-ml-tet	76	75	86	100		
5 DUBOIS-ls-phi	98	55	95	80	100	
6 DUBOIS-ls-tet	60	99	58	80	61	100
Mean	.97	.93	.95	.92	.97	.93
s.d.	.04	.10	.06	.11	.05	.10

Table 14

Means, standard deviations and correlations of indices based on component and factor analysis

Index	1	2	3	4	5	6	7	8	9	Mean	s.d.
1 VARIANCE-pc-phi	100									22.11	6.17
2 VARIANCE-pc-tet	92	100								35.74	11.14
3 VARIANCE-ml-phi	100	93	100							16.83	6.53
4 VARIANCE-ml-tet	99	100	99	100						29.60	11.24
5 VARIANCE-ls-phi	100	93	100	99	100					16.82	6.55
6 VARIANCE-ls-tet	87	98	87	100	87	100				32.85	12.17
7 E1/E2-phi	93	82	93	94	93	75	100			2.58	1.07
8 E1/E2-tet	92	86	92	93	92	81	96	100		3.80	2.01
9 KELLEY-phi	-38	-04	-37	-41	-37	08	-40	-28	100	.37	.06
10 KELLEY-tet	-12	-24	-12	-17	-12	37	-20	-06	94	.60	.15
11 EDIFF-phi	29	19	29	29	29	14	41	34	-28	19.07	29.66
12 EDIFF-tet	39	34	39	39	39	31	47	51	-18	32.54	52.03
13 EGIN>1-phi	-84	-78	-84	-80	-84	-73	-68	-69	40	3.62	1.29
14 EGIN>1-tet	-82	-67	-82	-83	-82	-60	-71	-68	54	3.65	1.35

	10	11	12	13	14
10 KELLEY-tet	100				
11 EDIFF-phi	-19	100			
12 EDIFF-tet	-10	-38	100		
13 EGIN>1-phi	14	-14	-27	100	
14 EGIN>1-tet	34	-20	-30	89	100

Table 15

Means, standard deviations and correlations of indices based on chi-squares and residuals

Index	1	2	3	4	5	6	7	8	9	10	11	12	Mean	s.d.
1 CHI-phi	100												200.91	110.51
2 CHI-tet	87	100											796.54	385.78
3 CH2-CH1-phi	88	72	100										96.59	93.16
4 CH2-CH1-tet	83	76	94	100									305.93	295.33
5 RES-pc-phi	14	24	14	24	100								7.96	.65
6 RES-pc-tet	08	22	05	21	68	100							8.00	1.27
7 RES-ml-phi	96	82	93	88	09	04	100						20.19	1.49
8 RES-ml-tet	92	91	88	87	20	18	95	100					24.75	2.52
9 RES-ls-phi	94	80	97	91	09	04	99	94	100				20.35	1.88
10 RES-ls-tet	18	84	26	93	-08	-09	28	96	25	100			33.00	13.18
11 T&L-phi	-75	-74	-50	-50	-38	-33	-65	-68	-61	02	100		.79	.22
12 T&L-tet	-65	-74	-41	-45	-43	-45	-56	-63	-51	-52	95	100	.62	.24

When using BICAL, it did not matter whether subjects were omitted or not when calculating the mean, standard deviation, or alpha (Table 16). There were 1.88% who had either zero or perfect scores and 1.44% who were "misfitting". These persons were not included in further analysis. A MANOVA of the number of misfitting persons across levels of difficulty, guessing, and factors indicated that there were more misfitting persons as guessing increased (Tables 17 and 18). The other fit indices from BICAL were not highly related to each other. The indices retained were MEAN, SD, MISFIT, ERROR, BETWEEN, TOTAL, WITHIN, and DISCR.

Indices based on the sum of residuals rather than the number of residuals greater than .01 from the NONLIN, NOHARM, and FADIV methods were retained. These sums of residuals correlated highly with the number of residuals greater than .01 (Table 19). The indices from NOHARM with guessing based on the lower 10% correlated highly with the estimates based on guessing fixed at .0 and .2. Only the sum of residuals based on guessing fixed at .0 and .2 were kept. Lord's index was also retained.

Thus only 35 of the original 87 indices were left at this point.

Means

The means on the one factor (mixed and all one discrimination), two factor (.1 and .5 intercorrelation), and five factor (.1 and .5 intercorrelation) for the remaining 35 indices are presented in Table 20. Many of these indices fail the second criterion: That the means on the one factor data should be either all larger or all smaller than

Table 16

Means, standard deviations and correlations of indices based on the one-parameter model

Index	1	2	3	4	5	6	7	8	Mean	s.d.
1 ZEROSC	100								2.39	4.27
2 PERFSC	29	100							7.02	6.91
3 MEANB4	-52	08	100						.26	.23
4 SDB4	52	23	-57	100					.96	.28
5 MISFIT	-37	-47	28	-23	100				7.22	4.02
6 MEAN	-52	06	100	-56	31	100			.27	.24
7 S.D.	51	20	-57	100	-18	-56	100		.99	.28
8 ERROR	06	33	19	27	-17	19	25	100	.01	.005
9 BETWEEN	-11	20	37	12	03	38	10	86	.81	.55
10 TOTAL	-29	-00	12	-56	-05	09	-56	-54	-.92	.12
11 WITHIN	07	37	-06	-34	-49	-09	-37	-24	.94	.01
12 DISCR	-11	-52	-17	38	52	-15	41	-04	1.07	.01

	9	10	11	12
9 BETWEEN	100			
10 TOTAL	-57	100		
11 WITHIN	-35	69	100	
12 DISCR	01	-42	-87	100

Note. There were no zero or perfect items.

Table 17

Degrees of freedom, F's, and significance levels from a MANOVA of the number of misfitting persons

Source	df	F	Sig.
Difficulty	1	1.70	.192
Guessing	2	8895.46	<.001
Factors	5	3.93	.002
D x G	2	1.18	.307
D x F	5	1.27	.276
G x F	10	1.42	.166
D x G x F	10	.75	.673
Within	828	.018	

Table 18

Mean number of misfitting persons cross-tabulated with difficulty, guessing, number of factors, and discrimination (or correlation)

		No. of factors					
		1		2		5	
Difficulty	Guessing	Mix	All.1	.1	.5	.1	.5
-1 1	.0	2.42	3.71	8.13	4.33	4.38	3.38
	.2	4.42	5.83	6.08	4.42	4.63	5.21
	Mix	3.71	3.04	7.92	3.63	5.50	3.08
-2 2	.0	6.62	7.83	13.42	6.83	9.25	8.25
	.2	11.63	12.08	11.67	12.00	9.46	11.88
	Mix	7.83	9.13	11.96	8.54	9.38	8.46

Table 19

Means, standard deviations and correlations of indices based on NONLIN, FADIV, NOHARM, and Lord

Index	1	2	3	4	5	6	7	8	Mean	s.d.
1 NONRES	100								.016	.011
2 NON>01	86	100							20.50	10.15
3 NH1ORES	45	39	100						.013	.014
4 NH1O>01	48	53	91	100					25.18	18.36
5 NHORES	45	39	100	91	100				.013	.014
6 NHO>01	48	53	91	100	91	100			25.28	18.32
7 NH2RES	20	15	71	60	71	59	100		.021	.016
8 NH2>01	30	34	75	79	75	79	86	100	33.33	16.42
9 FADRES	44	38	100	91	100	91	71	74	.013	.015
10 FAD>01	44	53	90	98	90	98	58	78	25.31	18.55
11 LORD	16	-01	13	09	13	09	-21	-19	44.26	7.37

	9	10	11
9 FADRES	100		
10 FAD>01	90	100	
11 LORD	14	10	100

Table 20

Means of the 35 indices retained after criterion 1 broken down by number of factors and discrimination (or correlation)

	1 factor		2 factor		5 factor	
Index	Mix	All	.1	.5	.1	.5
1 LOEV	.32	.35	.20	.34	.11	.31
2 REPB	.860	.868	.844	.864	.829	.858
3 MNALPHA	.184	.200	.116	.199	.065	.182
4 KR-21	.67	.68	.51	.69	.32	.67
5 GLM-r-pc-phi	1.54	1.47	4.35	1.47	5.56	1.61
6 GLM-r-ml-phi	1.16	1.16	2.27	1.14	2.17	1.22
7 GLM-r-ml-tet	1.12	1.12	1.61	1.10	1.79	1.18
8 GLM-u-ml-phi	1.007	1.007	1.217	1.006	1.086	1.008
9 DUBOIS-pc-phi	.982	.989	.943	.988	.944	.982
10 DUBOIS-pc-tet	.92	.95	.90	.96	.88	.96
11 DUBOIS-ml-phi	.98	.99	.90	.99	.89	.98
12 DUBOIS-ml-tet	.96	.98	.84	.97	.82	.96
13 VARIANCE-pc-phi	24.86	25.93	18.24	25.82	13.24	24.55
14 KELLEY-phi	.367	.367	.374	.363	.372	.363
15 EDIFF-phi	27.22	22.55	.94	16.29	11.58	35.82
16 EDIFF-tet *	55.51	58.31	.99	27.95	7.02	45.44
17 EGIN>01	3.21	2.76	3.60	2.97	5.52	3.63
18 CHI-phi *	140.10	130.71	348.68	155.20	281.97	148.80
19 RES-pc-phi *	7.74	7.70	8.53	7.71	8.32	7.78
20 RES-pc-tet *	7.65	7.62	8.64	7.65	8.78	7.67
21 RES-ls-tet	33.96	34.38	34.90	34.40	28.08	32.29
22 T&L-phi *	.94	.95	.59	.93	.40	.93
23 MEAN	.29	.28	.29	.28	.23	.27
24 S.D.	1.09	1.16	.84	1.15	.60	1.08
25 MISFIT	6.11	6.94	9.86	6.63	7.10	6.71
26 ERROR	.013	.007	.004	.007	.002	.008
27 BETWEEN	1.39	.78	.58	.80	.24	1.01
28 TOTAL	-.988	-.969	-.847	-.953	-.780	-.973
29 WITHIN *	.936	.938	.947	.940	.952	.942
30 DISCR	1.072	1.069	1.069	1.070	1.068	1.069
31 NONRES *	.010	.009	.015	.013	.033	.014
32 NHORES *	.005	.004	.033	.006	.021	.007
33 NH2RES	.0174	.0168	.0351	.0169	.0214	.0153
34 FADRES *	.005	.004	.034	.006	.021	.007
35 LORD	42.84	40.62	48.62	40.40	51.50	41.28

Note. * indicates the indices that passed the second criterion. (See p. 111 for more details.)

the means on the two and five factor data. Only 9 indices (starred) met this criterion. When difficulty and guessing were taken into account, then CHI-ml-phi, RES-pc-tet, RES-ls-tet, and T&L-phi were not able to meet the second criterion when there was a wide range of difficulty values (-2, 2). After these indices were excluded only five remained: EDIFF-pc-tet, WITHIN, NLRES, NHORES, FADRES.

It should be noted that all indices deleted so far also failed the second criterion within at least one level of guessing and difficulty.

MANOVA

A three-way MANOVA was calculated on the values of the five remaining indices. The mean-squares, degrees of freedom (df), and F-statistics are presented in Table 21. Also reported are the indices for which an effect was not significant. Specific post-hoc comparisons (see pp. 93) resulted only in EDIFF being excluded, as hypothesis 5 was not significant (Table 22).

Overlap

The final criterion related to the amount of overlap between the one factor and the two and five factor cases (treating correlations of .1 and .5 separately). On no occasion was an index lower than the minimum value for a one factor case, hence only the percentages exceeding the upper bound are reported (Table 23). WITHIN overlaps substantially and can thus be excluded. The NONRES is not as distinct as FADRES or NHORES. Box-plots for WITHIN (Figure 1), NONLIN (Figure 2), FADRES (Figure 3), and NHORES (Figure 4) graphically display the

Table 21

Degrees of freedom, F's, and significance levels

from a MANOVA of EDIFF, WITHIN, NONRES, NHORES, and FADRES

Source	df	F	Variable Not Sig.	Sig. Level
Difficulty	5	3121.91		
Guessing	10	36.81	EDIFF FADRES HMORES	.282 .273 .040
Factors	25	321.21		
D x G	10	53.27	EDIFF	.235
D x F	25	38.97		
G x F	50	25.46		
D x G x F	50	5.00	EDIFF HMORES	.015 .001

Table 22

Specific Scheffe contrasts to test the five

hypotheses regarding factor means

	Contrast				
Index	1	2	3	4	5
1 EDIFF-phi	-.49	74.83	11.51	42.43	3.65
2 WITHIN	-1507.12	-7267.45	-1877.82	-11280.80	-3606.03
3 NLRES	14.78	-105.31	-72.81	-293.43	-97.77
4 NHORES	25.04	-673.64	-111.56	-526.02	-142.51
5 FADRES	22.21	-646.20	-107.14	-502.84	-136.32

Table 23

Percentage exceeding the upper bound of the

1-factor data

No. of factors

	2		5	
Index	.1	.5	.1	.5
1 BICAL	22	0	49	3
2 NONRES	26	22	89	34
3 NHORES	100	24	98	34
4 FADRES	100	24	98	35

Figure 1

Box-plot of Bical WITHIN

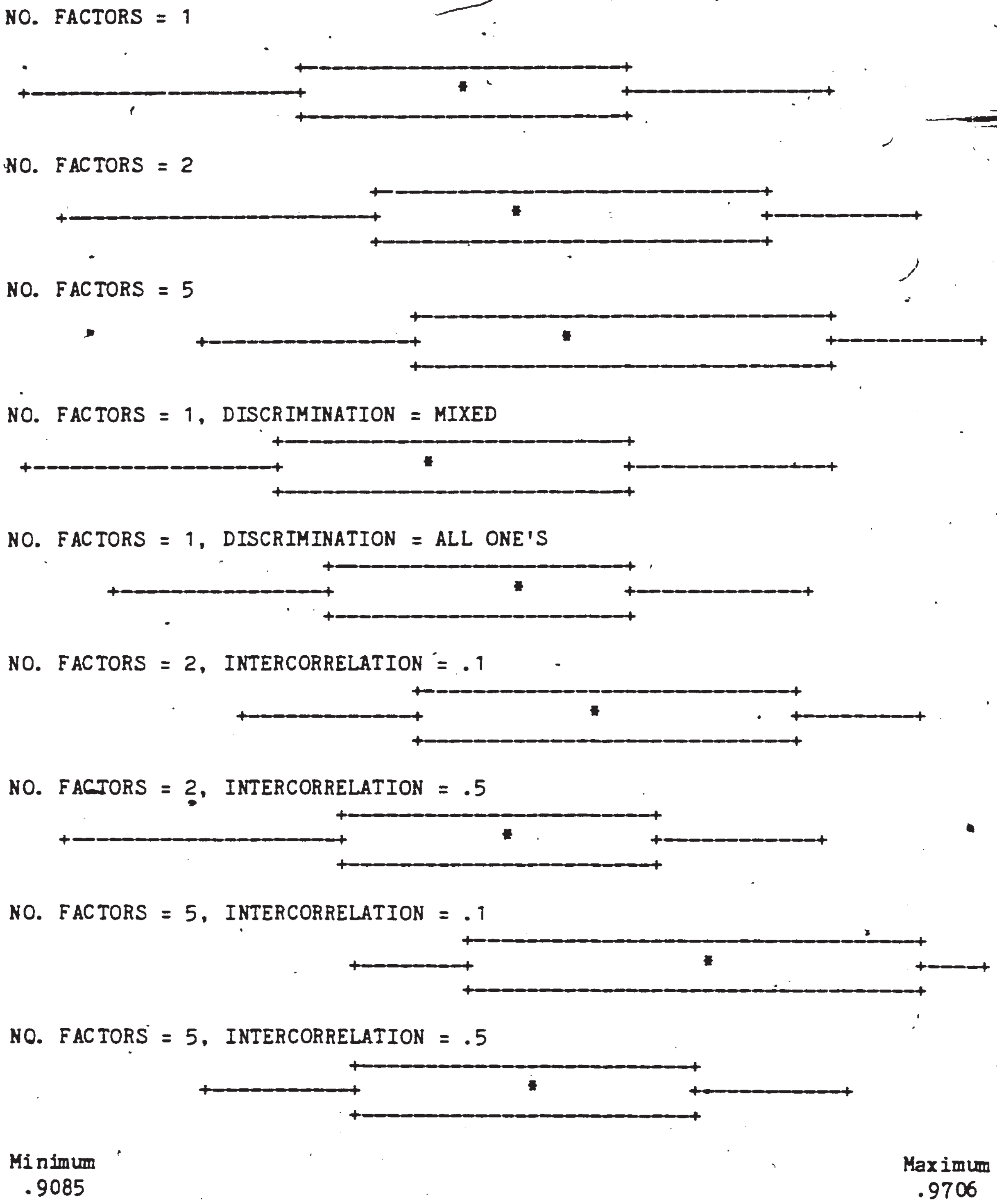

Figure 2

Box-plot of NONRES

```
NO. FACTORS = 1
     +-----+
 +----+ * +-------+
     +-----+

NO. FACTORS = 2
        +--------+
   +-----+   *   +---------+ 0    00 0   X                                   X
        +--------+
                             2    3

NO. FACTORS = 5
            +---------------------+
    +--------+       *           +---------------------------+  00 000  0 0    0
            +---------------------+
                                                                 3 2       3    2

NO. FACTORS = 1, DISCRIMINATION = MIXED
     +-----+
  +---+  * +-------+
     +-----+

NO. FACTORS = 1, DISCRIMINATION = ALL ONE'S
     +-----+
 +----+  * +------+
     +-----+

NO. FACTORS = 2, INTERCORRELATION = .1
          +------+
    +-----+  *  +--------+ 00 0   XX  X   X                                   X
          +------+
                             2    22

NO. FACTORS = 2, INTERCORRELATION = .5
        +---------+
    +---+   *    +--------+
        +---------+

NO. FACTORS = 5, INTERCORRELATION = .1
                        +-------------------------+
          +--------------+       *                +-----------------------------+
                        +-------------------------+

NO. FACTORS = 5, INTERCORRELATION = .5
         +-----------+
     +---+    *     +----------+
         +-----------+

Minimum                                                                Maximum
 .0033                                                                  .0689
```

Figure 3

Box-plot of FADRES

```
NO. FACTORS = 1
  +-+
 +-*+-+
  +-+

NO. FACTORS = 2
     +--------------------+
 +--+   *                +------------------------------+0000000 00           0 0  0
     +--------------------+                               4 2   3 22

NO. FACTORS = 5
     +----------+
  +-+   *        +--------------+0 000000 0 0
     +----------+                  63 252   2

NO. FACTORS = 1, DISCRIMINATION = MIXED
  +-+
 +-*+-+
  +-+

NO. FACTORS = 1, DISCRIMINATION = ALL ONE'S
  +-+
 +-*+-+
  +-+

NO. FACTORS = 2, INTERCORRELATION = .1
               +------------------------+
      +-----+           *               +----------------------------+          0 0  0
               +------------------------+

NO. FACTORS = 2, INTERCORRELATION = .5
   +--+
 +- *+--+
   +--+

NO. FACTORS = 5, INTERCORRELATION = .1
         +-----------------+
    +--+       *            +---------------+
         +-----------------+

NO. FACTORS = 5, INTERCORRELATION = .5
   +--+
 +-* +---+
   +--+

Minimum                                                                    Maximum
 .0014                                                                      .0924
```

Figure 4

Box-plot of NHORES

```
NO. FACTORS = 1
  +-+
 +-*+-+
  +-+

NO. FACTORS = 2
    +------------------+
 +--+   *              +---------------------------+  000 0 0 0000  0     0    X
    +------------------+                              43   3 2 2   2

NO. FACTORS = 5
    +----------+
  +-+  *       +-------------+000000000 000
    +----------+              3 252 44

NO. FACTORS = 1, DISCRIMINATION = MIXED
  +-+
 +-*+-+
  +-+

NO. FACTORS = 1, DISCRIMINATION = ALL ONE'S
  +-+
 +-*+-+
  +-+

NO. FACTORS = 2, INTERCORRELATION = .1
            +-------------------------+
        +---+         *               +------------------------------+    0    0
            +-------------------------+

NO. FACTORS = 2, INTERCORRELATION = .5
  +---+
 +- *+---+
  +---+

NO. FACTORS = 5, INTERCORRELATION = .1
        +---------------+
     +--+     *         +----------------+
        +---------------+

NO. FACTORS = 5, INTERCORRELATION = .5
   +---+
  +-* +---+
   +---+

Minimum                                                                    Maximum
 .0014                                                                      .0922
```

overlaps. In these plots the median is represented by a star (*) and the boundaries to the "box" are the upper and lower hinges (the 25th and 75th percentiles). The "H-spread" is the distance between the upper and lower hinges. The observation farthest from the median that still remains within one step (1.5 "H-spread") from each hinge is marked by '+'. The values in the second step (between 1.5 and 3 "H-spreads") from the hinges are marked with the letter 'O' and the values beyond the second step are marked 'X'. Where multiple data points are at positions marked X or O, the number of multiple points is noted.

It is quite clear that the best decision criterion for determining unidimensionality is either the FADIV or NOHARM absolute sum of residuals. These two indices are highly related (.998) and from Table 24, it can be seen that the means are very similar. The means from the two factor data (intercorrelation .1) are higher than the five factor (.1) data. Thus there does not seem to be a monotone relationship between the sum of residuals and the number of factors. The NONRES had more overlap than either NHORES or FADRES, particularly regarding the two factor (.1 intercorrelation) case. But the means for NONRES over the one, two, and five factors do increase monotonically (Table 24).

Table 24

Means of NHORES, FADRES, and NONRES cross-tabulated with difficulty, guessing, number of factors, and discrimination (or correlation)

		No. of factors					
		1		2		5	
Difficulty	Guessing	Mix	All 1	.1	.5	.1	.5
		NHORES					
	.0	.0046	.0044	.0688	.0085	.0401	.0087
-1 1	.2	.0064	.0061	.0325	.0078	.0214	.0083
	Mix	.0057	.0051	.0455	.0077	.0289	.0089
	.0	.0025	.0020	.0218	.0033	.0142	.0040
-2 2	.2	.0049	.0045	.0126	.0051	.0097	.0053
	Mix	.0035	.0035	.0162	.0043	.0111	.0046
		FADRES					
	.0	.0044	.0044	.0701	.0085	.0406	.0087
-1 1	.2	.0063	.0061	.0326	.0078	.0215	.0083
	Mix	.0056	.0051	.0467	.0077	.0292	.0090
	.0	.0025	.0020	.0252	.0034	.0150	.0040
-2 2	.2	.0050	.0046	.0128	.0052	.0099	.0053
	Mix	.0035	.0036	.0171	.0044	.0117	.0046
		NONRES					
	.0	.0099	.0100	.0150	.0178	.0532	.0188
-1 1	.2	.0138	.0134	.0165	.0165	.0373	.0180
	Mix	.0120	.0107	.0150	.0164	.0424	.0196
	.0	.0059	.0048	.0107	.0074	.0268	.0087
-2 2	.2	.0098	.0096	.0145	.0108	.0207	.0116
	Mix	.0072	.0068	.0177	.0090	.0194	.0098

Chapter V

Summary and Conclusions

The purposes of this study were: (1) to review various methods for determining unidimensionality and to assess the rationale of these methods; (2) to attempt to clarify the term unidimensionality, and to show how it differs from other terms often used interchangeably with it; and (3) to assess the effectiveness of various indices proposed to determine unidimensionality. The aim was not to invent new procedures but to attempt to justify existing ones.

Summary of the literature review

Unidimensionality is a fundamental assumption of measurement and Chapter II outlined 87 indices that have been suggested as decision criteria for determining unidimensionality. Many of the methods lacked a rationale, the sampling distributions of some were not known, and some were adjustments to an established index to take into account criticisms of the established index. The indices were grouped into four sections based respectively on answer patterns, reliability, component or factor analysis, and latent traits.

In Appendix A it was demonstrated how the terms unidimensionality, reliability, internal consistency, and homogeneity have been used interchangeably. At the end of Chapter II definitions were proposed that seemed consistent with how many authors have used the terms (or at least, what they appear to have intended when they have used them): these definitions aimed at setting clearer boundaries between the terms. Reliability was defined as the ratio of true score variance to observed score variance. Internal consistency seemed to label a group of methods for estimating the reliability. Such methods are based on, the variances and covariances of test items, and depend on only one administration of a test. Homogeneity, on one view, is a synonym for unidimensionality, but has also been used to refer more specifically to the similarity of item correlations. Unidimensionality was defined as the existence of one latent trait underlying the data. The usefulness of the terms internal consistency and homogeneity was questioned.

Thus unidimensionality is not defined in terms of unit rank, percentage of variance explained by the first component or factor, in terms of deviations from a perfect scale, in terms of the type of correlation, or by the number of common factors. While these have been used as methods to determine unidimensionality, they do not define it. A unidimensional test is not necessarily reliable, internally consistent, nor homogeneous. Indeed a unidimensional test may be factorially complex in terms of the linear common factor model. While the principle of local independence is fundamental to the definition of latent traits, and therefore to the definition of dimensionality, unidimensionality is itself defined as the existence of only <u>one</u> latent trait underlying the set of items.

On theoretical grounds, it was shown that some indices must fail, while others looked promising. A Monte Carlo simulation was suggested as a method for assessing the various indices under known conditions.

Summary of the simulation and results

The simulation

A multivariate logistic latent trait model was used as the basis of the simulation. Parameters were chosen to reflect previous studies using empirical and simulated data and included difficulty (a wide and narrow range), guessing (fixed at .0, .2, and a mixture of .0, .1, and .2), and one, two, and five factors. There were two sets of discrimination values for the one factor data. One set consisted of a mixture of values (of .6, 1, and 1.4), and the other set had equal discrimination values (all of one). The latter is easily recognised as the one-parameter, or Rasch model. The intercorrelations between the two and five factors were all .1 or all .5, and a simple structure, positive manifold pattern of loadings (all 1) was used; one additional advantage of this choice of loadings is that there were no linear dependencies between the factors.

The model has not been used before and hence it was not possible to capitalize on experience. The choice of intercorrelations turned out to be important. Many indices could distinguish between one and two factors when the intercorrelations were .1 but not when the intercorrelations were .5. More simulations with intercorrelations between .1 and .5 are necessary to assess if there is a cutoff point where the indices appear to fail. If the indices are not effective with constant loadings, then they almost certainly would not be effective when the loadings were varied.

Limitations of the simulation

That the ability parameters were drawn from a unit normal distribution restricts the generalizations from the simulation. This is the best possible condition, given that the development of many of the indices assumes normality, and it would be worthwhile to undertake further simulations based on different distributions. Also the number of items and the number of persons could be altered, even though differences probably would be more noticeable if there were fewer items and fewer people, rather than if there were more items and more people.

Thus the generalizations that can be made on the basis of this research are limited by the choice of parameters, and by the use of a normal distribution of ability. In light of these restrictions, the main points developed in the previous chapters are summarised.

Summary of results

The methods based on answer patterns required the strict assumption of scalability. Loevinger's H_t correlated .98 with Green's consistency index and both correlated highly with alpha. These methods are not indices of unidimensionality and seem more to be measures of internal consistency. Green's RepB did not correlate as highly with the other methods based on answer patterns, nor with alpha.

There is a high correlation between alpha and the various modifications. Cronbach appears to have been correct when he claimed that many of the modifications to alpha would lead to negligible differences. The means ranged from .59 (for KR-21) to .90 (for omega based on phi's), but to assess which is the most accurate estimate of reliability needs further discussion of the theoretical derivations and assumptions. This presumably would lead to a different simulation. The number of significant correlations also related highly to alpha, and occurred regardless of the choice of correlation. The mean of tetrachorics is over twice as large as the mean of the phis, but not surprisingly the means of the number of significant correlations do not differ markedly. Guessing does lead to a decrease in alpha, yet contrary to Ryan (1979), alpha increased as discrimination became more varied, it decreased as difficulty increased in range, and had no relation to the number of factors underlying the data.

Indices based on component or factor analysis do not aid in determining unidimensionality. The indices based on the amount of variance explained were highly correlated yet the means differed depending on the choice of correlation. The tetrachoric means were always higher than the phis, particularly when principal component methods were used. The means for the maximum likelihood and least squares were very similar.

A major problem with using tetrachorics in the maximum likelihood solutions was that there were many cases when the matrix was not positive-definite. These non-Gramian matrices occurred most often in cases with wide difficulty and low or zero guessing. In such cases it is more likely that an item will occur that is really difficult (i.e., no one will get it correct). Then the tetrachorics are very poorly estimated (e.g., see Guilford's example on p. 37 above).

The number of eigenvalues greater than one generally resulted in an overestimation of the number of factors in cases where it was known that there was only one factor, and underestimation in the other instances. Generally the values were larger when guessing occurred, when there was a wide range of difficulty, and when discriminations were varied. Generally, values based on tetrachorics were lower than values based on phis. Using the number of eigenvalues greater than one to estimate the number of factors appears to lack justification.

The Green, Lissitz and Mulaik indices are too often above their theoretical maximum. Generally, their intercorrelations were not as high as was expected. The Dubois indices were not effective criteria. Dubois' suggestion of a .95 cutoff point above which a set of items could be called unidimensional, was attained in the majority (63.97%) of cases. Of the cases below the suggested cutoff point, 21.25% were below .95 for the one factor data, 41.22% for the two factor, and 37.53% for the five factor data. The Kelley index and the ratio of the first and second eigenvalues were not effective indices. The ratio of the differences of the first and second, and the second and third eigenvalues based on tetrachoric correlations was not able to distinguish between one factor and five factors (.5 intercorrelation) and further, the theoretical problems (see p. 29) make the use of this index questionable.

There were substantial differences in the chi-square statistic from the maximum likelihood solution depending on whether phis or tetrachorics were used. If we do in fact attempt to employ the chi-square probability to test the hypothesis of one factor, the hypothesis could not be rejected in most cases where there was a narrow difficulty range and mixed or .2 guessing, and the chi-square was based on phis. However, the chi-square could not be rejected with wide difficulty and .0 or mixed guessing when the chi-square was based on tetrachorics. The difference between the two factor and one factor chi-square was probably not effective because the probabilities were nearly always in the extreme tails (note the difference for phis was 96.59 and for tetrachorics was 305.93, with 14 degrees of freedom). The maximum likelihood and least squares residuals were highly correlated with the chi-square and with the Tucker and Lewis indices. (See Montanelli, 1974, for a more detailed discussion of the relation between the Tucker and Lewis index, the chi-square and the residuals.) Overall, linear factor analysis is not appropriate for determining unidimensionality.

Except for the within-t test, it is not correct to claim that the tests of fit from the one-parameter estimation procedures provide "the most direct test of the unidimensionality assumption" (Rentz & Rentz, 1979, p.5). The within-t test, which summarizes the degree of misfit remaining within ability groups after the between-group misfit has been removed from the total, is the only index related to the number of factors. Yet even this index has much overlap and the means are all close to the expected value of one. There is little relationship between the various statistics, and it makes little difference whether misfitting people are excluded or not. There were 5.32% who had zero or perfect scores, or who were misfitting. Predictably, there were more zero scores with narrow difficulty and zero guessing, and more perfect scores with narrow difficulty and .2 guessing. The more a person guessed, the more likely that person would not fit. This confirms Ryan, Garcia-Quintanta and Hamm's (1980) findings that "in the vast majority of cases person misfit occurred when low ability students correctly responded to relatively difficult items" (p. 10). In the models with one factor, only the between-fit t was sensitive to the assumption of equal discriminations. With no guessing, the between-fit t values were close to the expected value of zero, otherwise they were at least one standard deviation away and there were no marked differences between the values. Error impact also was sensitive to varied discrimination, only when there was no guessing, whereas the between-fit and total-fit t statistics were similar regardless of guessing.

The three effective indices

Nonlinear factor analysis is satisfactory as the sum of absolute residuals does relate to the number of known factors. The overlap between the one factor, and the two and five factor residuals may be high but it may be expected that the nonlinear method will be more robust to violations of normality than the NOHARM and FADIV methods. More simulations are clearly needed to confirm this conjecture. The sum of absolute residuals from the nonlinear solution increased monotonically as the number of factors increased. The values decreased as the difficulty widened, and increased as guessing increased.

The two methods for estimating the parameters of the two-parameter latent trait model were effective in discriminating between one and more than one factor. The NOHARM and FADIV sum of absolute residuals were remarkably similar in means and were highly correlated. McDonald and Fraser (in prep.) are presently following up this similarity and they are assessing which method is most accurate in recovering known values, and which is robust to differing underlying distributions, and they are detailing the differences between the methods. Of interest is that guessing based on the bottom 10% of persons (total score) and estimates based on guessing fixed at .0 resulted in similar means. When guessing is fixed at .2, then the sum of absolute residuals failed the criterion relating to discriminating between one and more than one factor. It seems that fixing guessing at .0, even when the data were known to have .2 guessing, makes little difference to the NOHARM or FADIV means. There seems little difference between using FADIV and NOHARM except in time (NOHARM is twice as fast as FADIV), and in the size of problem that can be handled by the programme (FADIV is restricted to 40, whereas NOHARM is at present restricted to 300 items).

Perhaps it is not surprising that the two-parameter model was among the most effective decision criteria given that the data were generated via a three-parameter model. Simulations where data are generated by alternative models might seem necessary. The fundamental problem is to find alternative models that might imply clear alternative definitions of unidimensionality. Yet, at a minimum, it is possible to conclude that the sum of absolute residuals from nonlinear factor analysis, NOHARM or FADIV are effective under the conditions used in this thesis. These conditions are not atypical and they are the basis of many measurement issues (see Lord, 1980).

The nonlinear and two-parameter estimation procedures are not altogether dissimilar. The nonlinear method attempts to fit a polynomial function to a scatter plot of binary data (via least squares in the present simulation). In the NOHARM and FADIV methods, it is assumed that the data follow a normal ogive curve and the parameters of the normal ogive function are estimated by fitting some polynomial functions. It is thus expected that the NOHARM and FADIV methods provide better fit than the nonlinear method as the data were generated based on the logistic model (with d = 1.7), which meant that the data were generated almost as a normal ogive (see p. 52, above). Yet if the data are generated by alternative methods or based on a non-normal distribution of ability, the nonlinear method probably will be more effective.

Conclusion

It thus appears that there are three indices that can be used as decision criteria for determining unidimensionality. These indices are the sum of absolute residuals from a nonlinear factor analysis (specifying one factor with cubic terms), from Christoffersson and Muthen's two-parameter latent trait estimation procedure, and from McDonald and Fraser's "normal ogive harmonic analysis robust method". Clearly, more simulations are necessary to assess whether these three indices are robust, particularly when there is a non-normal distribution of ability.

It should be noted that it was not an objective of this study to derive a decision criterion for determining unidimensionality: rather, it was to study the effectiveness of various existing decision criteria. While it has been shown that almost all of the indices discussed in Chapter II were not effective, no attempt was made to arrive at a "rule" regarding the three indices that passed the four criteria outlined in Chapter III. Given a set of data of unknown dimensionality, a first step would be to use nonlinear factor analysis specifying one factor with cubic terms, or use FADIV or NOHARM, and inspect the residuals. Further research is needed before we can recommend specific cutoff points below which it can be concluded that a set of items is unidimensional.

From practical considerations, FADIV is restricted to a small number of items, and the data capacity of the nonlinear programme is dependent on sample size. NOHARM, however, is not sample dependent and can handle, at present, up to 300 items. On the other hand, both NONLIN and FADIV can be used to test structures that are not unidimensional.

Certainly, the findings reported in this thesis point to the value of the nonlinear factor analysis methods as providing an effective decision criterion for determining unidimensionality. This is consistent with the theoretical argument (see McDonald, 1981) for basing an index directly on misfit to the nonlinear factor analysis model.

References

Ahlawat, K. *Alternative methods of estimating the parameters of the normal ogive model*. Unpublished doctoral dissertation, University of Toronto, 1976.

American Psychological Association, American Educational Research Association, & National Council for Measurement in Education. *Standards for educational and psychological tests and manuals*. Washington, D.C.: American Psychological Association, 1966.

Anderson, C.C. Function fluctuation. *British Journal of Psychology*, 1958, *30*. (Monograph).

Anderson, T.W. Some scaling models and estimation procedures in the latent class model. In O. Grenander (Ed.), *Probability and statistics*. New York: Wiley, 1959.

Andersson, C.G. & Christoffersson, A., & Muthen, B. *FADIV. A computer program for factor analysis of dichotomized variables* (Report No. 74-1). Uppsala, Sweden: Uppsala University, Statistics Department, 1974.

Aoyama, H. Sampling fluctuations of the test reliability. *Annals of the Institute of Statistical Mathematics Tokyo*. 1957, *8*, 129.

Armor, D.J. Theta reliability and factor scaling. In H.L. Costner, (Ed.), *Sociological methodology*. San Francisco: Jossey-Bass, 1974.

Baker, F.B. *Empirical determination of sampling distribution of item discrimination indices and a reliability coefficient*. Final Report Contract OE-2-10-071, U.S. Office of Education, 1962.

Baker, F.B. Advances in item analysis. *Review of Educational Research*, 1977, *47*, 151-178.

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

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

Birnbaum, A. Statistical theory for logistic mental test models with a prior distribution of ability. *Journal of Mathematical Psychology*, 1969, *6*, 258-276.

Bock, R.D. & Aitken, M. *Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm*. Manuscript submitted for publication, 1981.

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

Brooks, R.D. An empirical investigation of the Rasch ratio-scale model for item difficulty indexes. (Doctoral dissertation, University of Iowa, 1965). *Dissertation Abstracts*, 1965, *26*, 2047A. (University Microfilms No. 65-434)

Brown, M. & Beneditti, J.K. On the mean and variance of the tetrachoric correlation coefficient. *Psychometrika*, 1977, *42*,347-355.

Carmines, E.G. & Zeller, R.A. *Reliability and validity assessment*. Beverley Hills, Cal.: Sage, 1979.

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

Carroll, J.B. *Evaluation of achievement tests in terms of internal statistics*. Paper presented at the Invitational Conference on Testing Problems, Washington, D.C., 1950.

Castellan, N.J. On the estimation of the tetrachoric correlation coefficient. *Psychometrika*, 1966, *31*, 67-73.

Cattell, R.B. Validity and reliability: A proposed more basic set of concepts. *Journal of Educational Psychology*, 1964, *55*, 1-22.

Cattell, R.B. *The scientific use of factor analysis in behavioral and life sciences*. New York: Plenum, 1978.

Cattell, R.B. & Butchers, H.J. *The prediction of achievement and creativity*. Indianapolis: Bobbs-Merrill, 1968.

Cattell, R.B. & Tsujioka, B. The importance of factor-trueness and validity, versus homogeneity and orthogonality, in test scales. *Educational and Psychological Measurement*, 1964, *24*, 3-30.

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

Cleary, T.A. & Linn, R.L. *Variability of Kuder-Richardson formula 20*. (ETS RM 68-7). Princeton, N.J.: Educational Testing Service, 1968.

Cohen, J. *Statistical power for the behavioral sciences*. New York: Academic Press, 1969.

Cohen, L. *A modified logistic response model for item analysis*. Unpublished manuscript, 1976.

Comrey, A.L. & Levonian, E. A comparison of three point coefficients in factor analysis of MMPI items. *Educational and Psychological Measurement*, 1958, *18*, 739-755.

Conrad, H.S & Sanford, R.N. Some specific war-attitudes of college students. Journal of Psychology, 1944, 17, 153-186.

Coombs, C.H. A theory of data. New York: Wiley, 1964.

Coombs, C.H. & Kao, R.C. On the multidimensional analysis of monotone single stimuli data. In C.H. Coombs, R.M. Thrall & R.C. Davies (Eds.), Decision processes. New York: Wiley, 1954.

Coombs, C.H. & Kao, R.C. Non-metric factor analysis. Engineering Research Bulletin, No. 38. Ann Arbor, Mich: University of Michigan Press, 1955.

Corballis, M.C. Some difficulties with difficulty: Note on Horst's "matrix factoring and test theory". Psychological Reports, 1968, 22, 15-22.

Cotton, J.W., Campbell, D.T. & Malone, R.D. The relationship between tetrachoric composition of test items and measures of test reliability. Psychometrika, 1957, 22, 347-357.

Coughlan, J.R. A multidimensional extension of the normal ogive, logistic and linear latent trait models. Unpublished masters' dissertation, University of Toronto, 1974.

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

Dinero, T.E. & Haertel, E. Applicability of the Rasch model with varying item discriminations. Applied Psychological Measurement, 1977, 4, 581-592.

Dingman, H.F. The relation between coefficients of correlation and difficulty factors. British Journal of Statistical Psychology, 1958, 11, 13-18.

Divgi, D.R. Calculation of the tetrachoric correlation coefficient. Psychometrika, 1979, 44, 169-172.

Divgi, D.R. Dimensionality of binary items: Use of a mixed model. Paper presented at the 1980 Annual meeting of the National Council on Measurement in Education. Boston, April, 1980.

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

Dubois, P.H., Loevinger, J. & Gleser, G.C. The construction of homogeneous keys for a biographical inventory (Report 52-18). San Antonio, Texas: Lackland Airforce Base, Human Resources Research Center, Personnel Research Laboratory, May, 1952.

Elderton, W.R. Frequency curves and correlation. Washington, D.C.: Harren Press, 1906.

Etezadi, J. *A general polynomial model for nonlinear factor analysis.* Unpublished doctoral thesis, University of Toronto, 1981.

Feldt, L.S. The approximate sampling distribution of Kuder-Richardson reliability coefficient twenty. *Psychometrika*, 1965, *30*, 357-370.

Finn, J. *MULTIVARIANCE: Univariate and multivariate analysis of variance, covariance, regression and repeated measures.* Version VI, Release 2. Chicago: National Educational Resources, 1978.

Fraser, C. *NOHARM: A program for estimating parameters of the normal ogive by harmonic analysis robust method.* University of Toronto, 1980.

Freeman, F.S. *Theory and practice of psychological testing* (3rd ed.). New York: Henry Holt, 1962.

Froemel, E. A comparison of computer routines for the calculation of the tetrachoric correlation coefficient. *Psychometrika*, 1971, *36*, 165-174.

Fuller, E.L. & Hemmerle, W.J. Robustness of the maximum-likelihood estimation procedure in factor analysis. *Psychometrika*, 1966, *31*, 255-266.

Gage, N.L. & Damrin, D.E. Reliability, homogeneity, and number of choices. *Journal of Educational Psychology*, 1950, *41*, 385-404.

Garrison, W.M. & White, K.R. *A simulation study on the utility of Rasch and Classical test analysis procedures.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April, 1979.

George, A.A. *Theoretical and practical consequences of the use of standardized residuals as Rasch model fit statistics.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April, 1979.

Gibson, W.A. Three multivariate models: Factor analysis, latent structure analysis, and latent profile analysis. *Psychometrika*, 1959, *24*, 229-252.

Gibson, W.A. Nonlinear factors in two dimensions. *Psychometrika*, 1960, *25*, 381-392.

Goldstein, H. Dimensionality, bias, independence and measurement scale problems in latent trait test score models. *British Journal of Mathematical and Statistical Psychology*, 1981, *33*, 234-246.

Gorsuch, R.L. *Factor analysis.* Philadephia: W.B. Saunders, 1974.

Gourlay, N. Difficulty factors arising from the use of the tetrachoric correlations in factor analysis. *British Journal of Statistical Psychology*, 1951, *4*, 65-72.

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

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

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

Guilford, J.P. The difficulty of a test and its factor composition. *Psychometrika*, 1941, *6*, 67-77.

Guilford, J.P. *Fundamental statistics in psychology and education* (4th ed.). New York: McGraw-Hill, 1965.

Gulliksen, H. *Theory of mental tests*. New York: Wiley, 1950.

Gupta, R.K. & Burnett, J.D. A program to carry out cluster analysis by homogeneous grouping. *Educational and Psychological Measurement*, 1972, *32*, 185-190.

Gustafsson, J.E. Testing and obtaining fit of data to the Rasch model. *British Journal of Mathematical and Statistical Psychology*, 1980, *33*, 205-233.

Gustafsson, J.E. & Lindblad, T. *The Rasch model for dichotomous items: A solution of the conditional estimation problem for long tests and some thoughts about item screening procedures*. Reports from the Institute of Education, University of Goteborg, No. 67.

Guttman, L. A basis for scaling qualitative data. *American Sociological Review*, 1944, *80*, 139-150.

Guttman, L. A basis for analyzing test-retest reliability. *Psychometrika*, 1945, *10*, 255-282.

Guttman, L. The principal components of scale analysis. In S.S. Stouffer (Ed.), *Measurement and prediction*. Princeton, N.J.: University Press, 1950.

Guttman, L. A new approach to factor analysis: The radex. In P.F. Lazarsfeld (Ed.), *Mathematical thinking in the social sciences*. New York: Columbia Press, 1954.

Guttman, L. Measurement as structural theory. *Psychometrika*, 1971, *36*, 329-347.

Haley, D.C. *Estimation of dosage mortality relationships when the dose is subject to error*. (Report No. 15). Stanford, Cal.: Stanford University, 1952.

Hambleton, R.K. *An empirical investigation of the Rasch test theory model*. Unpublished doctoral dissertation, University of Toronto, 1969.

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

Hambleton, R.K., Swaminathan, H., Cook, L.L., Eignor, D.R., & Gifford, J.A. Developments in latent trait theory: Models, technical issues, and applications. *Review of Educational Research*, 1978, *48*, 467-510.

Hambleton, R.K. & Traub, R.E. Information curves and efficiency of three logistic test models. *British Journal of Mathematical and Statistical Psychology*, 1971, *24*, 273-281.

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

Harman, H.H. *Modern factor analysis* (3rd ed.). Chicago: University of Chicago Press, 1979.

Hattie, J.A. A confirmatory factor analysis of the Progressive Achievement Tests: Reading Comprehension, reading vocabulary, and listening comprehension tests. *New Zealand Journal of Educational Studies*, 1979, *14*, 172-188. (Errata, 1980, *15*, 109.)

Hattie, J.A. The Progressive Achievement Tests revisited. *New Zealand Journal of Educational Studies*, 1980, *15*, in press. (a)

Hattie, J.A. *UNIDIM: A Fortran program to calculate indices of unidimensionality based on answer patterns, reliability, and factor analysis*. University of Toronto, 1980. (b)

Hattie, J.A. *LORD: A Fortran program to calculate an index of unidimensionality based on Lord's chi-square test*. University of Toronto, 1980. (c)

Hattie, J.A. A four stage factor analytic approach to studying behavioral domains. *Applied Psychological Measurement*, in press, 1981.

Hattie, J.A. & Hansford, B.F. *Evaluating communication apprehension: Implications for teachers and teaching*. Paper presented at the annual meeting of the Australian Association for Research in Education Conference, Melbourne, November, 1979.

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

Helsley, T.L., Suber, J.S. & Ryan, J.P. Item homogeneity, unidimensionality and test reliability: Comparing classical and latent trait psychometric procedures. Paper presented at the meeting of the Southeastern Psychological Association, Hollywood, Florida, May, 1977.

Henrysson, S. The relation between factor loadings and biserial correlation in item analysis. Psychometrika, 1962, 27, 419-424.

Hertzman, M. The effects of the relative difficulty of mental tests on patterns of mental organization. Archives of Psychology, 1936, No. 197.

Hoffman, R.J. The concept of efficiency in item analysis. Educational and Psychological Measurement, 1975, 35, 621-640.

Hoffman, R.J. & Gray, W.M. On partialling a simplex out of binary data. Multivariate Behavioral Research, 1978, 13, 223-227.

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

Horn, J.L. On the internal consistency and reliability of factors. Multivariate Behavioral Research, 1969, 4, 115-125.

Horrocks, J.E. & Schoonover, T.I. Measurement for teachers. Columbus, Ohio: Merrills, 1968.

Horst, P. Correcting the Kuder-Richardson reliability for dispersion of item difficulties. Psychological Bulletin, 1953, 50, 371-374.

Horst, P. Matrix factoring and test theory. In N. Frederickson & H. Gulliksen (Eds.), Contributions to mathematical psychology. New York: Holt, Rinehart & Winston, 1964.

Hoyt, C. Test reliability estimated by analysis of variance. Psychometrika, 1941, 6, 153-160.

Humphreys, L.G. Test homogeneity and its measurement. American Psychologist, 1949, 4, 245. (Abstract).

Humphreys, L.G. Individual differences. In C.P. Stone & D.W. Taylor (Eds.), Annual Review of Psychology. Stanford: Annual Reviews Inc., 1952.

Humphreys, L.G. The normal curve and the attenuation paradox in test theory. Psychological Bulletin, 1956, 53, 472-476.

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

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

IBM Scientific Subroutines. New York: IBM Corporation, 1970.

IMSL library, (8th ed.). Houston, Texas, 1980.

Indow, T. & Samejima, F. On the results obtained by the absolute scaling model and the Lord model in the field of intelligence. Yokohama: Psychological Labatory, Hiyoshi Campus, Keio University, 1966.

Jackson, J.M. A simple and more rigorous technique for scale analysis. In A manual of scale analysis, Part II. Montreal: McGill University, 1949.

Jackson, R.W.B. & Ferguson, G.A. Studies in the reliability of tests. Bulletin 12, Department of Educational Research, University of Toronto, 1941.

Jensema, C.J. A simple technique for estimating latent trait mental test parameters. Educational and Psychological Measurement, 1976, 36, 705-715.

Joreskog, K.G. Structural analysis of covariance and correlation matrices. Psychometrika, 1978, 43, 443-447.

Kaiser, H.F. A measure of the average intercorrelation. Educational and Psychological Measurement, 1968, 28, 245-247.

Kaiser, H.F. A second generation little jiffy. Psychometrika, 1970, 35, 401-415.

van der Kamp, L.J.Th. Studies in reliability. (Research Report 002-76). Leiden, Netherlands: University of Leiden, Department of Psychology, 1976.

Kelley, T.L. Essential traits of mental life. Harvard Studies in Education, 1935, 26, 146-153.

Kelley, T.L. The reliability coefficient. Psychometrika, 1942, 7, 75-83.

Kim, J. & Rabjohn, J. Binary variables and index construction. In K.F. Schuessler (Ed.), Sociological methodology. San Francisco: Jossey-Bass, 1980.

Kingsbury, G.G. & Weiss, D.J. Relationships among achievement level estimates from three item characteristic curve scoring methods. (Research Report 79-3). Minneapolis, University of Minnesota, Department of Psychology, Psychometrics methods program, April, 1979.

Kirk, D.B. On the numerical approximation of the bivariate normal (tetrachoric) correlation coefficient. *Psychometrika*, 1973, *38*, 259-268.

Koch, W.R. & Reckase, M.D. *Problems in application of latent trait models to tailored testing*. (Research Report 79-1). Columbia, Missouri, University of Missouri, Educational Psychology department, Tailored testing laboratory, September, 1979.

Kristof, W. *Statistical notes on reliability estimation*. (ETS RM 69-25). Princeton, N.J.: Educational Testing Service, 1969.

Kuder, G.F. & Richardson, M.W. The theory of the estimation of test reliability. *Psychometrika*, 1937, *2*, 151-160.

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

Lam, R. *An empirical study of the unidimensionality of Ravens Progressive Matrices*. Unpublished Masters' thesis, University of Toronto, 1980.

Lawley, D.N. The factorial analysis of multiple item tests. *Procedures of the Royal Society of Edinburgh*, 1944, *62-A*, Part I, 74-82.

Lazarsfeld, P.F. Latent structure analysis. In S. Koch (Ed.), *Psychology: A study of a science*, Vol. 3. New York: McGraw-Hill, 1959.

Levine, M. Psychometric models and appropriateness measurement. In D.J.Weiss (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference*, University of Minnesota, 1977.

Levine, M. & Rubin, D. Measuring the appropriateness of multiple-ohoice test scores. *Journal of Educational Statistics*, 1979, *4*, 269-290.

Lim, T.P. *Estimation of probabilities of dichotomous response patterns using a simple linear model*. Unpublished doctoral dissertation, University of Toronto, 1974.

Linn, R.L. A Monte Carlo approach to the number of factors problem. *Psychometrika*, 1968, *33*, 37-71.

Loevinger, J. *A systematic approach to the construction of tests of ability*. Unpublished doctoral dissertation, University of California, 1948.

Loevinger, J. A systematic approach to the construction and evaluation of tests of ability. *Psychological Monograph*, 1947, *61*, No. 4 (Whole No. 285).

Loevinger, J. The technique of homogeneous tests. *Psychological Bulletin*, 1948, *45*, 507-529.

Loevinger, J. The attenuation paradox in test theory. Psychological Bulletin, 1954, 51, 493-504.

Lord, F.M. A theory of test scores. Psychometrika, No.7, 1952. (Monograph)

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

Lord, F.M. An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. Psychometrika, 1953, 23, 56-76. (b)

Lord, F.M. Some relations between Guttman's principal components of scale analysis and other psychometric theory. Psychometrika, 1958, 23, 291-296.

Lord, F.M. An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. Psychometrika, 1968, 28, 989-1020.

Lord, F.M. Estimation of latent ability and item parameters when there are omitted responses. Psychometrika, 1974, 39, 247-264.

Lord, F.M. The 'ability' scale in item characteristic curve theory. Psychometrika, 1975, 40, 205-217.

Lord, F.M. Discussion on alternative models for adaptive testing. In D.J. Weiss (Ed.), Proceedings of the 1977 Computerized Adaptive Testing Conference, University of Minnesota, 1977.

Lord, F.M. Applications of item response theory to practical testing problems. New York: Erlbaum Associates, 1980.

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

Lumsden, J. A factorial approach to unidimensionality. Australian Journal of Psychology, 1957, 9, 105-111.

Lumsden, J. The construction of unidimensional tests. Unpublished Master's thesis. University of Western Australia, 1959.

Lumsden, J. The construction of unidimensional tests. Psychological Bulletin, 1961, 58, 122-131.

Lumsden, J. Test theory. Annual Review of Psychology, 1976, 27, 251-280.

Lumsden, J. Person reliability. Applied Psychological Measurement, 1977, 1, 477-482.

Lumsden, J. Tests are perfectly reliable. British Journal of Mathematical and Statistical Psychology, 1978, 31, 19-26.

Lyman, H.B. *Test scores and what they mean*. Englewood Cliffs, N.J.: Prentice-Hall, 1963.

Magnusson, P. *Test theory*. Reading, Mass.: Addison-Wesley, 1966.

McDonald, R.P. A general approach to nonlinear factor analysis. *Psychometrika*, 1962, *27*, 397-415. (a)

McDonald, R.P. A note on the derivation of the general latent class model. *Psychometrika*, 1962, *27*, 203-206. (b)

McDonald, R.P. Difficulty factors and non-linear factor analysis. *British Journal of Mathematical and Statistical Psychology*, 1965, *18*, 11-23. (a)

McDonald, R.P. *Numerical polynomial models in nonlinear factor analysis*. (ETS RM 65-32). Princeton, N.J.: Educational Testing Service, 1965. (b)

McDonald, R.P. Nonlinear factor analysis. *Psychometrika*, No. 15, 1967. (Monograph) (a)

McDonald, R.P. Numerical methods for polynomial models in non-linear factor analysis. *Psychometrika*, 1967, *32*, 77-112. (b)

McDonald, R.P. Factor interaction in nonlinear factor anaysis. *British Journal of Mathematical and Statistical Psychology*, 1967, *20*, 205-215. (c)

McDonald, R.P. *PROTEAN - A comprehensive CD3200/3600 program for nonlinear factor analysis* (ETS RM 67-26). Princeton, N.J.: Educational Testing Service, 1967. (d)

McDonald, R.P. The theoretical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis. *British Journal of Mathematical and Statistical Psychology*, 1970, *23*, 1-21.

McDonald, R.P. *Nonlinear and nonmetric common factor analysis*. Paper presented to the Psychometric Society, Murray Hill, S.C., April, 1976. (a)

McDonald, R.P. *Generalizability and the structure of data, with implications for factor indeterminancy and its measurement*. Unpublished manuscript, Toronto, 1976. (b)

McDonald, R.P. A simple comprehensive model for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, 1978, *31*, 59-72.

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

McDonald, R.P. Fitting a latent trait model by the analysis of covariance structures. Paper read at the 1980 Annual Convention of the Northeastern Educational Research Association, October, 1980. (a)

McDonald, R.P. A simple comprehensive model for the analysis of covariance structures: Some remarks on applications. British Journal of Mathematical and Statistical Psychology, 1980, 33, 161-183. (b)

McDonald, R.P. Some alternative approaches to the improvement of measurement in education and psychology: Fitting latent trait models. Australian Council for Educational Research Invitational Seminar, Melbourne, May, 1980. (c)

McDonald, R.P. The dimensionality of tests and items. British Journal of Mathematical and Statistical Psychology, in press, 1981.

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

McDonald, R.P. & Burr, E.J. A comparison of four methods of constructing factor scores. Psychometrika, 1967, 32, 381-401.

McDonald, R.P. & Fraser, C. COSAN: A Fortran program for the analysis of covariance structures. University of Toronto, 1978.

McDonald, R.P. & Fraser, C. A robustness study comparing estimation of the parameters of the two-parameter latent trait model. In preparation.

McDonald, R.P. & Leong, K.S. COFA: A Fortran program for unrestricted common factor analysis. University of Toronto, 1974.

McDonald, R.P. & Leong, K.S. COSA: A Fortran program for restricted common factor analysis. University of Toronto, 1976.

McDonald, R.P. & Mulaik, S.A. Determinacy of common factors: A nontechnical review. Psychological Bulletin, 1979, 86, 297-306.

McNemar, Q. Opinion-attitude methodology. Psychological Bulletin, 1946, 43, 289-374.

McNemar, Q. Psychological statistics. New York: Wiley, 1955.

Mead, R.J. Analysis of fit to the Rasch model. Unpublished doctoral dissertation, University of Chicago, 1975.

Montanelli, R.G. The goodness of fit of the maximum likelihood estimation procedure in factor analysis. Educational and Psychological Measurement, 1974, 34, 547-562.

Mosier, C.I. A note on item analysis and the criterion of internal consistency. Psychometrika, 1936, 1, 275-282.

Mosier, C.I. Psychophysics and mental test theory: Fundamental postulates and elementary theorems. *Psychological Review*, 1940, *47*, 355-366.

Mosier, C.I. Psychophysics and mental test theory II: The constant process. *Psychological Review*, 1942, *48*, 235-249.

Mosteller, F. & Tukey, J.W. *Data analysis and regression: A second course in statistics*. Reading: Addison-Wesley, 1977.

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

Muthen, B. Personal communication, October 22, 1980.

Muthen, B. Factor analysis of dichotomous variables: American attitudes toward abortion. In D.J. Jackson & E.F. Borgatta (Eds.), *Factor analysis and measurement in sociological research: A multi-dimensional perspective*. In press, 1981.

Muthen, B. & Christoffersson, A. Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika*, in press, 1981.

Nishisato, S. *Minimum entropy clustering of test-items*. Unpublished doctoral dissertation, University of North Carolina, 1966.

Nishisato, S. Structure and probability distribution of dichotomous response patterns. *Japanese Psychological Research*, 1970, *12*, 62-74.

Nishisato, S. Probability estimates of dichotomous response patterns by logistic fractional factorial representation. *Japanese Psychological Research*, 1970, *12*, 87-95.

Nishisato, S. Infomation analysis of binary response patterns. In S. Takagi (Ed.), *Modern psychology and quantification method*. Tokyo: University of Tokyo Press, 1971.

Nishisato, S. *Elements of scaling*. An informal publication of the Department of Measurement and Evaluation, O.I.S.E., Toronto, 1975.

Nishisato, S. Dual scaling and its variants. *New Directions for Testing and Measurement*, 1979, *4*, 1-12.

Nishisato, S. *Analysis of categorical data: Dual scaling and its applications*. Toronto: University of Toronto Press, 1980.

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

Nunnally, J.C. *Psychometric theory*. New York: McGraw-Hill, 1967.

Nunnally, J.C. *Introduction to Psychological Measurement*. New York: McGraw-Hill, 1970.

Owen, R.J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, 1975, *70*, 351-356.

Panchapakesan, N. *The simple logistic model and mental measurement*. Unpublished doctoral dissertation, University of Chicago, 1969.

Payne, D.A. *The specification and measurement of learning*. Waltham, Mass.: Blaisdell, 1968.

Pearson, K. Mathematical contributions to the theory of evolution - VII. On the correlation of characters not quantitatively measureable. *Philosophical Transactions of the Royal Society of London, Series A*, 1901, *195A*, 1-47.

Raju, N.S. *Kuder-Richardson Formula 20 and test homogeneity*. Paper presented at the National Council for Measurement in Education, Boston, April, 1980.

Rasch, G. *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Nielson & Lydiche, 1960.

Rasch, G. On general laws and the meaning of measurement in psychology. *Proceedings of the fourth Berkeley Symposium on Mathematical Statistics*. Berkeley: University of California Press, 1961.

Rasch, G. An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology*, 1966, *19*, 49-57. (a)

Rasch, G. An individualistic approach to item analysis. In P. Lazarsfeld & N.V. Henry, *Readings in Mathematical Social Science*. Chicago: Science Research Association, 1966. (b)

Rasch. G. *A mathematical theory of objectivity and its consequences for model construction*. Paper read at the European Meeting of Statistics, Econometrics and Management Science, Amsterdam, 1968.

Rasch, G. On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy*, 1977, *14*, 58-94.

Reckase, M.D. Development of a multivariate logistic latent trait model. (Doctoral dissertation, Syracuse University, 1972). *Dissertation Abstracts International*, 1972, *33*, 4495B. (University Microfilms. No. 73-7762)

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

Ree, M. Estimating item characteristic curves. Applied Psychological Measurement, 1979, 3, 371-385.

Remmers, H.H., Gage, N.L. & Rummel, J.F. A practical introduction to measurement and evaluation. New York: Harper & Row, 1965.

Rentz, R.R. & Rentz, C.C. Does the Rasch model really work. NCME Measurement in Education, 1979, 10, 1-11.

Richardson, M.W. Notes on the rationale of item analysis. Psychometrika, 1936, 1, 69-76.

Rotter J.B. Generalized expectancies for internal versus control of reinforcement. Psychological Monographs, 1966, 80, (Whole Number 609).

Ross, J. An empirical study of a logistic mental test model. Psychometrika, 1966, 31, 325-340.

Ryan, J.P. Assessing unidimensionality in latent trait models. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April, 1979.

Ryan, J.P., Garcia-Quintanta, R. & Hamm, D.W. Testing the fit of subjects to a latent trait model. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Boston, April 1980.

Ryán, J.P. & Hamm, D.W. Practical procedures for increasing the reliability of classroom tests by using the Rasch model. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, Mass.: 1976.

Sax, G. Principles of educational and psychological measurement and evaluation (2nd ed.). Belmont, California: Wadsworth, 1974.

Schmidt, F.L. The Urry method of approximating the item parameters of latent trait theory. Educational and Psychological Measurement, 1977, 37, 613-620.

Schrage, L. A more portable Fortran random number generator. ACM Transactions on Mathematical Software, 1979, 5, 132-138.

Shostrom, E.L. Personal Orientation Inventory: An inventory for the measurement of self-actualization. California: Edits, 1966, 1972.

Silverstein, A.B. Item intercorrelations, item-test correlations and test reliability. Educational and Psychological Measurement, 1980, 40, 353-355.

Smith, J.K. On the examination of test unidimensionality. Educational and Psychological Measurement, 1980, 40, 885-889.

Smith, K.W. On estimating the reliability of composite indexes through factor analysis. Sociological Methods and Research, 1974, 4, 485-510. (a)

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

Spearman, C. The abilities of man: Their nature and measurement. London: Macmillan, 1927.

Stanley, J.C. Measurement in today's schools (4th ed.). Englewood Cliffs, N.J.: Prentice-Hall, 1964.

Svoboda, M. Distribution of binary information in multidimensional space. Unpublished Masters' thesis, University of Toronto, 1972.

Swaminathan, H. Discussion on Achievement testing viewed as a trait measurement problem. In D.J. Weiss (Ed.), Proceedings of the 1977 Computerized Adaptive Testing Conference, University of Minnesota, 1977.

Swaminathan, H. & Gifford, J.A. Precision and item parameter estimation. Paper presented at the 1979 Computerized Adaptive Testing Conference, Minneapolis, June, 1979.

Sympson, J.B. Estimation of latent trait status in adaptive testing procedures. In D.J. Weiss (Ed.), Proceedings of the 1977 Computerized Adaptive Testing Conference, University of Minnesota, 1977.

Terwilliger, J.S. & Lele, K. Some relationships among internal consistency, reproducibility, and homogeneity. Journal of Educational Measurement, 1979, 16, 101-108.

Thorndike, R.L. & Hagen, E.P. Measurement and evaluation in psychology and education (4th ed.). New York: Wiley, 1977.

Thouless, R.H. Test unreliability and function fluctuation. British Journal of Psychology, 1936, 26, 325-343.

Thurstone, L.L. Unitary abilities. Journal of General Psychology, 1934, 11, 126-130.

Thurstone, L.L. The vectors of the mind. Chicago: University of Chicago Press, 1935.

Thurstone, L.L. & Thurstone, T.G. A neurotic inventory. Journal of Social Psychology, 1930, 1, 3-30.

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

Tucker, L.R., Koopman, R.F. & Linn, R.L. Evaluation of factor analytic research procedures by means of simulated correlation matrices. Psychometrika, 1969, 34, 421-459.

Urry, V.W. A monte carlo investigation of logistic mental test models. Unpublished doctoral dissertation, Purdue University, 1970.

Urry, V.W. Individualized testing by Bayesian estimation. Bureau of Testing, University of Washington, 1971.

Urry, V.W. Appproximations to item parameters of mental test models and their uses. Educational and Psychological Measurement, 1974, 34, 253-269.

Urry, V.W. OGIVIA: Item parameter estimation program with normal ogive and logistic three-parameter model options. Washington, D.C.: U.S. Civil Service Commission, Personnel Research and Development Center, 1977.

Urry, V.W. ANCILLES: Item parameter estimation program with normal ogive and logistic three-parameter model options. Washington, D.C.: U.S. Civil Service Commission, Personnel research and Development Center, 1978.

Vale, C.D. & Weiss, D.J. A study of computer-administered stradaptive ability testing. (Research Report 75-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric methods programme, October, 1975.

Vernon, P.E. The assessment of psychological scales by verbal methods. National Research Council and Industrial Health Research Board, Report No.58, 1938.

Walker, D.A. Answer-pattern and score scatter in tests and examinations. British Journal of Psychology, 1931, 22, 73-86.

Walker, D.A. Answer-pattern and score scatter in tests and examinations. British Journal of Psychology, 1936, 26, 301-308.

Walker, D.A. Answer-pattern and score scatter in tests and examinations. British Journal of Psychology, 1940, 30, 248-260.

Waller, M.I. Removing the effects of random guessing from latent trait ability estimates. (Research Report 74-32). Princeton, N.J.: Educational Testing Service, 1974.

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

Weiss, D.J. The stratified adaptive computerized ability test. (Research Report 73-3). Minneapolis, Minn.: University of Minnesota, Psychometrics Methods Program, Department of Psychology, 1973.

Wherry, R.J. & Gaylord, R.H. Factor pattern of test items and tests as a function of the correlation coefficient: Content, difficulty, and constant error factors. Psychometrika, 1944, 9, 237-244.

Wherry, R.J., Naylor, J.C., Wherry, R.J. & Fallis, R.F. Generating multiple samples of multivariate data with arbitrary population parameters. Psychometrika, 1963, 30, 303-313.

White, B.W. & Saltz, E. Measurement of reproducibility. Psychological Bulletin, 1957, 54, 81-99.

Willoughby, R.R. The concept of reliability. Psychological Review, 1935, 42, 153-165.

Wittgenstein, L. Philosophical investigations. (2nd ed.). (G.E.M. Anscombe, Trans.). Oxford, Eng.: B. Blackwell, 1958.

Wood, R., Wingersky, M. & Lord, F. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters. (Research Report 76-8). Princeton, N.J.: Educational Testing Service, 1976.

Wright, B.D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.

Wright, B.D. & Douglas, G.A. Best procedures for sample-free item analysis. Applied Psychological Measurement, 1977, 1, 281-295.

Wright, B.D., Mead, R.J. & Bell, S.R. BICAL: Calibrating items with the Rasch model. (Research Report 23-B). Chicago: University of Chicago, Department of Education, Statistical Laboratory, June, 1979.

Wright, B.D. & Panchapakesan, N. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969, 29, 23-48.

Wright, B.D. & Stone, M.H. Best test design. Chicago: MESA Press, 1979.

Appendix A: The claims that unidimensionality indices are synonymous with indices of reliability, internal consistency and homogeneity

The purpose of this appendix is to demonstrate the confusion as to the distinction between the concepts of reliability, internal consistency, homogeneity, and unidimensionality. The tenor of many writings is that reliability, homogeneity and internal consistency are interchangeable terms and indices that have been proposed based on these concepts can be used as decision criteria for assessing dimensionality.

There are six comparisons that can be made between the four concepts.

Reliability and internal consistency.

When writing about internal consistency, Horrocks and Schoonover (1968, p. 72) stated that "this measure of reliability simply shows how consistently the test measured a single population at a single time on a unitary trait or behaviour".

Reliability and homogeneity

Horst (1953) claimed that "Loevinger has insisted, quite rightly, that the Kuder-Richardson reliability formulae are actually estimates of item homogeneity as well as of test reliability".. (I was unable to discover where Loevinger made such an explicit claim.)

Reliability and unidimensionality

Payne (1968, p. 137) stated that "intuitively it would seem that if one were really interested in whether or not the items in a test were measuring the same thing, he should examine the item scores as well as the examinee scores. A reliability estimation procedure of Kuder and Richardson does just this". Armor (1974, p. 23), when discussing a method of maximizing the alpha reliability argues that these methods "are all based on the same basic and self-evident assumption: if a set of items is measuring the same or similar properties and the property comprises a single continuum or dimension, the items should all covary to some extent". Freeman (1962, p. 77) interchanges the Kuder-Richardson methods, homogeneity, and unidimensionality: "The appropriate use of this method (i.e., K-R) requires that all items in the test should be psychologically homogeneous; that is, every item should measure the same factor, or a combination of factors in the same proportion as every other item". Willoughby (1935, p. 164) concluded a discussion on the concept of reliability by stating that "a highly reliable test, therefore, is a test the elements of which are highly intercorrelated; and it turns out also to be a highly valid test in the sense that it is 'unitary' and therefore genuine".

Internal consistency and homogeneity

Remmers, Gage and Rummel (1965, p. 130) claimed that "methods of estimating the internal consistency, or homogeneity, of a test have been presented by Kuder and Richardson", and that alpha "provides some indication of how internally consistent, or homogeneous, a test is for the sample to which it is administered. If the test is not intended to be homogeneous, the coefficient is irrelevant". Cattell and Butchers (1968) in their interpretation of the history of homogeneity, contended that "this extent to which all items are so compressed about the central point of the test, and are therefore homoggneous in what they measure, can be expressed by coefficient alpha". Guilford (1965, pp. 449-450) is more explicit in stating that "we expect that homogeneous tests shall be internally consistent, that is, they should measure the same trait or ability and not attempt to include items uncorrelated with each other".

Internal consistency and unidimensionality

Stanley (1964, pp. 109-110) argued that Kuder and Richardson devised a procedure for estimating the internal consistency of a test, "i.e., the extent to which all items in the test measure the same abilities". Sax (1974, p. 285) stated that "all tests should be internally consistent, that is, they should measure the same trait or ability and not attempt to include items uncorrelated with each other". Gorsuch (1974, p. 192) wrote that "high internal consistency means there is a general factor underlying most of the items in the test".

Homogeneity and unidimensionality

Humphreys (1952, p. 136) contended that "the items of a homogeneous test should measure only one common factor. Measurement of two or more factors makes the test a heterogeneous one". Magnusson (1966, p. 18) defined "homogeneous to mean a variable which gives the extent to which the two conditions discussed, about unidimensionality and freedom from measurement errors, are satisfied."

Gulliksen (1950), probably the most influential proponent of classical test theory, wrote in a chapter called "Reliability estimated from item homogeneity", that Kuder and Richardson developed several methods of assessing the homogeneity of a set of items and presented coefficient alpha as a proper estimate of homogeneity (pp. 220-223). Richardson (1936) did claim that by following the Kuder-Richardson procedures, which includes estimating KR20 and KR21, a test would be made "more pure or homogeneous" (p. 74).

Authors of standards for tests have further confused the definitions. Lyman (1963) defined "content reliability" as evidence that the test items are measuring the same thing and suggested that the use of internal consistency measures such as one of the Kuder-Richardson formulae was appropriate (pp. 32-33). Under the guide lines for evaluating a test, Thorndike and Hagen (1977, pp. 109-110) ask "If the test purports to measure a generalised homogeneous trait, is evidence reported on the internal consistency (inter item or interpart correlations) of the parts that make up the test?" Addressing teachers, Sax (1974, p. 178) argued that "in most cases teachers want to estimate reliabilty from a single administration of a test. This desire has led to measures of internal consistency or homogeneity. The split-half technique and the K-R method are most often used for this purpose". The influential Standards for Educational and Psychological Tests and Manuals (A.P.A., A.E.R.A. and N.C.M.E., 1966) stated that it is essential that "if a test manual suggests that a score is a measure of a generalised homogeneous trait, evidence of internal consistency should be reported". Then they comment that "internal consistency is important if items are viewed as a sample from a relatively homogeneous universe" and it is recommended that split-half measures of the Kuder and Richardson type be used (pp. 30-31).

Appendix B: Indices calculated in the simulation, the programme in which they were calculated, their acronym, and any comments.

No.	Index	Programme	Acronym	Comments
1	Loevinger's H_t	UNIDIM	LOEV	
2	Green RepB	UNIDIM	REPB	
3	Green Consistency	UNIDIM	CONSIS	
4	Alpha	UNIDIM	ALPHA	
5	Mean-alpha	UNIDIM	MNALPHA	
6	KR-21	UNIDIM	KR-21	
7	Horst alpha	UNIDIM	HORST	
8	Mean phi	UNIDIM	MNPHI	
9	No. phi sign.(.05)	UNIDIM	PHI-05	
10	No. phi sign.(.01)	UNIDIM	PHI-01	
11	Mean tetrachoric	UNIDIM	MNTET	
12	No. tet sign.(.05)	UNIDIM	TET-05	
13	No. tet sign.(.01)	UNIDIM	TET-01	
14	Raju alpha	UNIDIM	RAJU	
15	% variance from first factor (or component)[a]	UNIDIM	VARIANCE-pc-phi	
16			VARIANCE-pc-tet	
17			VARIANCE-ml-phi	
18			VARIANCE-ml-tet	
19			VARIANCE-ls-phi	
20			VARIANCE-ls-tet	
21	Eigen 1/Eigen 2	UNIDIM	E1/E2-phi	
22			E1/E2-tet	
23	Green et al. r	UNIDIM	GLM-r-pc-phi	
24			GLM-r-pc-tet	
25			GLM-r-ml-phi	
26			GLM-r-ml-tet	
27	(Eigen 1 - Eigen 2)/ (Eigen 2 - Eigen 3)	UNIDIM	EDIFF-phi	
28			EDIFF-tet	
29	No. Eigen > 1	UNIDIM	EGIN>1-phi	
30			EGIN>1-tet	
31	Eigen/variance	UNIDIM	KELLEY-phi	
32			KELLEY-tet	
33	Theta	UNIDIM	THETA-phi	
34			THETA-tet	
35	Kaiser mean r	UNIDIM	KAISER-phi	
36			KAISER-tet	
37	Sum residuals[b]	UNIDIM	RES-pc-phi	These are normalised
38			RES-pc-tet	residuals - see IMSL
39			RES-ml-phi	
40			RES-ml-tet	
41			RES-ls-phi	
42			RES-ls-tet	
43	Correlation raw & factor scores	UNIDIM	DUBOIS-pc-phi	Bartlett factor scores (see McDonald & Burr, 1967)
44			DUBOIS-pc-tet	

45			DUBOIS-ml-phi	
46			DUBOIS-ml-tet	
47			DUBOIS-ls-phi	
48			DUBOIS-ls-tet	
49	Chi-square	UNIDIM	CHI-phi	
50			CHI-tet	
51	Green et al. u	UNIDIM	GLM-u-ml-phi	
52			GLM-u-ml-tet	
53	Omega	UNIDIM	OMEGA-phi	
54			OMEGA-tet	
55	Tucker and Lewis	UNIDIM	T&L-phi	
56			T&L-tet	
57	Chi-2 - Chi-1	UNIDIM	CH2-CH1-phi	
58			CH2-CH1-tet	
59	No. tet <.16 & >.84	UNIDIM	TETBND	
60	No. zero scores	BICAL	ZEROSC	
61	No. perfect scores	BICAL	PERFSC	
62	No. zero items	BICAL	ZEROIT	
63	No. perfect items	BICAL	PERFIT	
64	KR-20 Before removal	BICAL	KR-20B4	
65	Mean Before removal	BICAL	MEANB4	
66	s.d. Before removal	BICAL	SDB4	
67	No. misfitting people	BICAL	MISFIT	Based on C=2.0 (see Wright, Mead & Bell, 1979, p.15)
68	KR-20 After removal	BICAL	KR-20	
69	Mean After removal	BICAL	MEAN	
70	s.d. After removal	BICAL	S.D.	
71	Error impact	BICAL	ERROR	
72	Between-fit	BICAL	BETWEEN	
73	Total-fit	BICAL	TOTAL	
74	Within-fit	BICAL	WITHIN	
75	Discrimination mean	BICAL	DISCR	
76	Point-biserial mean	BICAL	PBIS	
77	Nonlinear Sum residuals	NONLIN	NONRES	Based on the covariances
78	Nonlinear No.res.>.01	NONLIN	NON>01	
79	NOHARM 10 Sum residuals	NOHARM	NH10RES	Guessing determined from no. in bottom 10% (based on total scores) who passed the item
80	NOHARM 10 No.res.>.01	NOHARM	NH10>01	
81	NOHARM .0 Sum residuals	NOHARM	NH0RES	All guessing input as .0
82	NOHARM .0 No.res.>.01	NOHARM	NH0>01	
83	NOHARM .2 Sum residuals	NOHARM	NH2RES	All guessing input as .2
84	NOHARM .2 No.res.>.01	NOHARM	NH2>01	
85	FADIV Sum residuals	FADIV	FADRES	Based on P-P* (see manual)
86	FADIV No.res.>.01	FADIV	FAD>01	
87	Lord Chi/df	LORD	LORD	A closeness index rather than chi-square as only pairs of items that had non-zero cells were included.

Notes

a. pc = principal components; ml = maximum likelihood; ls = least squares; phi = phi coefficients; tet = tetrachoric coefficients

b. Residuals are the sum of absolute residuals

CURRICULUM VITAE

John Allan Hattie

PERSONAL

Residence:

"Brooklyn"
Thalgarrah
Via Armidale
New South Wales
AUSTRALIA 2350

Date of Birth:	2 February, 1950
Place of Birth:	Timaru, New Zealand
Health:	Excellent
Citizenship:	New Zealand

EDUCATION

Degrees and Certificates

Date	Institution	Field	Degree or Certificate
1974-1976	Ontario Institute for Studies in Education	Measurement and Evaluation	Ph.d course work
1974	University of Otago	Education	Master of Arts First class Honours: Distinction
1972	N.Z. Department of Education	Teaching	Diploma in Teaching
1971	University of Otago	Education	Post-Graduate Diploma in Arts (Credit)
1970	Dunedin Teachers' College (NZ)	Teaching	Diploma of Dunedin Teachers' College (Distinction)
1970	University of Otago	Education	Bachelor in Arts
1963-1966	Timaru Boys' High School		School Certificate University Entrance

Courses

Degree	Department	Course
Ph.d	Measurement & Evaluation	Advanced Statistics & Experimental Design
	Measurement & Evaluation	Advanced Test Theory
	Measurement & Evaluation	Elements of Scaling
	Measurement & Evaluation	Personality Measurement
	Measurement & Evaluation	Factor analysis & Related Techniques
	Measurement & Evaluation	Intermediate Statistics & Research Design
	Measurement & Evaluation	Advanced Measurement & Experimental Design
	Applied Statistics	Design of Experiments
	Applied Statistics	Regression Analysis
	Applied Statistics	Sample Survey Theory (Audit)
	Mathematics	Time Series Analysis (Audit)
	Computer Applications	Simulation models of cognition & learning - Artificial Intelligence (Audit)
	Computer Applications	Research Seminar on Information Processing in Education (Audit)
Post-Graduate Diploma in Arts	Education	Psychology of Education
	Education	Philosophy of Education
	Education	Sociology of Education
	Education	Guidance & Counselling
	Education	Nineteenth Century Education in England
Diploma in Education	Education	Psychology of Adolescence
	Education	Principles of Teaching
	Education	New Zealand Education System
	Education	Administration of Education
Bachelor of Arts	Education	Years 1/2/3
	History	Years 1/2
	Political Studies	Year 1
	Mathematics & Statistics	Year 1
	Modern Pacific History	Year 1

Theses

Master of Arts	Conditions for administering creativity tests.
Diploma in Education	An analysis of Mednicks' Remote Associates Test.

PROFESSIONAL BACKGROUND

Positions

1977- Lecturer, Centre for Behavioural Studies - University of New England, Armidale, Australia

1976-1977 Lecturer in Elements of Statistics - Ontario Institute for Studies in Education, Canada.

1975-1976 Lecturer in Descriptive Statistics - Ontario Institute for Studies in Education, Canada

1974 Teacher of English, Music, and Liberal Studies - Timaru Boys' High School, New Zealand

1974 Part-time Lecturer in Special Education and Educational Tests and Measurement - University of Otago, New Zealand

1973 Senior tutor in Education - University of Otago, New Zealand

1973 Lecturer - New Zealand Physiotherapy School, New Zealand

1972 Teacher at Macandrew Intermediate School, Dunedin, New Zealand

1972 Part-time Lecturer in Educational Tests and Measurement - University of Otago, New Zealand

1971-1972 Tutor in Education - University of Otago, New Zealand

University Teaching

1977- Introduction to Behavioural Research (3rd year course)

Emphasizes the choosing and interpretation of common descriptive and inferential statistics. These statistics form the basis for developing experimental designs. Research studies will be interpreted and evaluated. The abilities developed in the statistics and experimental design sections will be applied to test construction and analysis.

1977- Multivariate Analysis in Educational Research (post-graduate course)

This course introduces multiple regression as a technique encompassing experimental and survey research design. Topics include principal components, discriminant analysis, path analysis, stepwise regression, use of dummy variables, analysis of variance, and factor analysis. The application of these techniques to such problems as the measurement of change, time series, and assessing interaction will be discussed.

1976-1977 Elements of Statistics (Post-graduate course)

A brief review of descriptive statistics, including simple correlation and regression, probability distributions, sampling distributions, expectation and moments, inferences about population parameters, basic principles of experimental design, one-way analysis of variance, and multiple comparison procedures.

1975-1976 Descriptive Statistics (post-graduate course: Audit)

Students enrolling in other courses with no background knowledge of elementary statistics are advised to take this non-credit course.

Research Experience

1980 Prison reform: Prisoners perceptions of the legal system (with G. McGrath)
Health and Longevity: Establishing a physical health logistic equation to motivate health consciousness (with M.P. Barbato)
Meta-analysis of self concept and academic ability (with B.F. Hansford)
Models of processing information and creative thinking (with D. Fitzgerald)
Evaluating the introduction of micro computers in schools (with D. Fitzgerald, G. McGrath, & A. Iakin)

1979 Study skills behaviour (with D. Watkins)
Communication apprehension (with B. Hansford)
Rasch models and reading tests (with J. Irvine)
Factor analytic models and their applications

1978 Conditions for administering creativity tests
Personal Orientation - The psychometric aspects of self actualization
Evaluating client-centred pscyhotherapy outcomes
The difference between factor and component analysis

1977 Scoring and administering creativity tests
Applications of factor analytic models

1976 Secondary Post-secondary interface project (Director: R. Traub, OISE. Research Assistant)

Factor analysis and related techniques (Director: R.P. McDonald, OISE. Research Assistant)
Random and fixed regressor models
Identifiability in factor models

Differences between Harman's and Comrey's factor analyses
General programming and data analysis duties

1974- Open school project: Administering tests, computer programming and data analysis (Director: R, Traub. Research Assistant)

1974- 1976 Drug taking behaviour among adolsecents. (Director: J. Hundleby, University of Guelph. Research Assistant)

Administrative

1979- Chairman: Teaching and Assessment Subcommittee - University of New England. CBSE.
Faculty of Education Safety Officer
1978 Member of Teaching and Assessment Subcommittee - University of New England
1976 Member of Search Committee for MECA Chairperson - OISE
1975- Member of Research and Development Committee - OISE
1976 Member of Departmental Executive Committee - OISE

Affiliations

Psychometrika Society
American Educational Research Association
National Council for Measurement in Education
Canadian Society for the Study of Education
Canadian Educational Researchers' Association

Community Activites

1977- Founding member of Armidale CARELINE - a telephone listening and referral service. As member, management committee, and training officer (Organised and run five courses involving 200 para-professional counsellors)
1977- Founding member of Armidale Search and Rescue Squad
1977 Vice Captain - Bush and Climbing section of Armidale Search and Rescue Squad
1973- 1974 Member of YOUTHLINE (Dunedin, NZ) - a telephone counselling service for young people. As member, management committee, and involved in training.
1972- 1973 Co-leader of wilderness trips, New Zealand

Thesis Supervision

Completed

Irvine, J. The role of play in pre-school behaviour. 1980. Ph.d.

Hancock, P. An evaluation of a retreat programme. 1980. M. Ed. Second class honours (Cosupervised)

Connor, P. Delay of gratification and equity theory. 1980. M. Ed. First class honours.

Dollinson, J. Parent training: Comparison of two approaches. 1979. M. Ed. Second class honours.

Huxley, I. A study of unliked children. 1979. M. Ed. Second class honours. (Co-supervised)

Sharpley, A. Cross-age tutoring. 1978. M. Ed. First class honours.

Sharpley, C. Variability in the reinforcing properties of rewards. 1978. M. Ed. First class honours and university medal.

Holdgate, G. A study of internal assessment and external examinations in secondary education in N.S.W. 1977. M. Ed. Second class honours.

Present

Sharpley, C. Direct and indirect rewards in the classroom. Ph.d.

Klich, Z. Relation between the learning styles of Aboriginal and Australian children. Ph.d. (Co-supervisor)

Ransley, W. Luria's concept of attention. Ph.d.

In-Sub Song Self-concept, environmental processes, learning hierarchies and academic ability. Ph.d.

Morrison, M. Creativity, fluid and crystallized intelligence and their relationship to simultaneous and successive processing. M. Ed.

Commons, A. Developing computer interface equipment to monitor classroom behaviour. M. Ed.

Gay, J. Word frequency counts: Tests for reading readiness. M. Ed. (Co-supervisor)

Dettrick, A. Self-concept in first year classrooms: Do teachers make a difference. M. Ed. (Co-supervisor)

Computer programming

Since arriving in Armidale (1977) many computer programmes have been written and others modified to work on the ICL1901, DEC2060, PDP1134, and PDP1105. These include basic statistical routines, factor analysis programmes, item analysis programmes, graphing, exploratory data analysis, regression analysis, and many one-off programmes. These programmes have all been interfaced and a manual written for users. The programming is primarily in Fortran and Basic.

Others

1980 Advisory editorial consultant *Journal of Educational Psychology*

1975-1976 Member of the editorial board, *Interchange*, Canada

PUBLICATIONS

Published

Hattie, J.A. A four stage factor analytic approach to studying behavioural domains. *Applied Psychological Measurement*, in press.

Hattie, J.A. The Progressive Achievement Tests revisited. *New Zealand Journal of Educational Studies*, in press.

Ozdowski, S.A. & Hattie, J.A. The impact of divorce laws on divorce rate in Australia: A time series analysis. *Australian Journal of Social Issues*, in press.

Watkins, D. & Hattie, J.A. The learning processes of Australian university students: Investigation of contextual and personological factors. *British Journal of Educational Research*, in press.

Hattie, J.A. & Watkins, D. Australian and Filipino investigation of the internal structure of Biggs New study process questionnaire. *British Journal of Educational Psychology*, in press.

Watkins, D. & Hattie, J.A. The internal structure and predictive validity of the Schmeck inventory of learning processes: Some Australian and Filipino data. *Educational and Psychological Measurement*, in press.

Watkins D. & Hattie, J.A. An investigation of the internal structure of the Biggs study process questionnaire. *Educational and Psychological Measurement*, 1980, *40*, 1125-1130.

Watkins, D. & Hattie, J.A. An investigation of the construct validity of three recently developed personality instruments: An application of confirmatory factor analysis to multi-trait, multi-method matrices. *Australian Journal of Psychology*, in press.

Sharpley, C., Irvine, J.W. & Hattie, J.A. Changes in performance of children's handwriting as a result of varying contingency conditions. *Alberta Journal of Educational Research*, 1980, *26*, 183-193.

Hattie, J.A. The place of statistics, research design, and knowledge of computers in the Masters' degree. *Educational Bulletin*, 1980, *1*, 3-5.

Hattie, J.A. Should creativity tests be administered under test like conditions? An empirical study of three alternative conditions. *Journal of Educational Psychology*, 1980, *72*, 85-96.

Hattie, J.A. *Beneath the words*. A one-and-a-half hour video film for training client centred counsellors. Audio-Visual Centre, UNE, 1979.

Hattie, J.A. & Holden, N.F. *A manual to accompany Beneath the Words*. UNE, 1979. pp. 187.

Hattie, J.A. A confirmatory factor analysis of the Progressive Achievement Tests: Reading Comprehension, reading vocabulary, and listening comprehension tests. *New Zealand Journal of Educational Studies*, 1979, *14*, 172-188.

Hattie, J.A. Sex as a moderator in personality inventories: The Personal Orientation Inventory. *Journal of Personality Assessment*, 1979, *43*, 627-628.

Hattie, J.A. *The development of a telephone listening and referral service*. UNE, 1978. pp. 153.

Hattie, J.A. An evaluation of Diploma in Education students' perceptions of their teacher education programme. *Faculty of Education Report*, 1978. pp. 235.

Hattie, J.A. Conditions for administering creativity tests. *Psychological Bulletin*, 1977, *80*, 1249-1260.

Hattie, J.A. Student evaluations of courses and teaching: Uses, problems and value. *Experiences in teaching external students*, 1977, *2*, 5-11.

Hattie, J.A. *Technical Report of the Physics Achievement Test: Secondary Postsecondary Interface Project II: Nature of Students*, Vol. 2. Traub, R.E., Wolfe, R., Wolfe, C., Evans, P., Russell, H.H., and Wayne, K. Toronto: Ministry of Education and Ministry of Colleges and Universities, Ontario, 1976.

Hattie, J.A. Knowledge of the Road Code in New Zealand: A research note. *New Zealand Journal of Educational Studies*, 1976, *11*, 189-190.

Hattie, J.A. Youthline, Dunedin. An evaluation of a telephone counselling service. *New Zealand Medical Journal*, 1975, *82*, 80-81.

Hattie, J.A. An attitudinal study about old people. *New Zealand Medical Journal*, 1975, *82*, 251-254.

Hattie, J.A. Sex stereotyping in secondary schools' literature books. New Zealand Post-Primary Teachers' Association Journal, 1975, October, 34.

Hattie, J.A. Youthline, Dunedin: The first two years. An assessment, recommendations, and predictions. Dunedin, 1974.

Papers presented

Hattie, J.A. & Hansford, B.C. Evaluating the relationship between self and performance/achievement. Paper presented at the Australian Association for Research in Education Conference, Sydney, November, 1980.

Hansford, B. & Hattie, J.A. Meta analysis: A reflection on problems. Paper presented at the Australian Association for Research in Education Conference, Sydney, November, 1980.

Ozdowski, S.A. & Hattie, J.A. An investigation of the effects of divorce laws in Australia. Paper presented at the Sociological Association of Australia and New Zealand, University of Tasmania, August, 1980.

Hattie, J.A. & Fitzgerald, D. A one-day workshop on Multivariance procedures. University of Queensland, May, 1980.

Hattie, J.A. How to conduct a meta-analysis. Invited paper to the Conference on Measurement, University of Queensland, April, 1980.

Hattie, J.A. Some thoughts on establishing a telephone listening and referral service. Invited address to Goulbourn Careline, Goulbourn, March, 1980.

Hattie, J.A. & Hansford, B.C. Evaluating communication apprehension: Implications for teachers and teaching. Paper presented at the Australian Association for Research in Education Conference, Melbourne, November, 1979.

Hattie, J.A. & Watkins, D. Study skills and personality. Paper presented at the Australian Association for Research in Education Conference, Melbourne, November, 1979.

Hattie, J.A. Meta-analysis: The philosophy and methodology. Invited paper to Measurement and evaluation seminar. University of Queensland, Brisbane, October, 1979.

Hansford, B.C., Hattie, J.A., Miller, P.K., & Aveyard, B.C. An assessment of Communication Apprehension in various envixonments. Paper presented at the Australian Conference on Communication, July, 1979.

Hattie, J.A. "Sheer surmise and conjecture, and perhaps wishful thinking": Another look at whether Sir Cyril Burt faked his research on the heritability of intelligence. Paper presented at the Teaching and Assessment Workshop, May, 1979.

Hattie, J.A. & Fitzgerald, D. A use of confirmatory factor analysis in the evaluation of intelligence testing models. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April, 1979.

Hattie, J.A. The first year's experiences in the development of a telephone counselling service. Invited paper presented to Australian Psychological Association (New England), May, 1978.

Bush, P.A. & Hattie, J.A. The reliabilities of a battery of tests of divergent thinking under two conditions of testing. Paper presented to the Symposium on Educational Psychology, Dunedin, New Zealand, October, 1972.

Submitted and in Review

Hattie, J.A. & Swanson, M.S. The Remote Associates Test: An evaluation. New Zealand Psychologist.

Hattie, J.A. & Hansford, B.C. The dimensionality of communication apprehension. Communications Yearbook.

Hattie, J.A. & Fitzgerald, D. Modelling the WISC: Which theory best fits the data. British Journal of Mathematical and Statistical Psychology.

Sharpley, C.F. & Hattie, J.A. The differences and similarities between husbands and wives on the Tennessee Self-Concept Scale. Journal of Counseling Education.

Sharpley, C.F. & Hattie, J.A. A psychometric evaluation of a scale for assessing counseling orientation. Counselor Education and Supervision.

Hattie, J.A. & Klich, M. An evaluation of a test of counselor competence: The Potential Interpersonal Competence Scale. Counselor Education and Supervision.

Hattie, J.A., Olphert, W.B., & Cole, B. The assessment of student teachers, American Educational Research Journal.

Hattie, J.A. & Hansford, B.C. A meta-analysis of self concept and academic ability. Review of Educational Research.

Hansford, B.C. & Hattie, J.A. Problems in conducting a meta-analysis. American Psychologist.

In preparation

Watkins, D. & Hattie, J.A. The mismatch between students as to their motive and strategy for studying.

Fitzgerald, D. & Hattie, J.A. Laterality and creativity.

Hattie, J.A. & Hattie, J.A. Do the Torrance Tests of Creative Thinking measure similar latent abilities as the Torrance measure of "Your style of learning and thinking"?

Nolan, J. & Hattie, J.A. Teaching parents to modify their child's behaviour: A time series analysis.

Hattie, J.A. & Hancock, P.A. The relationship between the Personal Orientation Inventory and the Personal Orientation Dimensions.

Hattie, J.A. Test validity or group discrimination: Comparing new and old samples to norm samples.

Hattie, J.A. The relationship between creativity and intelligence: Using restricted factor analysis with constrained uniquenesses.

Hattie, J.A. & Irvine, J. A restandardization of the GAP reading test using the Rasch model.

Irvine, J. & Hattie, J.A. A restandardization of the St. Lucia reading test using the Rasch model.

Books (In preparation)

Hansford, B.C. & Hattie, J.A. Self concept and academic ability.

Law, H.G., Synder, C.W., McDonald, R.P. & Hattie, J.A. Multi-mode models for data analysis.

McGrath, G., Chapman, R. & Hattie, J.A. The prisoner's perceptions of the legal system.

GENERAL INTERESTS

Hockey, Tennis, Golf, Hiking, Climbing, and Skiing.

END

01 03 82

FIN