Methodology Review: Assessing Unidimensionality of Tests and Items

John Hattie
University of New England, Australia

Various methods for determining unidimensionality are reviewed and the rationale of these methods is assessed. Indices based on answer patterns, reliability, components and factor analysis, and latent traits are reviewed. It is shown that many of the indices lack a rationale, and that many are adjustments of a previous index to take into account some criticisms of it. After reviewing many indices, it is suggested that those based on the size of residuals after fitting a two- or three-parameter latent trait model may be the most useful to detect unidimensionality. An attempt is made to clarify the term unidimensional, and it is shown how it differs from other terms often used interchangeably such as reliability, internal consistency, and homogeneity. Reliability is defined as the ratio of true score variance to observed score variance. Internal consistency denotes a group of methods that are intended to estimate reliability, are based on the variances and the 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. The usefulness of the terms internal consistency and homogeneity is questioned. Unidimensionality is defined as the existence of one latent trait underlying the data.

One of the most critical and basic assumptions of measurement theory is that a set of items forming an instrument all measure just one thing in common. This assumption provides the basis of most mathematical measurement models. Further, 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, the variable must be unidimensional; that is, the various items must measure the *same* ability, achievement, attitude, or other psychological variable. As an example, it seems desirable that a test of mathematical ability be not confounded by varying levels of verbal ability on the part of persons completing the test. Yet despite its importance, there is not an accepted and effective index of the unidimensionality of a set of items. Lord (1980) contended 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 under a latent trait model.

Initially identified as a desirable property of tests in the 1940s and 1950s (the earliest mention was in Walker, 1931), unidimensionality was used in the same way as was homogeneity and internal consistency until the recent increase of interest in latent trait models constrained a clearer and more precise definition. One aim of this paper is to make the boundaries of the terms less fuzzy by detailing variations in the use of the terms and by proposing definitions that seem consistent with usage by many authors (or at least, what the authors appear to have intended when they used the terms). A further aim is to review the many indices that have at various

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 9, No. 2, June 1985, pp. 139–164
© Copyright 1985 Applied Psychological Measurement Inc.
0146-6216/85/020139-26$2.55

times been proposed for unidimensionality, with particular emphasis on their rationale. It is argued that most proposers do not offer a rationale for their choice of index, even fewer assess the performance of the index relative to other indices, and hardly anyone has tested the indices using data of known dimensionality.

As a working definition of unidimensionality, a set of items can be said to be unidimensional when it is possible to find a vector of values $\boldsymbol{\phi} = (\phi_i)$ such that the probability of correctly answering an item g is $\pi_{ig} = f_g(\phi_i)$ and local independence holds for each value of $\boldsymbol{\phi}$. The first section, on methods based on answer patterns, assumes f_g to be a step function. The second section, on methods based on reliability, assumes f_g to be linear. The next two sections, on methods based on principal components and factor analysis, assume f_g to be generally linear (though a nonlinear function is considered). The final section, on methods based on latent traits, assumes different forms for f_g.

Each section details various indices that have been proposed for unidimensionality. Given the large number of indices, it is not possible to discuss each one fully or to cite every researcher who has used it. Table 1 lists the indices discussed in this paper and cites references to the authors who have used or recommended each index.

Indices Based on Answer Patterns

Several indices of unidimensionality are based on the idea that a perfectly unidimensional test is a function of the amount by which a set of item responses deviates from the ideal scale pattern. The ideal scale pattern occurs when a total test score equal to n is composed of correct answers to the n easiest questions, and thereafter of correct answers to no other questions.

Guttman (1944, 1950) proposed a reproducibility coefficient that provides "a simple method for testing a series of qualitative items for unidimensionality" (1950, p. 46). This index of reproducibility is a function of the number of errors; these errors 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 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. Guttman has suggested reproducibility indices of .80 (1944) and .90 (1945) as acceptable approximations to a perfect scale.

Guttman's coefficient has a lower bound that is a function of the item difficulties and 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 formulas for dichotomous items and reported an average discrepancy between his and Guttman's formula of .002 (see Nishisato, 1980).

Loevinger (1944, 1947, 1948) argued that tests of ability are based on two assumptions: (1) scores at different points of the scale reflect different levels of the same ability, and (2) for any two items in the same test, the 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 (1944) developed what she termed an index of homogeneity. When all the test items are arranged in order of increasing difficulty, the proportion of examinees who have answered correctly both items, P_{ij}, is calculated for all pairs of items. From this, the theoretical proportion, (P_iP_j), who would have answered correctly both items had they been independent, is subtracted. These differences are summed over the $n(n - 1)$ pairs of items:

$$S = \sum\sum(P_{ij} - P_iP_j) \quad . \tag{1}$$

Table 1
Indices and Researchers Who Have
Recommended Each Index

Indices Based On Answer Patterns	
Green RepB	Green (1956), White & Saltz (1957)
Consistency	Green (1956), White & Saltz (1957)
Loevinger's index	Loevinger (1944), Gage & Damrin (1950), Lumsden (1959, 1961), Hoffman (1975)
Indices Based On Reliability	
Alpha	Gage & Damrin (1950), Freeman (1962), Guilford (1965), Horrocks & Schoonover (1968), Lumsden (1959), Payne (1968), Ryan (1979)
Horst's alpha	Horst (1953), Lumsden (1959)
Raju's alpha	Terwilliger & Lele (1979), Raju (1980)
Mean item correlation	Cronbach (1951), Cattell & Tsujioka (1964), Humphreys (1956), Cattell (1978)
Mean correlation	Cronbach (1951), Kaiser (1968), Silverstein (1980)
No. zero correlations	Armor (1974), Mosier (1936), Humphreys (1952)
Item-test correlation	Mosier (1940), Kelley (1942), Nunnally (1970)
KR-21	Kuder & Richardson (1937), Gage & Damrin (1950), Raju (1980)
Indices Based On Principal Components	
Percent variance	Cattell & Tsujioka (1964), Carmines & Zeller (1979), Hambleton & Traub (1973), Hutten (1979), Reckase (1979)
Number of eigenvalues > 1	Armor (1974), Laforge (1965), Koch & Reckase (1979), Birenbaum & Tatsuoka (1982)
Eigenvalue 1/Eigenvalue 2	Bentler (1972), Lumsden (1957, 1961), Hambleton (1980), Hutten (1980), Lord (1980)

(continued on next page)

Table 1, continued
Indices and Researchers Who Have Recommended Each Index

(Eigenvalue 1 − Eigenvalue 2)/ (Eigenvalue 2 − Eigenvalue 3)	Divgi (1980)
Eigenvalue/variance	Kelley (1942)
Sum of residuals	Lumsden (1959), McDonald (1981)
No. residuals > .01	Lumsden (1959)
Correlation of raw & factor scores	Dubois (1970)
Indices Based on Factor Analysis	
Chi-square (1 factor)	Bock & Lieberman (1970)
Chi-square (2 factor − 1 factor)	Jöreskog (1978)
Tucker & Lewis	Tucker & Lewis (1973)
Theta	Armor (1974), Carmines & Zeller (1979), Bentler (1972), Greene & Carmines (1980)
Omega	Armor (1974), Heise & Bohrnstedt (1970), Carmines & Zeller (1979), Smith (1974a, 1974b), Greene & Carmines (1980)
Communalities	Green, Lissitz & Mulaik (1977), Hattie & Hansford (1982), Watkins & Hattie (1980)
Nonlinear factor analysis	McDonald (1967b, 1981, 1982), Etezadi (1981), Lam (1980), McDonald & Ahlawat (1974)
Indices Based On Latent Trait Models	
Christoffersson-Muthén	Christoffersson (1975), Muthén (1978, 1981), Lord (1980), Muthén & Christoffersson (1981)
Two-parameter latent trait model	McDonald (1981), McDonald & Fraser (1985)
One-parameter latent trait model	Andrich & Godfrey (1978–1979), Reckase (1979), Rentz & Rentz (1979), Wright (1977), Wright, Mead, & Bell (1979), Wright & Stone (1979)
Lord's chi-square	Lord (1953), Hambleton, Swaminathan, Cook, Eignor, & Gifford (1978)

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 answering correctly the more difficult item in each pair (P_j):

$$S_{\max} = \sum\sum(P_j - P_iP_j) \quad . \qquad (2)$$

The index of homogeneity thus proposed by Loevinger (1944) was the ratio of the Equations 1 and 2 ($H = S/S_{\max}$). The coefficient is unity for a perfectly homogeneous test and zero for a perfectly nonhomogeneous test. Yet for some sets of items, such as those that allow guessing, the lower bound may not necessarily be zero. (Incidentally, Loevinger's method of identifying errors is the same as Green's (1956), but Green sums over those item pairs whose members are adjacent in difficulty level and not over all pairs of items, as does Loevinger.)

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 (i.e., a perfect scale) is made (Lumsden, 1959). Another major criticism is that there is nothing in the methods that enables a test of just one trait to be distinguished from a test composed of an equally weighted composite of abilities. Loevinger (1944), who recognized 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. Guilford (1965) argued that it is possible to say that a test that measures, say, two abilities may be considered a unidimensional test provided that each and every item measures both abilities. An example is an arithmetic-reasoning test or a figure-analogies test.

A further objection is that it is possible to construct a set of items that forms a perfect scale yet would appear not to be unidimensional. For example, consider a set of 10 items testing different abilities, one item at a level of difficulty appropriate to each grade from 1 through 10. The test is given to a group of 10 children, each of which is an average student at each grade level from 1 to 10. A perfect scale very probably would result. Thus, perfect reproducibility may not necessarily imply that the items are unidimensional. It seems that using these examples, as have Loevinger (1947) and Humphreys (1949) among others, confuses the method of assessing dimensionality with the identification of the dimension(s) measured. It is possible to say that the items in the above example measure one characteristic, which could be labeled development. Saying that a test is unidimensional does not identify that dimension, in the same way that saying that a test is reliable does not determine what it is reliably measuring.

Because of these criticisms, there was a decline in the use of scaling methods to index unidimensionality. Over the past decade the indices have reemerged under different names. Loevinger's (1944) H has been termed c_{t3} by Cliff (1977) and Reynolds (1981). A number of simulations (Wise, 1982, 1983) have compared various reproducibility indices (or order statistics as they are now called) and rediscovered many of the above problems. Wise (1983), for example, compared Loevinger's H and c_{i1} described by Cliff (1977) and Reynolds (1981) using six simulated and three real data sets. He found that c_{t1} was a poor index of dimensionality for data sets in which there were items of similar difficulty and in which items were of disparate difficulty but did not belong to the same factor. Loevinger's (1944) index also did not perform well in data sets in which there were correlations between factors. In these cases, Loevinger's index tended to overestimate the dimensionality and was not able to distinguish between items loading on different factors.

Indices Based on Reliability

Perhaps the most widely used index of unidimensionality has been coefficient alpha (or KR-20). In his description of alpha Cronbach (1951) 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 stated that ''for a test to be interpretable, however, it is not essential that all items be factorially similar. What is required is that a large proportion of the test variance should be attributable to the principal factor running through the test'' (p. 320). He then stated that alpha ''estimates the proportion of the test variance due to all common factors among the items,'' and indicates ''how much the test score depends upon the general and group, rather than on the item specific, factors'' (p. 320). Cronbach claimed that this was true provided that the inter-item correlation matrix was of unit rank, otherwise alpha was an underestimate of the common factor variance. (The rank of a matrix is the number of positive, nonzero characteristic roots of the correlation matrix.) However, Cronbach suggested that this underestimation was not serious unless the test contained distinct clusters (whereas Green, Lissitz, & Mulaik, 1977, demonstrated that with distinct clusters alpha can be overestimated). These statements led Cronbach, and subsequently many other researchers, to make the further claim that a high alpha indicated a ''high first factor saturation'' (p. 330), or that alpha was an index of ''common factor concentration'' (p. 331), and the implication was that alpha was related to dimensionality.

Underlying the intention to use alpha as an index of unidimensionality is the belief that it is somehow related to the rank of a matrix of item intercorrelations. Lumsden (1957) argued that a necessary condition for a unidimensional test (i.e., ''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). He means that the matrix fits the Spearman case of the common factor model. From a more systematic set of proofs, Novick and Lewis (1967) established that alpha is less than or equal to the reliability of the test, and equality occurs when the off-diagonal elements of the variance-covariance matrix of observed scores are all equal. This implies the weaker property of unit rank. (However, the converse is not true, i.e., unit rank does not imply that the off-diagonals are all equal.) Thus, there exists a diagonal matrix that, on being subtracted 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. Alpha is not a monotonic function of unidimensionality.

Green et al. (1977) observed that though 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. 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. The factors were also well determined with four items having high loadings on each factor. Each item had a communality of .90. Green et al. calculated alpha to be .811, 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, it can be determined that 15 of the 45 distinct inter-item correlations are zero. This should be a cause for concern.

In a monte carlo simulation, Green et al. (1977) found: (1) that alpha increases as the number of items *(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 concluded that the chief defect of alpha as an index of dimensionality is its tendency to increase as the number of items increase.

Green et al. (1977) were careful to note that Cronbach realized the effect of the number of items and that he recommended the average inter-item correlation. Yet in their monte carlo study, Green et al. found that the average inter-item correlation is unduly influenced by the communalities of the items and by negative inter-item correlations.

Index/Index-Max Formulas

Because alpha is based on product-moment correlations, it cannot attain unity unless the items are all of equal difficulty. This limitation has led other authors to modify alpha in various ways, and to propose other indices of unidimensionality. Generally, these modifications have involved dividing the index by its maximum value (index/index-max). Loevinger's (1944) *H* is one such ratio. Horst (1953), who reconceptualized Loevinger's index in terms of intercorrelations between items, 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 the item difficulties. The problem in such an index is to obtain plausible estimates of item reliability.

Both Loevinger (1944) and Horst (1953) defined 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 maximized the item covariances when test variances were fixed but the item difficulties were allowed to vary. (Thus, these indices are not estimates of the classical definition of reliability, i.e., as the ratio of true score variance to observed score variance.)

Although his work predated these criticisms, Cronbach (1951) was aware of the problems of alpha, and reported that the difference between alpha and indices controlling for the dispersion of item difficulties was not practically important.

Indices Based on Correcting for the Number of Items

Alpha is dependent on the length of the test, and this leads to a new problem—and to more indices. Conceptually, the unidimensionality 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 this mean correlation is based on the case in which the items have equal variances and equal covariances, and then is applied more generally. Cronbach recommended the formula as an overall index of internal consistency. If the mean correlation 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 correlation could be indicative of a nonhomogeneous test and recommended that when the mean correlation is low, only a study of the correlations among items would indicate whether a test could be broken into more homogeneous subtests.

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 since it contained all information needed to decide on the dimensionality (also see Mosier, 1936). 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.

Another problem in using an average inter-item correlation is that if the distribution of the correlations is 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 (Mosier, 1940). Clearly, despite its common usage as an index of unidimensionality, alpha is extremely suspect.

Indices Based on Principal Components

Defining a unidimensional test 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.

Since the first principal component explains the maximum variance, then this variance, usually expressed as the percentage 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 the set of items is to being unidimensional. An obvious problem is how "high" does this variance need to be before concluding that a test is unidimensional. Carmines and Zeller (1979), without any rationale, contended that at least 40% of the total variance should be accounted for by the first component before it can be said that a set of items is measuring a single dimension. Reckase (1979) recommended that the first component should account for at least 20% 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 does a unidimensional test.

Since 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." A common strategy is to retain only those components with eigenvalues greater than 1.0 (see Kaiser, 1970, for a justification, though there are many critics of his argument, e.g., Gorsuch, 1974; Horn, 1965, 1969; Linn, 1968; Tucker, Koopman, & Linn, 1969). The number of eigenvalues greater than 1.0 has been used as an index of unidimensionality.

Lumsden (1957, 1961), without giving reasons, suggested that the ratio of the first and second eigenvalues would give a reasonable index of unidimensionality, though he realized that besides having no fixed maximum value, little is known about the extent to which such an index may be affected by errors of sampling or measurement. Hutten (1980) also assessed unidimensionality on the basis of the ratio of the first and second largest eigenvalues of matrices of tetrachoric correlations. Without citing evidence, Hutten wrote 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).

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 eigenvalue is not much larger than any of the others. A possible index to operationalize 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, then 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 1.0, this index will be small. It is not difficult, however, to construct cases when 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 identify the four-factor case as unidimensional, but not the three-factor case!

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 the test is unidimensional. It has been suggested 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 $(N - 1)^{1/2}$, 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 components analysis primarily in that it estimates a uniqueness for each item given a hypothesis as to the number of factors.

There are other differences between the two methods and Hattie (1979, 1980) 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. The Fuller and Hemmerle 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 for the present purpose the use of the chi-square applied to the matrix of tetrachorics from binary data. (However, 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. Jöreskog (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. Given the important effect of sample size in determining the chi-square values in factor analysis, the chi-square values are commonly very large relative to degrees of freedom. Consequently, even trivial departures lead to rejection of the hypothesis. Jöreskog and Sörbom (1981) suggested that the chi-square measures are better thought of as goodness-of-fit measures rather than as test statistics.

Instead of using chi-squares, McDonald (1982) 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, is surely false and will be proved so by a significant chi-square if only the sample size is made sufficiently large. If the residuals are small the fit of the hypothesis can still be judged to be 'satisfactory' " (p. 385). This raises the issue that it is probably more meaningful to ask the degree to which a set of items departs from unidimensionality than to ask whether a set of items *is* unidimensional.

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 from which to conclude that a set of items is unidimensional. It is also not clear why the authors did not condone using the statistic to indicate how well the hypothesized number of factors explains the interrelationships, yet they did claim that it summarizes the quality of interrelationships (see Tucker & Lewis, 1973, p. 9).

Maximizing Reliability

An index based on the largest eigenvalue (λ_1), that has been often rediscovered is maximized-alpha (Armor, 1974; Lord, 1958):

$$\text{maximized-alpha} = [n/(n-1)](1 - 1/\lambda_1) \quad . \qquad (3)$$

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, since alpha and maximized-alpha do not differ by much when only high loadings 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 unidimensionality, though 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. Armor argued that maximized-alpha was better than alpha, but in his examples, the differences between these two coefficients were .05, .02, .002, and 0.0, none of which justifies his claim of ''substantial'' differences.

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 specificities and these can be satisfactorily estimated only by fitting the common factor model efficiently. Approximations from components analysis would generally yield underestimates of the uniquenesses (and therefore of the specificities) and thus lead to spuriously high values of omega.

Should Factor Analysis Be Used on Binary Items?

Many of the tests for which indices of unidimensionality have been derived are scored correct or incorrect. A problem in using the usual factor model on binary-scored items is the presence of ''difficulty factors.'' This problem has a 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: (1) it may have something to do with the choice of a measure of association; (2) it may be that difficulty factors are related to distinct human abilities; or (3) it may have something to do with chance or guessing. From each account various indices of unidimensionality have been proposed.

Choice of a measure of association. Wherry and Gaylord (1944) argued that difficulty factors were obtained because phi and not tetrachoric correlations were used. This was, they argued, because tetrachorics would be 1.0 in cases of items measuring the same ability regardless of differences in difficulty, whereas the sizes of phis are contingent upon difficulty. They contended that if difficulty factors were found even when tetrachorics were used (as in Guilford's, 1941, case), then this must be considered disproof of the unidimensional claim. Empirically, the sample matrix of tetrachorics is often not positive-definite (i.e., non-Gramian). This may be a problem of using (perhaps without justification) a maximum likelihood computer program 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.

Carroll (1945) has further 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 was guessing.

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 led to an important discovery by Gibson (1959, 1960) who recognized that difficulty factors were caused by nonlinear regressions of tests on factors. However, a general treatment of nonlinearity was not given until McDonald's series of publications (1962a, 1965a, 1965b, 1967a, 1967b, 1967c, 1976).

Nonlinear factor analysis. McDonald (1965a, 1967c) 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. A major consequence of using linear factor analysis on binary items is to distort the loadings of the very easy and very difficult items and to make it appear that such items do not measure the same underlying dimension as the other items.

McDonald and Ahlawat (1974), in a monte carlo study, convincingly demonstrated (1) that if the regressions of the items on the latent traits are linear, there are no spurious factors; (2) that a factor whose loadings are correlated with difficulty need not be spurious; (3) 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 (4) 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 nonlinearity."

McDonald (1979) described a method of nonlinear factor analysis using a fixed factor score model which, unlike the earlier random model (McDonald, 1967c), obtains estimates of the parameters of the model by minimizing a loss of function based either on a likelihood ratio or on a least squares function. Etezadi and McDonald (1983) have investigated numerical methods for the multifactor cubic case of this method with first-order interactions.

Although 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 (1967c), McDonald and Ahlawat (1974), and Etezadi (1981), 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 nonlinear factor analysis specifying 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.

Communality Indices

Green et al. (1977) have suggested two indices based on communalities. The first they called u,

$$u = \sum_{i \neq j}\sum |\bar{r}_{ij}| / \sum_{i \neq j}\sum (h_i^2\, h_j^2)^{1/2} \quad , \qquad (4)$$

where h^2 is a communality from a principal components analysis. Unlike many other proposers of indices, Green et al. did offer a rationale for these indices. When there is a single common factor among the items, the loadings on this factor equal the square roots 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. suggested that the inequality $|\bar{r}_{ij}| < (h_i^2\ h_j^2)^{1/2}$ 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. (1977) calculated the value of u in a monte carlo study and found u to be relatively independent of both the number of items and the communality of the items. Although u did increase as the number of factors loading on an item increased, Green et al. reported that u did not increase as much as alpha. From an inspection of their summary statistics it appears that, contrary to the claim of Green et al., u is affected as much as alpha, though the values are much lower than alpha. Further, although u does seem to distinguish between a one-factor solution and a more-than-one-factor solution, it does not relate in any systematic way to the number of extra factors involved.

A serious problem with u is that it requires knowledge of the communalities of the items, which depends on knowledge of the correct dimensionality. In the simulation of Green et al. (1977), they provided the communalities, but in practice these are not known and u can be (and nearly always is) larger than 1, as $|\bar{r}_{ij}| > (h_i^2\ h_j^2)^{1/2}$ (because the communalities are usually underestimated). The usefulness of u is therefore most questionable.

Green et al. (1977) suggested a further index. Given that r 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 a correction for attenuation would be calculated, and the resulting values averaged, then an index not affected by the communalities is obtained. Green et al. contended that such an index also takes on values 0 to 1, with 1 indicating unidimensionality. This index also depends on knowledge of the communalities and is affected by the number of factors determining a variable.

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:

$G + V + Sv + E$ (e.g., verbal analogy = v),
$G + N + Sn + E$ (e.g., number series = n),
$G + K + Sk + E$ (e.g., matrix completion = k),
$G + M + Sm + E$ (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" (1957, pp. 107–108).

It can be argued that these items may 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, 1972). The hypothesis of one second-order factor could not be rejected in 9 of the 11 data sets and Hattie concluded that "It thus seems that there is a unidimensional construct underlying the major factors of the POI" (p. 80), which could be identified as self-actualization. Any method for assessing dimensionality does not necessarily identify the nature of the unidimensional set. The naming of the dimension is an independent task.

Indices Based on Latent Trait Models

Before outlining indices based on latent trait models, it is necessary to present some basic theory. This basic theory also provides a clear definition for the concept of unidimensionality.

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 can be denoted by $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$. The vector $\boldsymbol{\theta}$ 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 θ 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:

$$P(\theta) = c + \frac{1 - c}{(1 + e^{-da(\theta - b)})} \quad , \tag{5}$$

or alternatively by a normal ogive function

$$P(\theta) = c + (1 - c) \int_{-\infty}^{a(\theta - b)} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \, dz. \tag{6}$$

The difference between Equations 5 and 6 is less than .01 for every set of parameter values when d is chosen as 1.7 (Haley, 1952).

The parameter a in Equations 5 and 6 represents the discriminating power of the item, the degree to which the item response varies with θ level. Parameter b is a location or difficulty parameter and determines the position of the item characteristic curve along the θ scale. The parameter c is the height of the lower asymptote, the probability that a person infinitely low on θ will answer the item correctly. It is called the guessing or pseudochance level. If an item cannot be answered correctly by guessing, then $c = 0$. 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.

The most critical and fundamental assumption of the latent trait models is that of local independence (Anderson, 1959; McDonald, 1962b, 1981). Lord (1953) has stated that this is almost indispensable for any theory of measurement. The principle of local independence requires that any two items be uncorrelated when θ is fixed and does *not* require that items be uncorrelated over groups in which θ varies. Lord and Novick (1968) gave the definition of local independence more substantive meaning by writing 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 persons' value on the trait is determined, the behavior is random, in the sense of statistical independence. (p. 538)

McDonald (1962b) argued that the statement of the principle of local independence contained 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 the principle of local independence is rejected in favor of some less restrictive principle then it is not possible to retain the definition of latent traits, since it is by that principle that latent traits are defined. McDonald pointed out, however, that it is possible to 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 (as have 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 mutually statistically independent. It holds for two latent traits, yet such a case would not be considered unidimensional, that is, measuring the same construct. *Unidimensionality can be rigorously defined as the existence of one latent trait underlying the set of items.*

At this point the working definition of unidimensionality presented in the opening paragraph can be clarified. It was claimed that a set of items can be said to be unidimensional when it is possible to find a vector of values $\phi = (\phi_i)$ such that the probability of correctly answering an item g is $\pi_{ig} = f_g(\phi_i)$ and local independence holds for each value of ϕ. This definition is not equating unidimensionality with local independence, because it can further require that it is necessary to condition only on one θ dimension and that the probabilities π_{ig} can be expressed in terms of only one θ dimension.

Lord (1953) 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. First, 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 $df = df_1 + df_2 + . . . + df_n$ degrees of freedom), the resulting combined chi-square may be tested for significance. If the combined chi-square is significant, then Lord argued it must be considered that the test is not unidimensional. When assessing how this statistic behaved (for use in Hattie, 1984a), it soon became obvious that it does not matter whether a chi-square or an exact test is used, since in both cases the probability is nearly always close to 1.0.

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. 225, 544–545). 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 that can be said about the latent traits, though strictly it weakens their definition.

Indices Based On the One-Parameter Model

Using a combination of difficulty, discrimination, and guessing, various latent trait models have been proposed. 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 (1) that there is no guessing, (2) that the discrimination parameter a is constant, and (3) that the principle of local independence applies.

One of the major advantages of the Rasch model often cited is that there are many indices of how adequately the data "fits" the model. In one of the earliest statements, Wright and Panchapakesan (1969) contended 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" (p. 25). These sentiments have often been requoted and an earnest effort made to find and delete misfitting items and people. Rentz and Rentz (1979) stated 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).

From the one-parameter model there has been a vast array of indices suggested (e.g., see Table 2), but they appear to lack theoretical bases, little is known of their sampling distributions, and they seem to be ad hoc attempts to understand the behavior of items. Their developers have paid insufficient attention to helping psychometricians understand the assumptions, the methods of calculation, and the justification for them (see Rogers, 1984 and below, for a detailed account of the derivations and behavior of these indices). 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). The tests are motivated primarily by a desire to assess the various assumptions, particularly the unidimensionality assumption.

To illustrate the various "fit statistics," consider the following indices that are typically used. The Between-Fit t has been called the "most natural" in that it "is derived directly from the 'sample-free' requirements of the model" (Wright, Mead,

Table 2
Some Fit Statistics Based On the One-Parameter IRT Model Commonly Used to Detect Departures From Assumptions (Including Unidimensionality)

Name	Formula	
Mean-Squared Residual (summed over persons or items)	$\sum\{(X_{ij} - P_{ij})/[P_{ij}(1 - P_{ij})]^{1/2}\}^2$	(7)
Total-t (summed over person or items)	$\sum(x_{ij} - P_{ij})^2/\sum P_{ij}(1 - P_{ij})$	(8)
Between-t for items	$\sum_{r}^{m}[(x_{rj} - \sum_{i \varepsilon r} n_i P_{ij})^2/\sum_{i \varepsilon r} n_i P_{ij}\,(1 - P_{ij})]k/(m - 1)(k - 1)$	(9)
Yen's Q_1	$\sum_{m}^{10}\{n_m(O_{mj} - En_{mj})^2/[E_{mj}\,(1 - E_{mj})]\}$	(10)
van den Wollenberg Q_1	$\sum^{k}\{\sum^{k-1}[n_r(n_{rj} - En_{rj})^2]/[En_{rj}(n_r - E_{rj})]\}$	(11)

X_{ij} = observed response of person i to item j
P_{ij} = probability of a correct response for person i to item j, obtained from the model equation using estimates of the person and item parameters
N = number of persons
m = number of score groups (usually 6)
O_{ij} = observed proportion of examinees in group m who answer correctly item j
E_{ij} = expected proportion of examinees in group m who answer correctly item j
n_{rj} = number of correct responses at each score level K for each item j
En_{rj} = expected number of correct responses at each score level k for each item j

& Bell, 1979, p. 10). The statistic, however, is sample-dependent. The sample is divided into subgroups based on score level according to estimated θ, and then, the observed proportion of successes on each item in each estimated θ subgroup is compared with that predicted from the estimates of the item difficulties given by the total sample. Wright et al. derived a standardized mean-square statistic that has an expected value of 0 and variance of 1. 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. (1979) noted that any *t* value larger than 1.5 ought to be examined for response irregularities and that 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.

There have been critics of the claims made by the Rasch proponents. Hambleton (1969) included five items in 15-, 30-, and 45-item simulations that were constructed to measure a second ability orthogonal to the first ability. (He did 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)

Gustafsson and Lindblad (1978) found that tests of fit, like those used by Wright et al. (1979), 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).

Van den Wollenberg (1982a, 1982b) has demonstrated that many of the commonly used indices of unidimensionality based on the Rasch model are insensitive to violations of the unidimensionality axiom. Instead he proposed two new indices, one of which (his Q_1) is easy to compute but seems to lack sensitivity, and the other (Q_2) requires much computer time to calculate but does seem to be sensitive to violations of unidimensionality under some conditions. Van den Wollenberg discussed the conditions when Q_2 seems to be useful, and he promised more systematic inquiries of Q_2. Q_2 is, however, more a global test of the fit of the one-parameter model rather than a specific index of unidimensionality. If Q_2 is large relative to its expected value (yet to be accurately determined), it could be because of violations of other assumptions such as equal discriminations and/or no guessing.

Rogers (1984) investigated the performance of the Between-*t* and Total-*t* for persons and items and the Mean-Square Residual for persons and items, along with many other one-parameter indices to detect unidimensionality (including all those listed in Table 2). She varied test length, sample size, dimensionality, discrimination, and guessing. Rogers used the programs BICAL (Wright et al. 1979) and NOHARM (a method that fits the unidimensional and multidimensional normal ogive one-, two-, and three-parameter models by finding the best approximation to a normal ogive in terms of the Hermite-Tchebycheff polynomial series, using harmonic analysis; see Fraser, 1981; McDonald, 1982, for details). When either BICAL or NOHARM one-parameter methods were used, all indices were insensitive to multidimensionality, except for Total-*t* for persons and Mean-Square for items when the dimensions were close to orthogonal. Rogers concluded that the inability of all indices "to detect multidimensionality under the one-parameter model restricts their use to conditions where unidimensionality is assured" (p. 92).

Indices Based On the Two-Parameter Model

The two-parameter model allows for estimation of difficulty and discrimination and assumes zero guessing. Bock and Lieberman (1970) detailed a method of estimating these two parameters by considering the pattern of a person's item responses across all items in the test. When compared to the more usual method of factoring tetrachoric correlations, Bock and Lieberman reported trivial differences in the estimates of item difficulties and discrimination, but there were differences in the assessment of dimensionality. Using two sets of data, they reported a sharp break between the size of the first eigenvalue and that of the remaining eigenvalues, and they concluded that "these results would ordinarily be taken to support the assumption of unidimensionality" (p. 191). The usual chi-square tests of one factor from a maximum likelihood factor analysis were 25.88 and 53.74, each with 5 degrees of freedom for the two data sets employed.

Clearly in both cases, a one-factor solution must be rejected. The chi-square approximations based on Bock and Lieberman's (1970) two-parameter solution were 21.28 and 31.59, each based on 21 degrees of freedom. In these cases there is evidence for a unidimensional solution. Bock and Lieberman concluded that the "test of unidimensionality provided by maximum likelihood factor analysis cannot be relied upon when tetrachoric correlations are used" instead of their estimates (p. 191).

A major obstacle to using the Bock and Lieberman (1970) method is a practical one. The method requires the use of 2^n possible response patterns across all items which, for example, would require 32,768 possible patterns for a 15-item test, and more than 107 million response patterns for a 30-item test.

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 (1970) approach. Muthén (1978) developed an estimation method that was computationally faster than Christoffersson's. These methods solve many of the computation problems, and the resulting indices of fit, after fitting one factor (based on the size of residuals), are most promising in that they are based on sound theoretical considerations.

Bock and Aitkin (1981) pointed out the practical limitations of the Bock and Lieberman (1970) work, but stated that from a statistical point of view the Christoffersson (1975) and Muthén (1978) method "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. 444). Instead, Bock and Aitkin proposed to estimate the item parameters by integrating over the empirical distribution. The Bock and Aitkin method was applied to the same data as was the Bock and Lieberman, Christoffersson, and Muthén methods (see Bock & Aitkin, 1981). The differences in the estimates were 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.

The fit statistics listed in Table 2 can be made applicable to the two- and three-parameter models. Rogers (1984) derived an appropriate formula for each index and assessed their sensitivity to the same factors listed above as for the one-parameter model. For all indices there was an increased sensitivity to multidimensionality as more parameters were fitted. It appears that the restrictiveness and consequent ''superficiality'' of the one-parameter model ''may allow violations of the unidimensionality assumption to go undetected, while the greater detail and accuracy of prediction provided by the two-parameter model effectively exposes its presence. . . . The three-parameter model, at least as fitted under NOHARM, appears to offer little improvement on the two-parameter model'' in assessing unidimensionality (Rogers, 1984, pp. 93–94).

Other Approaches

Hulin, Drasgow, and Parsons (1983) suggested a procedure that combined latent trait methods and factoring tetrachoric correlations. First, they obtain the eigenvalues of the matrix of item tetrachorics (replacing noncomputable tetrachorics by an ad hoc approximation; see Hulin, Drasgow, & Parsons, pp. 248–249). Using a two- or three-parameter estimation program, they then estimate the item parameters from this correlation matrix. Next, they generate a truly unidimensional item pool by using the estimated item parameters as the parameters of the simulated items (having the same number of simulated examinees and items as the real data set). The eigenvalues of the matrix of tetrachoric correlations obtained from this synthetic data set are then computed. Finally, and most importantly, the second eigenvalues of the real and synthetic data sets are compared. If the difference is large, this suggests a nonunidimensional set. Suggested magnitudes of differences are provided in Drasgow and Lissak (1983). The guidelines are for some very limited cases. If further simulations support the method, it may prove very useful.

Rosenbaum (1984) presented a theorem that seems critical in assessing unidimensionality. This theorem is based on monotone nondecreasing functions of latent variables. These nondecreasing functions include most applications, such as total number correct and positively weighted scores. That is, any score which, if an additional correct response is added, does not decrease the previous sum of items. Rosenbaum's theorem is:

> Let $(\mathbf{Y}, \mathbf{Z})$ be a partition of the item responses in $\mathbf{X}$ into two nonoverlapping groups of responses. Then local independence implies that for every function $\mathbf{h}(\mathbf{Z})$, the responses in $\mathbf{Y}$ are *associated* given $\mathbf{h}(\mathbf{Z}) = \mathbf{a}$ for all $\mathbf{a}$; that is, for all nondecreasing functions $g_1(\mathbf{Y})$ and $g_2(\mathbf{Y})$, cov $[g_1(\mathbf{Y}), g_2(\mathbf{Y})|\mathbf{h}(\mathbf{Z}) = \mathbf{a}] \geqslant 0$. (p. 427)

Rosenbaum provided many illustrations of applications of his theorem. For example, from a test of 120 multiple-choice items, the responses from two items were assessed using the above theorem. The 15,982 candidates were divided into 119 subgroups based on their total score on the remaining 118 items. By the above theorem, local independence implies a nonnegative population correlation (e.g., using the Goodman-Kruskal tau; Goodman & Kruskal, 1979) or equivalently, a population odds ratio of at least 1 (e.g., using the Mantel-Haenszel ratio; Mantel & Haenszel, 1959) within each class.

Given the sound theoretical bases of Rosenbaum's (1984) procedure, it is expected that his procedure will become widely used. It must be noted, however, that Rosenbaum's theorem relates to assessing the assumption of local independence, which is not necessarily the same as the assumption of unidimensionality (see above). Other problems of the method relate to the choice of sample statistic (e.g., the Goodman-Kruskal tau tends to overestimate a relationship; see Reynolds, 1977, pp. 74–75), the existence of not too many cells with small frequencies, the adequacy of the sample, and the use of multiple significance tests.

Generally, it seems that though latent trait theory provides a precise definition of unidimensionality, there is still debate as to the efficacy of the many proposed methods for determining decision criteria for unidimensionality.

Conclusions and Recommendations

At the outset it was argued that unidimensionality was a critical and basic assumption of measurement theory. There have been two major issues that have pervaded this review. First, there is a paucity of understanding as to the meaning of the term unidimensionality and how it is distinguished from related terms. Second, there are too many indices that have been developed on an ad hoc basis with little reference to their rationale or behavior and little, if any, comparison with other indices.

Definition

A major problem in assessing indices of unidimensionality has been that unidimensionality has been confused and used interchangeably with other terms such as reliability, internal consistency, and homogeneity. Consequently, an index is developed from some estimate of reliability, and then it is claimed that the index relates to unidimensionality. It is important that the meaning of these terms is clarified.

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. Yet, there are methods that satisfy these criteria that have not been classified as internal consistency measures (e.g., omega). 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. In the latter case a perfectly homogeneous test is one in which all the items intercorrelate equally. That is, the items all measure the construct or constructs equally. Thus, homogeneity is often a desirable quality, but there have been authors who have advocated that test constructors should not aim for high homogeneity. Cattell (1964, 1978; Cattell & Tsujioka, 1964) has been a principal adversary of 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 argued 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 and Tsujioka's colorful 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. (p. 8)

Internal consistency relates more to a set of methods and seems of limited usefulness. Homogeneity has been used in two senses, one as a synonym for unidimensionality and the other as a measure of equality of item intercorrelations. In the first sense, the term homogeneity is redundant and may be confusing, and in the second sense it may not be desirable. Whether internal consistency and homogeneity are meaningful terms in describing attributes of items and/or tests remains questionable.

Unidimensionality can be defined as the existence of one latent trait underlying the data. This definition is based on latent trait theory and is a specific instance of the principle of local independence, though it is not synonymous with it. As a consequence of this definition, it is probable that indices based on the goodness-of-fit of data to a comprehensive latent trait model may be effective indices of unidimensionality. The problems of such indices relate to ensuring that the correct latent trait model is chosen and that the parameters of the model are satisfactorily estimated. Rosenbaum's (1984) theorem is also worth further investigation because it is so soundly based. Thus, a unidimensional test is one that has one latent trait underlying the data, and such a test may be or may not necessarily be reliable, internally consistent, or homogeneous.

The Indices

Altogether, over 30 indices of unidimensionality have been identified and these were grouped into five sections: methods based on (1) answer patterns, (2) reliability, (3) principal components, (4) factor analysis, and (5) latent traits. Some indices must fail (e.g., ratio of eigenvalues), some are clearly suspect (e.g., alpha), others seem more appropriate in specific conditions (e.g., fit statistics from factor analyzing tetrachoric correlations), while others look promising (e.g., fit statistics from the Christoffersson, 1975, and Muthén, 1978, methods). The major reasons for many indices not being adequate indices of unidimensionality are that unit rank is desired and/or they are based on linear models.

It has been argued above that if a one-factor cubic provides good fit (from a nonlinear factor analysis), then the rank of the inter-item correlation or covariance matrix is three. The claim that unit rank is a necessary condition for unidimensionality is incorrect. Of the numerous methods based on unit rank, alpha has been used most often as an index of unidimensionality. There is, however, no systematic relationship between the rank of a set of variables and how far alpha is below the true reliability. Further, alpha can be high even if there is no general factor, since (1) it is influenced by the number of items and parallel repetitions of items, (2) it increases as the number of factors pertaining to each item increases, and (3) it decreases moderately as the item communalities increase. It seems that modifications of alpha also suffer the same problems. Despite the common use of alpha as an index of unidimensionality, it does not seem to be justified.

Beside the nonnecessity of unit rank, a further problem of many procedures is that a linear model cannot be assumed. When items are scored dichotomously, then the use of a linear factor model and the use of phi or tetrachoric correlations are not appropriate since they assume linearly related variables. Nonlinear factor analysis may be appropriate, but the present problems appear to relate to efficient computer programs for estimating the parameters and a lack of understanding of the behavior of indices based on nonlinear methods. Recent research and computer programs by Etezadi (1981), however, could change this situation.

The indices based on latent trait methods seem to be more justifiable. Yet, if the incorrect model is used, then the resulting indices must fail. It seems unlikely that indices based on the one-parameter or Rasch model will prove useful. The increasing number of indices based on the Rasch model is of concern, particularly since most seem to lack a clear (if any) rationale and there is no supporting evidence, such as simulation studies, to demonstrate their effectiveness. The methods of Christoffersson (1975), Muthén (1978), and McDonald (1982) are based on a weaker form of the principle of local independence and it is likely that some function of the residuals after estimating the parameters may serve as adequate indices of unidimensionality.

Yet, there are still no known satisfactory indices. None of the attempts to investigate unidimensionality has provided clear decision criteria for determining it. What is needed is a monte carlo simulation to assess the various indices under known conditions. Such a simulation is outlined in Hattie (1984a). The simulation assessed the adequacy of most of the indices cited in this review. Data with known dimensionality were generated using a three-parameter latent trait model. Altogether, there were 36 models: two levels of difficulty (−2 to 2, −1 to 1), three levels of guessing (all .0, all .2, and a mixture of .0, .1 and .2), and six levels of dimensionality (one factor with mixed discrimination, one factor with discrimination all 1, two factors intercorrelated .1, two factors intercorrelated .5, five factors intercorrelated .1, and five factors intercorrelated .5).

Only four of the indices could consistently distinguish one-dimensional from more than one-dimensional data sets. The four indices were the sum of (absolute) residuals after fitting a two- or three-parameter latent trait model using NOHARM (Fraser, 1981; McDonald, 1982) or FADIV (Andersson, Christoffersson, & Muthén, 1974). The advantages of using NOHARM are that it is computationally faster and it can handle large data sets.

In subsequent simulations, Hattie (1984b) has investigated decision rules based on these four indices. It seems that if the sum of (absolute) residuals after specifying one dimension is reasonably small, *and* if the sum of residuals after specifying two dimensions is not much smaller, then it can be confidentally assumed that the set of items is unidimensional.

Final Caveat

Finally, it must be considered that it may be unrealistic to search for indices of unidimensionality or sets of unidimensional items. It may be that unidimensionality could be ignored and other desirable qualities of a set of items sought, such as whether the estimates are consistent, or whether the estimates provide a useful and effective summary of the data. Certainly, psychological measurement has so far done without these indices. Moreover, it may be that a set of items will not be unidimensional except for the most simple variables, yet it seems reasonable to claim that unidimensional tests can be factorially complex. Maybe it is meaningful to quest after an index if the question is rephrased from "Is a test unidimensional or not?" to "Are there decision criteria that determine how close a set of items is to being a unidimensional set?"

It may be that multidimensional tests can be confused as unidimensional tests if the multiple dimensions have proportional contributions to each item. In such a case, scores on a test would represent a weighted composite of the many underlying dimensions. In some cases the problem is theoretical in that the labeling of such a test is the major concern. For example, with an arithmetic-reasoning test it can be argued that such a test could be unidimensional, whereas others may wish to argue that it is multidimensional. This is a concern for the test developer and user. It would be advantageous for an index of unidimensionality to distinguish between tests involving one and more than one latent trait, but if the two dimensions contribute equally to the item variances, this might not be possible. The use of second-order factor analysis specifying one second-order factor may be necessary to detect such unidimensional sets of items. The existence of more than one second-order factor is convincing evidence of a multidimensional data set. Certainly, identifying a unidimensional set of items and labeling the set are separate processes.

Throughout this review the issue has been the unidimensionality of an item set. Obviously, it must be remembered that dimensionality is a joint property of the item set, or the pool of which it is a sample, and a particular sample of examinees from its underlying population. Much recent research has indicated that the same set of test items may be attacked by persons using different cognitive strategies (Klich & Davidson, 1984). Although agreeing that this occurs, it still seems worthwhile to devise tests that measure similar content and/or styles and thus aim for more dependable information about individual differences.

Further, it may be that an act of judgment and not an index is required. Kelly (1942) argued that embodied in such concepts as unidimensionality is a belief or point of view of the investigator such that an act of judgment is demanded when a researcher asserts that items measure the same thing. Thus, not only may it be possible to recognize by inspection whether one test appears to be unidimensional when compared to another, but also even if there is an index, then judgment must still be used 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 the degree to which a test is unidimensional.

References

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

Anderson, T. W. (1959). Some scaling models and estimation procedures in the latent class model. In U. Grenander (Ed.), *Probability and statistics* (pp. 9–38). New York: Wiley.

Andrich, D., & Godfrey, J. R. (1978–1979). Hierarchies in the skills of Davis' reading comprehension test, Form D: An empirical investigation using a latent trait model. *Reading Research Quarterly, 2,* 183–200.

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

Bejar, I. I. (1980). A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. *Journal of Educational Statistics, 17,* 282–296.

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

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

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika, 46,* 443–459.

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

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment.* Beverly Hills CA: Sage.

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

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

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

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

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

Cliff, N. (1977). A theory of consistency or ordering generalizable to tailored testing. *Psychometrika, 42,* 375–399.

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

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

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

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

Etezadi, J. (1981). A general polynomial model for nonlinear factor analysis. *Dissertation Abstracts International, 42,* 4342a.

Etezadi, J., & McDonald, R. P. (1983). A second generation nonlinear factor analysis. *Psychometrika, 48,* 315–342.

Fraser, C. (1981). *NOHARM: A FORTRAN program for non-linear analysis by a robust method for estimating the parameters of 1-, 2-, and 3-parameter latent trait models.* Armidale, Australia: University of New England, Centre for Behavioural Studies in Education.

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

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

George, A. A. (1979, April). *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 CA.

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

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

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

Goodman, L. A., & Kruskal, W. H. (1979). *Measures of association for cross classifications.* New York: Springer-Verlag.

Gorsuch, R. L. (1974). *Factor analysis.* Philadephia PA: Saunders.

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

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

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

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

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

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

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

Gustafsson, J. E., & Lindblad, T. (1978). *The Rasch* *model for dichotomous items: A solution of the conditional estimation problem for long tests and some thoughts about item screening procedures* (Report No. 67). Göteborg, Sweden: University of Göteborg, Institute of Education.

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

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

Guttman, L. (1950). The principal components of scale analysis. In S. S. Stouffer (Ed.), *Measurement and prediction* (pp. 312–361). Princeton NJ: University Press.

Haley, D. C. (1952). *Estimation of dosage mortality relationships when the dose is subject to error* (Report No. 15). Stanford CA: Stanford University, Applied Mathematics and Statistics Laboratory.

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

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

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

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

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

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

Hattie, J. A. (1980). The Progressive Achievement Tests revisited. *New Zealand Journal of Educational Studies, 15,* 194–197.

Hattie, J. A. (1981). A four stage factor analytic approach to studying behavioral domains. *Applied Psychological Measurement, 5,* 77–88.

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

Hattie, J. A. (1984b). *Some decision rules for determining unidimensionality.* Unpublished manuscript, University of New England, Centre for Behavioural Studies, Armidale, Australia.

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

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

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

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

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

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

Horrocks, J. E., & Schoonover, T. I. (1968). *Measurement for teachers.* Columbus OH: Merrill.

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

Hulin, C. L., Drasgow, F., & Parsons, C. (1983). *Item response theory: Applications to psychological measurement.* Homewood IL: Dow & Jones Irwin.

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

Humphreys, L. G. (1952). Individual differences. *Annual Review of Psychology, 3,* 131–150.

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

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

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

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

Jöreskog, K. G. (1978). Structural analysis of covariance and correlation matrices. *Psychometrika, 43,* 443–447.

Jöreskog, K. G., & Sörbom, D. (1981). *LISREL V: Analysis of linear structural relationships by the method of maximum likelihood.* Chicago: National Educational Resources.

Kaiser, H. F. (1968). A measure of the average intercorrelation. *Educational and Psychological Measurement, 28,* 245–247.

Kaiser, H. F. (1970). A second generation little jiffy. *Psychometrika, 35,* 401–415.

Kelley, T. L. (1935). Essential traits of mental life. *Harvard Studies in Education, 26,* 146–153.

Kelley, T. L. (1942). The reliability coefficient. *Psychometrika, 7,* 75–83.

Klich, L. Z., & Davidson, G. R. (1984). Toward a recognition of Australian Aboriginal Competence in cognitive formation. In J. Kirby (Ed.), *Cognitive strategies and educational performance* (pp. 155–202). New York: Academic Press.

Koch, W. R., & Reckase, M. D. (1979, September). *Problems in application of latent trait models to tailored testing* (Research Report 79–1). Columbia MO: University of Missouri, Educational Psychology Department, Tailored Testing Research Laboratory.

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

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

Lam, R. (1980). *An empirical study of the unidimensionality of Ravens Progressive Matrices.* Unpublished master's thesis, University of Toronto, Canada.

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

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

Loevinger, J. (1944). *A systematic approach to the construction of tests of ability.* Unpublished doctoral dissertation, University of California.

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

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

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

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

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

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

Lumsden, J. (1957). A factorial approach to unidimensionality. *Australian Journal of Psychology, 9,* 105–111.

Lumsden, J. (1959). *The construction of unidimensional tests.* Unpublished master's thesis, University of Western Australia, Perth, Australia.

Lumsden, J. (1961). The construction of unidimensional tests. *Psychological Bulletin, 58,* 122–131.

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

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

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

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

McDonald, R. P. (1965b). *Numerical polynomial models in nonlinear factor analysis* (Report No. 65–32). Princeton NJ: Educational Testing Service.

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

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

McDonald, R.P. (1967c). Nonlinear factor analysis. *Psychometric Monographs* (No. 15).

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

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

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

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

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

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

McDonald, R. P., & Fraser, C. (1985). *A robustness study comparing estimation of the parameters of the two-parameter latent trait model.* Unpublished manuscript.

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

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

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika, 43,* 551–560.

Muthén, B. (1981). 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 multidimensional perspective* (pp. 201–214). London: Sage.

Muthén, B., & Christoffersson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika, 46,* 407–419.

Nishisato, S. (1966). Minimum entropy clustering of test-items. *Dissertation Abstracts International, 27,* 2861B.

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

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

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

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

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

Nunnally, J. C. (1970). *Introduction to psychological measurement*. New York: McGraw-Hill.

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

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

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

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, 4,* 321–333.

Rasch, G. (1966a). An individualistic approach to item analysis. In P. Lazarsfeld & N. V. Henry (Eds.), *Readings in mathematical social science* (pp. 89–108). Chicago: Science Research Association.

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

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

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

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

Rentz, R. R., & Rentz, C. C. (1979). Does the Rasch model really work? *NCME Measurement in Education, 10,* 1–11.

Reynolds, H. T. (1977). *The analysis of cross-classifications*. New York: Free Press.

Reynolds, T. (1981). Ergo: A new approach to multidimensional item analysis. *Educational and Psychological Measurement, 41,* 643–660.

Rogers, H. J. (1984). *Fit statistics for latent trait models*. Unpublished master's thesis, University of New England, Armidale, Australia.

Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika, 49,* 425–435.

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

Shostrom, E. L. (1972). *Personal Orientation Inventory: An inventory for the measurement of self-actualization*. San Diego CA: EdITS.

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

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

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

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

Svoboda, M. (1972). *Distribution of binary information in multidimensional space*. Unpublished master's thesis, University of Toronto, Canada.

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

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

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

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

van den Wollenberg, A. L. (1982a). A simple and effective method to test the dimensionality axiom of the Rasch model. *Applied Psychological Measurement, 6,* 83–91.

van den Wollenberg, A. L. (1982b). Two new test statistics for the Rasch model. *Psychometrika, 42,* 123–140.

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

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

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

White, B. W., & Saltz, E. (1957). Measurement of reproducibility. *Psychological Bulletin, 54,* 81–99.

Wise, S. L. (1982, March). *Using partial orders to determine unidimensional item sets appropriate for item response theory.* Paper presented at the annual meeting of the National Council for Measurement in Education, New York.

Wise, S. L. (1983). Comparisons of order analysis and factor analysis in assessing the dimensionality of binary data. *Applied Psychological Measurement, 7,* 311–312.

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

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

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

Wright, B. D., & Stone, M. H. (1979). *Best test design.* Chicago: Mesa Press.

Acknowledgments

The author gratefully acknowledges the contributions of R. P. McDonald, S. Nishisato, R. Traub, M. Wahlstrom, W. Olphert, and reviewers whose criticisms greatly improved this paper.

Author's Address

Send requests for further information to John A. Hattie, Centre for Behavioural Studies, University of New England, Armidale, NSW 2351, Australia.

Reprints

Reprints of this article may be obtained *prepaid* for $2.50 (U.S. delivery) or $3.00 (outside U.S.; payment in U.S. funds drawn on a U.S. bank) from Applied Psychological Measurement, N658 Elliott Hall, University of Minnesota, Minneapolis, MN 55455, U.S.A.