Three Approaches to Determining the Dimensionality of Binary Items

Mary Roznowski
Ohio State University

Ledyard R Tucker and Lloyd G. Humphreys
University of Illinois

A monte carlo investigation of three approaches to assessing the dimensionality of binary items used a population model that allowed sampling of items and examinees and provided for variation and control of important parameters. The model was realistic of performance of binary items in current tests of cognitive abilities. Three indices were investigated: one based on the property of local independence of unidimensional tests (the independence index), one based on patterns of second factor loadings derived from simplex theory (the pattern index), and one that reflects the shape of the curve of successive eigenvalues (the ratio of differences index). The last index was used for matrices of phi coefficients, tetrachoric correlations, and variances-covariances. The local independence index reported here was the most accurate dimensionality index. The pattern index was accurate under many combinations of parameters, but decreased substantially at the highest level of factor correlations and the widest dispersion of item difficulties. None of the eigenvalue indices produced satisfactory accuracy, except under the most favorable combinations of parameters. Nonetheless, the eigenvalues of variance-covariance matrices provided a more accurate basis for dimensionality decisions than tetrachoric correlations, which have been the statistic of choice of many investigators. Recommendations for use are also given. *Index terms: binary items, dimensionality, factor analysis, phi correlations, tetrachoric correlations.*

The issue of dimensionality has come to the fore with the increase of interest in recent years in item response theory (IRT; see Hulin, Drasgow, & Parsons, 1983; Lord, 1980; Stout, 1987, 1990) and its use in adaptive testing and test construction. One important assumption in IRT is that of unidimensionality of the latent space defined by the measures. Dependable IRT item and person parameters can be obtained from item pools that are not unidimensional in the strict sense of the term—a dominant dimension defined by correlated group factors is sufficient (Drasgow & Parsons, 1983; Harrison, 1986). Such item pools are likely to be considerably more valid in most applications (Humphreys, 1985a). However, methods of test administration, such as adaptive testing, that potentially provide a different set of items for examinees and regularly provide different sets of items for examinee subgroups place a greater demand on the unidimensionality assumption.

A standard method of assessing the dimensionality of a pool of binary items is to compute tetrachoric correlations among the items, obtain principal components, and inspect the latent roots (Gorsuch, 1983; McDonald, 1985). If the item pool is unidimensional, it is expected that the first root will describe a major—although numerically uncertain—proportion of total variance. A large difference is also expected in the size of the second root relative to the first, and no sizable gap is expected subsequent to the first difference. However, these expectations have no firm quantitative basis.

Tetrachoric correlations are frequently recommended for component analysis because the size of product-moment correlations is affected by item difficulties, and because phi coefficients

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 15, No. 2, June 1991, pp. 109-127

0146-6216/91/020109-19$2.20

produce difficulty factors (Gorsuch, 1983; McDonald, 1985; Wherry & Gaylord, 1944). If items produced a perfect Guttman scale, the matrix of intercorrelations would be a perfect simplex, not a Spearman hierarchy.

Although the use of tetrachorics is fairly standard, the tetrachoric matrix has two undesirable properties for factor analysis and assessing dimensionality. The sampling errors of individual tetrachorics vary widely as a function of item difficulty levels, and the correlation matrix is likely to be non-Gramian. The effects of these properties become more severe as the spread of difficulties increases (Gorsuch, 1983; McDonald, 1985; Tucker, 1983).

It is important to determine the accuracy with which a decision contrasting unidimensionality with multidimensionality can be made, and to evaluate both traditional and nontraditional methods. It is also necessary to have a factor model of known dimensionality, to sample items and examinees, and to compare indices in the same dataset. Various methods of assessing dimensionality have been proposed and evaluated (see Hattie, 1985; McDonald, 1981; Reckase, 1985; van den Wollenberg, 1982; Zwick, 1987), as have various alternative frameworks for discussing dimensionality (e.g., Rosenbaum, 1984; van den Wollenberg, 1982). The goal of this study was to expand this research base and satisfy these various requirements in an evaluation of three heuristic methods of assessing the dimensionality of binary item pools.

Method

Characteristics of a Desirable Model

A main objective of the study was to use a model that realistically simulated psychological data, especially cognitive ability data. A starting point was the model developed by Tucker, Koopman, and Linn (1969), who discussed both a "formal" factor model and a "simulation" model. Their latter model added a relatively small amount of variance from a large number of small, overlapping factors ("minor factors"; see also Humphreys, 1985b; McDonald, 1981; Stout, 1990). The added variance simulates the very large number of determinants of responses to items, because neither items nor tests can be pure measures of a hypothetical factor. It is important to investigate both models as they affect the assessment of dimensionality, but results are reported here only for the formal model. Thus, one requirement was that the population model resemble the formal factor model of Tucker et al.

A second requirement was that the model furnish data with realistic psychometric properties. Distributions of item difficulties and item intercorrelations should be at levels typical in cognitive ability item pools. For instance, product-moment correlations among such items in wide ranges of talent are rarely greater than .25, and frequently are considerably lower (e.g., differences in the size of mean correlations between those found in a heterogeneous intelligence test and those from a homogeneous test of mechanical information are quite small).

A third requirement was that the data generation model be compensatory. In such a model, an item's variance is broken down into a linear combination of loadings on major group factors and random error. Intercorrelations are the sums of products of overlapping factors. A compensatory model fits psychological data rather well, and its success is reflected in the predictive validities of psychological tests for practical criteria.

A final feature of the model was that it be sufficiently general to be adapted to the several common models of item functioning. A continuous normal distribution is assumed to underlie the binary responses of items measuring cognitive abilities. This assumption, given appropriate constraints on the size and variability of loadings on a single factor, provides for a Guttman scale or for a one-, two-, or three-parameter IRT model.

Description of the Model

Data were generated so that the structural relations desired were prespecified, which allowed for evaluating the validity of the factor analytic

design. With such a generation method, a design plan can be adjusted in terms of the exact nature and quality of the data desired.

Basic notions. The model specified the existence of a major domain B, which represented the most important influences on observed scores. Every row i in factor matrix **B** represented a variable, and every column j represented a factor. These elements B_{ij} were the conceptual values in the major domain, and they represented theoretical notions concerning the composition and structural relations of the input variables. The major domain was presumed to have simple structure. Finally, a restriction of the factor matrix to positive manifold was required because cognitive ability item responses were being simulated.

Factor intercorrelations. An initial step in generating the factor model was to determine the phi matrix (**Φ**)—the intercorrelations among factors. Significant control of this parameter was available. First, a value was selected in the range (0,1) that indirectly determined the desired minimum level of factor correlations. A second, higher value in the same range was selected from which the first value was subtracted. This difference was multiplied by a random number in the interval (0,1), and that product was added to the initial value. The procedure was repeated for each factor by using the same first and second values, which gave a unique value for each factor in the desired range. Taking the square root of the values from the preceding step and obtaining the cross-products among factors gave the correlations between pairs of factors. In effect, these factor intercorrelations were the products of factor loadings on a single, general factor.

Factor correlations in the present application were determined by initial values (first and second values) of .20 and .40 for the low, .40 and .60 for the intermediate, and .50 and .70 for the high levels of intercorrelations. Thus, the average random increments to the base value were identical for the three levels.

Ensuring simple structure. The next step was to derive a first-order **B** having simple structure with a restriction to positive manifold. Initial B values were determined by selecting random numbers in the (0,1) interval. Several checks were carried out to ensure that the structural relations and general quality of the factor matrix were as desired.

1. To establish simple structure, a cutoff probability for establishing zeroes in **B** was used to determine the number of zeroes in each column (i.e., per factor).
2. The presence of at least one high (nonzero) entry in each row (i.e., per variable) of the factor was ensured by inserting a nonzero loading in the column position selected randomly, if no nonzero loading was present.
3. The number of high loadings in each column was counted. The minimum number of high loadings was determined a priori. As before, high entries were inserted until the desired minimum number of high loadings was achieved.
4. All pairs of columns (j,k) were compared to determine whether there was an appropriate (specified a priori) number of rows with high entries on j and 0 entry on k. This was further assurance that each variable was represented and that simple structure was achieved.

Size of factor loadings. Adjustments were implemented to influence the absolute size of loadings in the final input factor matrix. Two initial values were set to indicate the range from which loadings for the nonzero entries in the matrix were selected. The procedure presented earlier for factor intercorrelations was used here as well. A vector of loadings (referred to as *Gl* coefficients) was obtained, with one coefficient for each variable. These coefficients were randomly selected from a rectangular distribution in the range Gl_S to Gl_E (''start'' and ''end'' values, set a priori), which represented the hypothesized variance from the major domain.

An **A** matrix—the factor matrix transformed to uncorrelated factors—was obtained next. The first step in this process was obtaining the Cholesky decomposition (**v**) of the **Φ** matrix. **B**

is post-multiplied by $\mathbf{v}$ to obtain an initial $\mathbf{A}$ matrix. Next, rows of $\mathbf{A}$ and $\mathbf{B}$ are adjusted to take into account the vector of *GI* coefficients. After the rows of $\mathbf{A}$ are normalized, the same multiplier is applied to rescale $\mathbf{B}$ in accordance with the transformation to $\mathbf{A}$.

Application to binary scores. The model has thus far been described in terms of continuous variates. Model output was converted to binary data. Each sampled individual had a score on each of k dimensions. Definitions for scores and their components are as follows:

X_{ik} = score or measure of individual i on dimension k;
Z_{ij} = item score of individual i on item j; and
W_{jk} = weight for item j on dimension k.

A fundamental relation involving these components is

$$Z_{ij} = \sum_k X_{ik} W_{jk} \quad . \tag{1}$$

To generalize this relation,

$$\mathbf{Z}_i = \mathbf{X}_i \boldsymbol{\Omega} \quad , \tag{2}$$

where $\mathbf{X}_i$ is a vector of scores X_{ik}, and $\mathbf{Z}_i$ is a vector of scores Z_{ij}. $\boldsymbol{\Omega}$ is a matrix of factor weights W_{jk}.

The $\boldsymbol{\Omega}$ weight matrix includes *common*, *specific*, and *measurement error* factor weights. Specific and error factors are considered together as the uniqueness component $\mathbf{U}$. The common factor weight matrix is designated $\mathbf{A}$, as before. The weight matrix $\boldsymbol{\Omega}$ can be decomposed from one super matrix to two separate weight matrices, $\mathbf{A}$ and $\mathbf{U}$, representing common and unique influences. The vector $\mathbf{X}_i$ can be partitioned similarly into common $\mathbf{a}_i$ and unique components $\mathbf{u}_i$, with $\mathbf{X}_i = \{\mathbf{a}_i, \mathbf{u}_i\}$. Expanding Equation 2 results in

$$Z_{ij} = \mathbf{a}_i \mathbf{A} + \mathbf{u}_j \mathbf{U} \quad . \tag{3}$$

Using this basic factor model, binary scores can be derived that conform to the qualities of the simulation model in continuous form. Because the factor model is a linear model involving the addition of contributions to item scores, the factor weight matrix $\mathbf{A}$ can be applied to an underlying score distribution, and be combined with unique components to determine item scores. In computing Z_{ij} from Equation 1, a random normal deviate for individual i on each dimension k (X_{ik}) is selected and multiplied by the factor loading for item j on dimension k (a_{jk}). For instance, the following holds for a three-factor model:

$$Z_{ij} = X_{i1} a_{j1} + X_{i2} a_{j2} + X_{i3} a_{j3} \quad . \tag{4}$$

Uniqueness is factored in by selecting a random normal deviate for each item and multiplying it by that item's uniqueness. This product is added to the Z_{ij} from the common factor influences (from Equation 4).

Generation of binary responses. To generate binary scores (s_{ij}), a simulated person's (simulee) normally distributed continuous score for item j (Z_{ij}) is compared to a cutting score for item j. The cutting score mean and standard deviation (determined a priori) are used to set up the cutting score values (d_j)—one for each item. A vector of guessing parameters ($\mathbf{c}_j$) is set up with a desired range of guessing probabilities. The equations that characterize the probabilistic model used for generating binary responses are:

$$\text{If } (Z_{ij} \geq d_j),\ S_{ij} = 1 \tag{5}$$

and

$$\text{if } (Z_{ij} < d_j),\ S_{ij} = 0 \tag{6}$$
$$\text{with } P(S_{ij} = 1) = c_j \quad .$$

Thus, if the score of simulee i to item j is greater than or equal to the cutting score for item j, the simulee receives a correct score (1) for that item. If Z_{ij} is less than the threshold, the simulee receives a 0 to indicate an incorrect response. The guessing component is taken into account such that an incorrect answer to item j is changed to correct with probability c_j. This process is continued for all j responses and i simulees. The event of S_{ij} equalling 0 or 1 is independent from occasion to occasion. Finally, the model is probabilistic in that it involves each item as a sampling of the many influences on item responses.

Independent Parameters

Five model parameters were varied in a factorial design. Cells in the design were filled with 100 independent samples, with a new population model being generated for each sample.

Sample sizes. *N*s were 125, 500, and 2,000, providing 2-to-1 ratios of expected standard errors. Because this parameter is under the researcher's control in practice, an adequate *N* can usually be established in advance for purposes of determining dimensionality. All dimensionality indices should vary in accuracy as a function of sample size.

Number of items. This parameter *n* took on values of 20, 30, 40, 50, and 60. For a given number of factors, the factors are expected to be more accurately defined as the number of items increases.

Item difficulties. Three distributions of difficulties were used, varying from flat to peaked. The distributions defined in normal deviate units were as follows: $\mu = .10$, $\sigma = .80$ (wide); $\mu = .00$, $\sigma = .50$ (medium); $\mu = -.13$, $\sigma = .32$ (narrow). The positive sign of μ indicates a central tendency of *p* values less than .50 in the absence of guessing. However, the means for all distributions were greater than .50 after guessing was applied. An experienced test constructor can exercise some degree of control over this parameter in practice. More importantly, item difficulties can always be known before a decision must be made concerning dimensionality. Note that variation in the distribution of item difficulties affects the size of item product-moment intercorrelations, but it only affects the sampling variability of tetrachoric correlations.

In the final samples, mean proportions correct of .567, .580, and .615 were obtained with standard deviations of .215, .155, and .113, respectively. These values included the effects of guessing.

Factor intercorrelations. The manipulation of this parameter produced mean correlations of approximately .35 (low), .55 (intermediate), and .70 (high), which also produced variation in item intercorrelations. Standard deviations were approximately .08. For purposes of determining dimensionality, it is not possible to know the factor intercorrelations. However, the mean product-moment item correlation can be obtained prior to a decision; this value can be readily computed from an internal consistency estimate for total score.

Number of factors. This parameter was varied from 1 through 5. A test constructor can control number of factors only approximately by adherence to item specifications. However, good test construction practices will not increase the number of major factors as the number of items is increased. Conversely, when unidimensionality is not achieved by an item writer working toward that goal, the size of the correlations among the multiple factors is likely to be high. If unidimensionality can be rejected, it becomes imperative to decide on the number of group factors, although this decision is not central to the present research. Finally, variation of this parameter in the present model resulted in variation in the level of item intercorrelations (both product-moment and tetrachoric).

Given that oblique factors are always characteristic of cognitive ability data, increasing difficulty can be expected in distinguishing between unidimensionality and multidimensionality as the number of factors increases. Positively correlated first-order factors determine one or more second-order factors. In the present model there was a single second-order factor that described an increasing amount of total variance as the number of factors increased. When the second-order factor was held constant, the total contribution of five factors was less than the total contribution of two factors. Thus, as the number of homogeneously correlated group factors increased, data appeared to be more unidimensional, and the task of determining dimensionality became more difficult in both modeled and real data.

The Dimensionality Indices

Local independence. If a test is unidimensional, the items are independent of each other

in a sample of examinees at the same latent trait level. This property of the test is called local independence (see Lazarsfeld, 1959; Lord & Novick, 1968; McDonald, 1981).

Local independence may be approximated in fallible data by restricting analyses to members of the sample who have the same total score (see also van den Wollenberg, 1982). Total score may be used as an estimate of score on the latent trait. If the items are unidimensional, there is no systematic structure in the intercorrelations in which total score is held constant. In a relatively new conceptualization, Stout (1990) discusses "essential independence" and "essential dimensionality," which more appropriately reflect the nature of real data. These concepts take into account the existence of multiple determinants of items, and attempt to reflect the *dominant* dimension(s) in item pools.

Because an index of dimensionality cannot be formed by analyzing n different matrices independently, one for each level of total score, an aggregating procedure was required. This was followed by a method for finding structure if it exists. In greater detail, the procedure was as follows:[1]

1. An aggregate variance-covariance matrix was formed from the separate matrices computed from samples of simulees with the same total score. Each submatrix was weighted by sample size in forming the aggregate.
2. Signs of the aggregate covariances were changed in accordance with the sign-changing procedure of centroid factor analysis to maximize the algebraic sum of the covariance matrix (Thurstone, 1947). Because this matrix had total score held constant, there were approximately equal numbers of positive and negative signs in the aggregate matrix at the outset.
3. The ratio of the algebraic sum of covariances following the sign change to the absolute sum of covariances was formed, disregarding entries in the principal diagonal. Ratios that approached unity indicate the presence of more than one factor among the original item covariances (i.e., there is structure remaining in the matrix, although one factor has been removed).
4. The ratio in Step 3 was compared to the parallel ratio of algebraic to absolute sum of the values of the covariance matrix of scores in which total score was *not* held constant. The sign change was placed for this ratio so that the technique would be applicable to matrices composed of both positive and negative correlations among noncognitive items. This ratio equals 1.0 among cognitive ability items in the absence of occasional negative correlations in small samples.
5. The local independence index (LII) was computed as the difference between the ratios in Steps 3 and 4. The presence of multiple factors is indicated by small differences. A ratio of these ratios was also evaluated, but no consistent advantage appeared.

Pattern of second factor loadings. When the product-moment intercorrelations (phis) of a perfect Guttman scale are factored, loadings on the second factor have a distinctive pattern of signs and relative sizes. If the obtained intercorrelations are based on a single factor plus random error, the second factor loadings are expected to approximate this distinct pattern. Several means of obtaining a quantitative index of dimensionality from the pattern of second factor loadings were explored. The second index (PI) was thus based on the sizes and signs of second factor loadings.

The second component in the correlation matrix formed by a perfect Guttman scale produces a curve that approximates an ogive. Easy and difficult items have high loadings of opposite sign, and items of moderate difficulty have second factor loadings close to zero. The index was computed as follows:[2]

1'. The **R** matrix was factored after replacing the unities in the diagonal with squared multiple correlations.

[1]Credit for this index is due Ledyard Tucker.

[2]Credit for this index is due Lloyd Humphreys.

2. Items were ranked by difficulty and divided into easy and difficult halves.
3. Four sums of second factor loadings were obtained—one for each sign in each half.
4. The absolute sums of opposite sign from the upper and lower halves in the difficulty distribution were obtained.
5. The smaller of the sums was the dimensionality index, which approaches zero in a unidimensional test.

Ratio of eigenvalue differences. Indices of dimensionality in the intercorrelations of continuous variates have long depended on relations among the successive eigenvalues obtained from principal factors. Whether known as "root staring" or the "scree test," the assumption is that there will be a break in the eigenvalue curve at the point the last replicable factor has been extracted. Thereafter the relations among the latent roots should have little slope. For the one-factor case, the index (the ratio difference index, RDI) was computed as the ratio of the difference between the first two roots to subsequent differences (the average of the next two differences was chosen); this ratio should be large if the data are unidimensional. This principle was applied to the eigenvalues of tetrachoric, product-moment, and variance-covariance matrices.

This ratio of differences to differences reflects the shape of the curve of eigenvalues more accurately than the ratio of the first to the second root. The eigenvalue ratio of the variance-covariance matrix was found to be generally more accurate than that for the other matrices and its performance was evaluated relative to the above two indices. Results are thus given for the ratio index from variance-covariance matrices. The finding that this ratio was more accurate in variance-covariance matrices than the same ratio in tetrachoric and phi matrices presumably reflects two different principles. Variance-covariance matrices and matrices of phi correlations are more stable from sample to sample than tetrachoric matrices. However, less weight is given in the variance-covariance matrix to extreme items that produce difficulty factors when phis are analyzed.

Use of the "root 1" index. Even more popular than a break in the latent roots obtained in a principal factor analysis is the "root 1" criterion in principal components analysis. This criterion was not examined systematically when applied to binary items because it is inappropriate, given the low level of item intercorrelations in cognitive tests. Inspection of only a few samples is required to reject "root 1" as a basis for a decision concerning dimensionality. Its failure in binary data also indicates an important reason for failure in continuous data—that is, the criterion will fail when replicable factors are determined by relatively small correlations.

Overlap Criterion

In order to determine the effectiveness of the indices, a measure of separation was needed between unifactor versus multifactor distributions. Index distributions for each one-factor sample and each multiple-factor sample were ordered numerically. Each multifactor distribution was compared with the corresponding one-factor distribution. The maximum value in the numerically low distribution and the minimum value in the numerically high distribution were determined. Then the number of entries in the high distribution less than the maximum value in the low distribution, and the number of entries in the *low* distribution less than the minimum value in the *high* distribution were determined. The sum of these two values gave a numerical indication of the overlap of the two distributions.

Because there were 100 replications in each of the 675 cells of the factorial design, the maximum value of the measure of discrimination was 200, which indicates complete independence of distributions. Therefore, the previously obtained sum was subtracted from 200 to transform the metric to an index of separation rather than one of overlap. A value of 100 indicated complete overlap. For data in which overlap was small, there was a single maximum score or small range of contiguous maximum scores. In cases in which overlap was large, there was frequently more than a single maximum, but this was not a defect in the

measure of overlap, at least for applied purposes.

Experimental Variables

The parameters varied in a factorial design were as follows: sample size (3 levels), number of items (5 levels), distribution of item difficulties (3 levels), number of factors (5 levels), and levels of factor intercorrelations (3). Other parameters were set so that the model reflected realistic psychological data. Some items loaded on a single factor, and others were more complex. A sufficient number of factorially pure items was included to provide adequate factor definition, regardless of the number of factors.

The size of item intercorrelations was set at levels typical of cognitive tests in which substantial unique variance is the norm. However, as number of factors, level of factor correlations, and distributions of difficulties in the model were altered, variation in size of resultant item correlations could not be held constant. Resultant mean correlations for combinations of item difficulties and factor intercorrelations for the five factors are presented in Table 1.

Results

Factors in Continuous Variates

Insight into the nature of the model can be gained by factoring intercorrelations of continuous variates prior to dichotomization. Table 2 presents eigenvalues for three samples of 500 for 30 items and 1 through 5 factors in which squared multiple correlations served as communality estimates. Average intercorrelations were set at .55, an intermediate level of obliqueness. Also included are estimated eigenvalues for random data matrices based on the parallel analysis procedure of Humphreys and Montanelli (1975) and Montanelli and Humphreys (1976). In this and the following tables, "low," "intermediate," and "high" refer to the three levels of factor correlations. "Narrow," "medium," and "wide" refer to difficulty distributions.

For the continuous data in Table 2, there are breaks in the curve formed by successive eigenvalues for the proper number of factors in each sample. Parallel analysis also led to the expected number of factors. Eigenvalues for the same correlation matrices with unities in the diagonal (also shown in Table 2) have similar breaks for the expected number of common factors. However, the "root 1" criterion failed after four common factors. Furthermore, investigators advocating a parallel analysis criterion in principal components analysis would commit additional errors by frequently accepting only three factors in samples in which either four or five were required.

With more highly oblique factors or smaller *N*s, the number-of-factors decision would have been made with less confidence and probably less accuracy. The distribution of eigenvalues indicates increasing unidimensionality as the number of factors increases from 2 to 5. It is clear that the number-of-factors decision would be made with more errors in all combinations of parameters after the loss of information from the conversion of continuous variates to binary scores and from the introduction of success by guessing.

Table 3 contains Kuder-Richardson (KR-21) coefficients as a function of number of items and

Table 1
Mean Item Correlations for Combinations of Factor Obliqueness, (Low, Intermediate, and High), Distribution of Item Difficulties (Narrow, Medium, and Wide), and Number of Factors

No. of Factors	Low			Intermediate			High		
	Narrow	Medium	Wide	Narrow	Medium	Wide	Narrow	Medium	Wide
1	.24	.20	.18	.24	.20	.18	.24	.20	.18
2	.24	.20	.16	.27	.23	.18	.29	.24	.19
3	.21	.18	.15	.27	.21	.17	.28	.23	.18
4	.20	.17	.14	.24	.21	.16	.27	.23	.18
5	.20	.17	.14	.24	.20	.16	.27	.23	.18

Table 2
Successive Eigenvalues for One to Five Factors in the Continuous Model With Squared Multiple Correlations and With Unities in the Diagonal for Three Samples (N = 500 per Sample)

	Eigenvalue							
Sample	1	2	3	4	5	6	7	8
Squared Multiple Correlations								
1 Factor								
1	19.42	.21	.15	.14	.14	.10	.10	.10
2	17.68	.21	.18	.18	.12	.12	.10	.09
3	19.32	.16	.14	.14	.12	.11	.09	.08
2 Factors								
1	15.27	3.60	.17	.14	.13	.09	.06	.06
2	15.08	3.55	.14	.14	.10	.09	.08	.07
3	14.90	3.27	.18	.16	.13	.10	.09	.08
3 Factors								
1	14.75	2.11	1.51	.17	.12	.11	.10	.08
2	15.13	2.65	1.89	.15	.12	.11	.11	.08
3	15.90	1.68	1.41	.16	.13	.12	.10	.09
4 Factors								
1	14.05	1.91	1.46	.66	.15	.14	.10	.08
2	15.95	1.24	1.08	.74	.17	.14	.11	.10
3	15.65	1.08	.82	.78	.18	.17	.12	.10
5 Factors								
1	13.81	1.50	1.14	.63	.50	.18	.14	.10
2	15.22	1.22	.98	.79	.43	.12	.11	.09
3	16.79	1.19	.74	.62	.43	.15	.13	.11
Parallel Analysis								
Estimates	.57	.48	.42	.38	.34	.32	.27	.24
Unities								
1 Factor								
1	19.76	.58	.54	.53	.52	.50	.48	.45
2	18.07	.66	.66	.64	.58	.56	.53	.52
3	19.66	.59	.57	.52	.51	.49	.47	.44
2 Factors								
1	15.63	3.98	.61	.59	.58	.52	.49	.49
2	15.45	3.93	.58	.56	.56	.52	.50	.49
3	15.28	3.67	.62	.61	.60	.59	.54	.52
3 Factors								
1	15.14	2.50	1.93	.59	.56	.53	.51	.51
2	15.48	2.99	2.25	.55	.53	.52	.48	.46
3	16.26	2.08	1.79	.58	.56	.55	.53	.49
4 Factors								
1	14.46	2.31	1.87	1.10	.59	.58	.54	.52
2	16.33	1.62	1.47	1.13	.58	.55	.54	.52
3	16.04	1.48	1.25	1.19	.64	.60	.58	.55
5 Factors								
1	14.24	1.94	1.61	1.08	.97	.66	.63	.56
2	15.61	1.61	1.39	1.21	.89	.58	.53	.51
3	17.15	1.55	1.12	.98	.80	.54	.53	.50

Table 3
Kuder-Richardson Coefficients as a Function of Number of Items and Selected Values of Mean Item Correlations

Mean Item Correlation	Number of Items				
	20	30	40	50	60
.30	.90	.93	.94	.96	.96
.25	.87	.91	.93	.94	.95
.20	.83	.88	.91	.93	.94
.15	.78	.84	.88	.90	.91
.12	.73	.80	.84	.87	.89

selected values of mean item correlations. Values increase as correlations and number of items increase. This table shows the relation between mean item correlations and internal consistencies for total scores. From these values, it can be seen that the generation model produced realistic output. The mean item correlations presented are representative of the values obtained in the data, and they produced realistic homogeneity coefficients.

Indices

Table 4 contains the measure of discrimination for LII for all combinations of parameters for 20 through 60 items, respectively. The index's accuracy increases with sample size and number of items. It decreases as obliqueness of the factors, range of item difficulties, and number of factors increase.

Of interest is the decrease in accuracy with the increase in number of factors. As mentioned, the second-order factor defined by the oblique factors makes an increasing contribution to total variance as number of factors increases, whereas each separate factor contributes a smaller amount. With double the number of factors, high obliqueness, and a limited number of items, it would be difficult to distinguish between unidimensionality and multidimensionality. However, that difficulty would have little practical effect because such a multifactor test—for applied purposes—measures a single dominant dimension, and thus becomes essentially unidimensional.

Table 5 contains results for PI. This index is affected by the parameters similarly to LII, but there are both attractive and unattractive differences. One attraction is that the accuracy of PI is sometimes increased as the range of item difficulties increases. It is also less adversely affected by the increase from 2 to 5 factors in many combinations of parameters. However, presence of large interactions involving sample size, factor obliqueness, and item difficulties is an unattractive attribute, especially because sample size can adversely affect discrimination accuracy. Nevertheless, numerous combinations of parameters can be found for which PI is equal or superior in accuracy to LII. In addition, PI can be readily computed. When second factor loadings were weighted by the separation (distance) of the item from the center of the difficulty distribution, further modest increases in accuracy were obtained for wide distributions of item difficulties for a few combinations of parameters.

Table 6 contains overlap results for RDI from variance-covariance matrices. This index is the easiest of the three to compute, but it has little else to recommend it. It shares with PI the problem of interaction with *N*. However, it was on average less accurate. Exceptions to this trend are clustered in the small sample rows. The accuracy obtained from the eigenvalues of variance-covariance matrices, however, was superior to that from eigenvalues from tetrachorics. This finding extends to all but large *N* and the narrow range of item difficulties (Tucker, Humphreys, & Roznowski, 1986). Therefore, the results indicate that the eigenvalues of variance-covariance matrices provide the most accurate index in the family representing the "root staring" criterion of dimensionality, and that this best representative is quite inaccurate under a wide range of parameters.

Evaluation of Indices

Because the data were generated in a complete factorial design, analysis of variance was used to evaluate the strengths and weaknesses of the indices. Table 7 contains sums of squares for main effects.[3] All main effects were significant

[3]Means and standard deviations for the three indices are available from the first author.

Table 4
Discrimination of One Versus Multiple Factors by LII for 20, 30, 40, 50, and 60 Items and N = 125, 500, and 2,000, and for Low, Intermediate, and High Levels of Obliqueness and Tests With Narrow, Medium, and Wide Distributions of Item Difficulty

	Low			Intermediate			High		
1 Versus	Narrow	Medium	Wide	Narrow	Medium	Wide	Narrow	Medium	Wide
20 Items, N = 125									
2	188	186	174	173	169	145	149	132	121
3	181	174	158	149	143	133	127	116	118
4	163	162	148	141	135	133	126	113	110
5	160	164	152	139	132	132	121	108	108
20 Items, N = 500									
2	200	200	200	200	199	193	198	194	182
3	199	198	197	191	195	183	190	183	159
4	196	195	193	188	192	173	176	172	145
5	193	197	183	175	181	155	166	154	136
20 Items, N = 2,000									
2	200	200	200	200	200	197	200	200	195
3	200	200	200	200	200	196	200	197	191
4	200	200	198	199	200	194	199	199	181
5	200	200	196	199	200	189	199	195	180
30 Items, N = 125									
2	197	196	190	188	178	165	164	151	131
3	192	186	182	178	160	158	146	123	124
4	185	181	177	156	144	136	131	119	117
5	173	172	169	149	135	145	125	115	107
30 Items, N = 500									
2	200	200	200	200	200	200	199	200	197
3	200	200	200	196	200	195	195	188	177
4	198	199	199	194	195	188	187	195	157
5	195	199	197	192	190	176	170	171	154
30 Items, N = 2,000									
2	200	200	200	200	200	200	200	200	200
3	200	200	200	200	200	200	200	200	196
4	200	200	199	199	200	199	200	200	192
5	200	200	200	200	200	196	199	200	189
40 Items, N = 125									
2	197	199	196	195	189	181	175	168	140
3	194	191	187	184	176	165	147	148	123
4	188	183	179	161	164	139	136	142	112
5	177	187	169	156	154	123	128	125	111
40 Items, N = 500									
2	200	200	200	200	200	199	200	199	200
3	200	200	200	200	199	197	198	196	191
4	200	200	200	200	196	196	194	192	167
5	199	197	199	197	197	182	187	172	154
40 Items, N = 2,000									
2	200	200	200	200	200	200	200	200	200
3	200	200	200	200	200	200	200	200	199
4	200	200	200	200	200	199	200	200	198
5	200	200	200	200	200	199	200	200	196

continued on the next page

Table 4, continued

Discrimination of One Versus Multiple Factors by LII for 20, 30, 40, 50, and 60 Items and N = 125, 500, and 2,000, and for Low, Intermediate, and High Levels of Obliqueness and Tests With Narrow, Medium, and Wide Distributions of Item Difficulty

	Low			Intermediate			High		
1 Versus	Narrow	Medium	Wide	Narrow	Medium	Wide	Narrow	Medium	Wide
50 Items, N = 125									
2	200	200	198	196	191	174	176	169	149
3	200	197	187	187	184	162	162	149	131
4	197	191	183	174	168	138	152	133	113
5	187	184	177	170	151	134	141	124	106
50 Items, N = 500									
2	200	200	200	200	200	200	200	200	200
3	200	200	200	200	200	199	199	200	192
4	200	200	199	199	200	198	192	195	185
5	199	200	198	199	197	191	185	186	174
50 Items, N = 2,000									
2	200	200	200	200	200	200	200	200	200
3	200	200	200	200	200	200	200	200	200
4	200	200	200	200	200	200	200	200	200
5	200	200	200	200	200	200	200	200	200
60 Items, N = 125									
2	200	199	199	198	195	182	189	176	160
3	198	200	196	191	192	172	171	152	133
4	195	195	187	184	175	156	162	137	118
5	191	193	182	176	155	140	146	123	114
60 Items, N = 500									
2	200	200	200	200	200	200	200	200	200
3	200	200	200	200	200	200	200	200	198
4	200	200	200	200	200	199	199	199	187
5	199	200	200	200	199	197	192	190	176
60 Items, N = 2,000									
2	200	200	200	200	200	200	200	200	200
3	200	200	200	200	200	200	200	200	200
4	200	200	200	200	200	200	200	200	200
5	200	200	200	200	200	200	200	200	199

at $p < .01$, except for number of items using RDI. The mean square profiles were quite different for the indices. LII was more robust overall to change in levels of the parameters than were the other indices. However, all indices showed the expected sensitivity to sample size, obliqueness of factors, and number of items. LII was more affected by N than by any other source of variance. It was also more affected by sample size than were the other two indices, although these were also affected to some degree by N. PI was affected considerably by factor intercorrelations. RI was sensitive to number of factors, factor obliqueness, and item difficulty distribution. The less an index is affected by parameter changes, the more accurate is interpolation in reaching a decision about dimensionality in data that do not precisely match the parameters investigated here.

Given the low level of item correlations in tests of cognitive ability, important design considerations are large samples and a very large ratio of number of variables to number of factors by ordinary factor analytic standards. The level of factor correlations can be controlled only indirectly through careful item and test construction practices. Correlations can be increased much more

Table 5
Discrimination of One Versus Multiple Factors by PI for 20, 30, 40, 50, and 60 Items and N = 125, 500, and 2,000, and for Low, Intermediate, and High Levels of Obliqueness and Tests With Narrow, Medium, and Wide Distributions of Item Difficulty

	Low			Intermediate			High		
1 Versus	Narrow	Medium	Wide	Narrow	Medium	Wide	Narrow	Medium	Wide
20 Items, N = 125									
2	170	157	171	144	135	153	119	123	132
3	157	148	169	146	121	137	121	109	142
4	141	147	168	131	131	153	119	112	125
5	147	143	166	133	132	154	117	111	144
20 Items, N = 500									
2	190	193	195	177	181	187	170	172	159
3	186	190	198	169	180	174	159	155	138
4	181	185	195	158	174	163	147	147	131
5	171	186	194	164	169	166	145	154	136
20 Items, N = 2,000									
2	199	200	197	196	200	189	197	193	151
3	199	200	196	196	200	173	190	178	143
4	200	200	196	193	198	170	185	173	124
5	193	200	193	194	196	162	178	169	129
30 Items, N = 125									
2	172	173	189	167	149	168	140	128	139
3	169	161	186	157	144	170	117	116	130
4	159	153	190	143	127	161	122	114	138
5	157	149	176	132	125	162	110	112	134
30 Items, N = 500									
2	197	198	200	195	194	193	185	184	160
3	199	196	200	189	189	189	177	177	143
4	193	193	199	178	184	176	163	154	125
5	191	197	198	183	180	166	163	143	123
30 Items, N = 2,000									
2	200	200	200	200	200	190	199	200	159
3	199	200	200	199	198	188	199	187	132
4	199	200	198	199	199	172	196	182	115
5	199	200	193	196	194	153	194	176	105
40 Items, N = 125									
2	190	183	195	173	172	181	155	145	154
3	180	175	192	167	162	181	137	133	142
4	171	176	190	149	151	175	133	129	134
5	158	169	188	134	151	166	120	122	133
40 Items, N = 500									
2	200	200	200	198	200	189	193	193	154
3	199	199	199	197	198	192	187	187	149
4	196	200	200	190	194	169	181	170	121
5	197	199	195	190	190	163	172	147	108
40 Items, N = 2,000									
2	200	200	200	200	200	193	200	198	152
3	200	200	200	200	200	183	199	198	128
4	200	200	200	200	200	168	198	181	108
5	200	200	198	199	199	156	195	174	117

continued on the next page

Table 5, continued

Discrimination of One Versus Multiple Factors by PI for 20, 30, 40, 50, and 60 Items and N = 125, 500, and 2,000, and for Low, Intermediate, and High Levels of Obliqueness and Tests With Narrow, Medium, and Wide Distributions of Item Difficulty

	Low			Intermediate			High		
1 Versus	Narrow	Medium	Wide	Narrow	Medium	Wide	Narrow	Medium	Wide
50 Items, N = 125									
2	191	194	194	193	176	176	163	154	141
3	190	181	194	175	175	175	150	140	128
4	178	180	194	165	166	167	141	127	133
5	180	177	193	157	152	169	129	132	129
50 Items, N = 500									
2	200	200	200	199	200	193	196	198	154
3	200	200	200	198	200	187	192	192	137
4	199	200	200	197	197	182	186	179	122
5	200	200	198	191	196	161	176	170	121
50 Items, N = 2,000									
2	200	200	200	200	200	197	200	200	166
3	200	200	199	200	200	190	200	197	132
4	200	200	200	200	200	182	200	192	114
5	200	200	199	200	200	163	199	175	107
60 Items, N = 125									
2	193	194	198	187	182	185	171	169	152
3	190	192	198	176	180	176	149	156	135
4	189	189	196	166	170	172	144	143	127
5	178	179	196	162	170	160	127	132	129
60 Items, N = 500									
2	200	200	200	200	200	196	199	198	154
3	200	200	200	200	200	190	199	194	129
4	199	200	200	199	200	175	198	185	120
5	200	200	197	199	197	152	191	162	106
60 Items, N = 2,000									
2	200	200	200	200	200	195	200	200	149
3	200	200	200	200	200	188	200	191	129
4	200	200	200	200	200	169	200	183	106
5	200	200	199	200	200	157	200	170	102

easily than they can be decreased; the latter would typically be the desired outcome. If a test is not unidimensional in spite of an item writer's best efforts, the multiple factors will probably be highly intercorrelated.

Table 7 also shows η^2 values that were computed relative to the total sums of squares for the main effects total and all interactions. The table shows that the three-way and higher-order interactions were relatively small, even for LII, which has the smallest total sum of squares. However, the relatively large two-way interactions are troublesome and make interpolation difficult.

The data in Table 8 show that the sensitivity of the indices to sample size carries through to the two-, three-, and four-way interactions. Sample size is primarily responsible for the two-way interactions. Furthermore, N even adds to the size of the already small higher-order interactions. The sensitivity to sample size of LII produced substantial interactions. However, these interactions are largely spurious due to the scale of measurement of the separation index; the dependent measure did not contain adequate "top" for LII. Had an index with adequate "top" been used, the large two-way interactions would like-

Table 6
Discrimination of One Versus Multiple Factors by RDI for 20, 30, 40, 50, and 60 Items, N = 125, 500, and 2,000, for Low, Intermediate, and High Levels of Obliqueness, and Tests With Narrow, Medium, and Wide Distributions of Item Difficulty

	Low			Intermediate			High		
1 Versus	Narrow	Medium	Wide	Narrow	Medium	Wide	Narrow	Medium	Wide
20 Items, N = 125									
2	194	189	174	173	169	152	134	122	121
3	191	182	170	163	155	144	135	113	117
4	185	175	168	156	137	145	123	117	120
5	181	173	172	147	142	147	114	113	114
20 Items, N = 500									
2	200	200	193	200	187	170	191	179	148
3	200	200	190	198	188	162	173	157	117
4	198	192	179	186	182	145	156	146	106
5	195	193	168	178	166	130	141	122	101
20 Items, N = 2,000									
2	200	200	198	200	200	180	199	195	146
3	200	200	194	200	199	164	193	175	117
4	200	197	182	191	188	139	180	163	106
5	198	196	173	189	180	122	170	146	101
30 Items, N = 125									
2	200	199	185	190	175	162	164	138	129
3	199	193	183	186	160	157	141	119	113
4	194	183	172	167	145	136	129	110	112
5	186	170	162	163	141	129	125	108	102
30 Items, N = 500									
2	200	200	198	200	199	181	198	192	141
3	200	200	196	200	196	155	194	176	106
4	198	196	176	193	186	132	167	142	103
5	199	189	170	187	174	108	152	132	103
30 Items, N = 2,000									
2	200	200	200	200	200	183	200	197	154
3	200	200	193	200	198	159	200	176	102
4	200	197	165	199	184	119	188	146	100
5	199	194	144	195	166	102	174	139	100
40 Items, N = 125									
2	199	200	194	195	191	176	180	166	128
3	198	199	188	192	180	153	159	134	115
4	194	186	180	180	168	142	141	125	113
5	185	183	167	164	149	126	115	109	102
40 Items, N = 500									
2	200	200	197	200	199	178	200	195	141
3	200	200	194	200	200	153	197	174	105
4	199	198	174	195	188	118	180	157	100
5	200	195	145	193	175	101	169	142	100
40 Items, N = 2,000									
2	200	200	200	200	200	184	200	199	154
3	200	200	195	200	197	139	199	172	100
4	199	197	163	197	185	102	185	140	100
5	200	190	135	189	164	101	167	106	100

continued on the next page

Table 6, continued
Discrimination of One Versus Multiple Factors by PI for 20, 30, 40, 50, and 60 Items, N = 125, 500, and 2,000, for Low, Intermediate, and High Levels of Obliqueness, and Tests With Narrow, Medium, and Wide Distributions of Item Difficulty

	Low			Intermediate			High		
1 Versus	Narrow	Medium	Wide	Narrow	Medium	Wide	Narrow	Medium	Wide
50 Items, N = 125									
2	200	200	199	198	197	172	178	181	142
3	200	199	194	200	188	155	165	159	115
4	198	195	175	180	177	129	145	134	101
5	194	195	161	174	159	118	132	127	101
50 Items, N = 500									
2	200	200	200	200	200	183	200	200	141
3	200	200	192	200	200	141	198	179	101
4	200	198	172	197	185	107	188	148	100
5	200	191	142	189	174	100	172	130	100
50 Items, N = 2,000									
2	200	200	200	200	200	193	200	200	149
3	200	200	196	200	198	128	200	169	100
4	200	198	139	191	177	103	180	122	100
5	198	187	132	187	150	100	173	101	100
60 Items, N = 125									
2	200	200	198	200	196	176	193	181	145
3	200	200	192	197	193	152	174	154	110
4	199	196	169	184	170	128	159	141	107
5	199	189	164	176	162	118	135	125	103
60 Items, N = 500									
2	200	200	200	200	200	189	200	194	145
3	200	200	195	200	198	140	200	174	100
4	199	194	163	199	183	102	186	149	100
5	198	192	137	189	171	102	172	117	100
60 Items, N = 2,000									
2	200	200	200	200	200	191	200	199	141
3	200	200	195	200	195	128	199	163	100
4	199	194	141	194	171	102	178	117	100
5	197	190	117	186	146	100	153	100	100

ly largely disappear and the small higher-order interactions would become even smaller.

Discussion

This research has documented the complexity of determining dimensionality for binary items. The results show that it is important to have an index that is both robust to changes in levels of parameters and lacks substantial interactions among parameters. No single index examined performed best under all combinations of parameters. However, previous research points to one index that can be rejected without reservation (Tucker et al., 1986). That index is the difference in size of eigenvalues of the first two principal factors, and it is not recommended for use. Furthermore, tetrachorics are probably not dependable for any purpose when there is a wide range of difficulties, except in sample sizes substantially larger than 2,000.

The three indices examined here are fairly quick and inexpensive to compute. These characteristics make their use, as well as further work on them, attractive. An index of dimensionality cannot be recommended for use with small samples, small numbers of items, or realistic levels of

Table 7
ANOVA Mean Squares, Degrees of Freedom (*df*), and η^2 Ratios for LII, PI, and RDI Indices

		Mean Square			η^2		
Source	*df*	LII	PI	RDI	LII	PI	RDI
Main Effects	13				.745	.691	.776
Number of Items	4	3,822	5,044	236			
Factor Correlation	2	20,459	59,580	89,407			
Difficulty	2	4,123	8,712	87,733			
Sample Size	2	75,858	46,990	6,887			
Number of Factors	3	5,651	6,368	31,093			
Interactions							
2-way	66	1,020	1,360	1,333	.215	.230	.148
3-way	164	46	157	196	.024	.066	.054
4-way	200	20	20	61	.013	.010	.020
5-way	96	9	12	15	.003	.003	.002
Error	526	152	229	254			
Total	539						

obliqueness of multiple factors. The local independence index is recommended for large samples and many items. The pattern index is recommended for obtaining a quick estimate of dimensionality, except with high obliqueness, a wide distribution of item difficulties, and large sample sizes.

It may be questioned whether the data generation model is more complex than data likely to be encountered in practice. The answer to this question is negative. Investigators typically find clear-cut simple structure in analyses of continuous measures only when they start with a good deal of information about their measures and then carefully select a battery of tests on the basis of that information. Just as factorially pure tests are not common, factorially pure items are also rare.

Intercorrelations among ability factors defined by continuous variates also tend to be high in wide ranges of talent. For example, there are stable factor correlations of about .55 in the Armed Services Vocational Aptitude Battery (ASVAB), for which the factors are defined by tests having quite dissimilar content (e.g., Arithmetic Reasoning and Mathematics Knowledge on the one hand, and Vocabulary and Reading Comprehension on the other). Not only is it unlikely that

Table 8
Mean Squares (*MS*), Degrees of Freedom (*df*), and η^2 Ratios for the ANOVA Interactions for Those Involving *N* Versus All Others

	2-way			3-way			4-way		
Source	*MS*	*df*	η^2	*MS*	*df*	η^2	*MS*	*df*	η^2
Local Independence									
Involving *N*	2,754	22	.194	78	88	.022	25	152	.012
All others	153	44	.022	10	76	.002	5	48	.001
Pattern Index									
Involving *N*	2,046	22	.115	224	88	.050	21	152	.008
All others	1,017	44	.115	79	76	.015	16	48	.002
Ratio Index									
Involving *N*	1,435	22	.053	156	88	.023	57	152	.015
All others	1,282	44	.095	242	76	.031	73	48	.006

these four types of items would be found in a single item pool, it is also unlikely that the two types that defined separate factors among total scores would be found together. The selective factors imposed on item pools by the test constructor's conceptualization of the test are likely to produce high levels of obliqueness among multiple factors.

Finally, the emphasis in IRT applications on equivalent measurement accuracy at all θ levels requires that item pools have a wide range of difficulties. Without such a range, parameters are not well estimated. Indices of dimensionality are expected to be applied to pools of binary items—many of which are factorially complex, have highly correlated multiple factors, and include items that are widely different in levels of difficulty.

When cognitive data are viewed in terms of a hierarchical factor model, no serious error is committed by accepting a unidimensional hypothesis in the presence of multiple factors that are substantially oblique. Error becomes progressively less serious as the number of factors and factor intercorrelations increase. Both number of factors and factor intercorrelations increase the *general* factor's contribution to total score variance. The *group* factor's contribution to total score variance decreases as the number of factors increases. Consequently, the contribution to total variance from the sum of five factors is less than the overall contribution from the sum of two group factors of the same degree of obliqueness. It seems counterintuitive, but as long as each item measures the general factor and the factorial complexity of the test items is high, the total score will more closely reflect a single dimension. In most applications, the most valid dimension is the one defined by the factor correlations (Humphreys, 1985a). Furthermore, a test constructor should avoid choosing to measure only one of the correlated factors.

Properly weighted multiple dimensions that are substantially positively correlated are not a problem for IRT when each examinee is exposed to every item, as is the case in a standard printed test. Multiple dimensions do become a problem, however, in adaptive testing. The expected build-up of the general factor variance in the total score occurs only when all secondary factors are adequately sampled. An algorithm for the selection of items that depends only on the *a* and *b* item parameters cannot be relied on. Secondary factors must be known or estimated, and that information used in item selection in order to maximize the validity of the test and to avoid the bias that results if all examinees are not measured on the same dominant dimension.

References

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

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale NJ: Erlbaum.

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

Hattie, J. A. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9,* 139–164.

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

Humphreys, L. G. (1985a). Correlations in psychological research. In D. K. Detterman (Ed.), *Current topics in human intelligence: Vol. 1. Research methodology* (pp. 1–24). Norwood NJ: Ablex.

Humphreys, L. G. (1985b). General intelligence: An integration of factor, test, and simplex theory. In B. B. Wolman (Ed.), *Handbook of intelligence: Theories, measurements, and applications* (pp. 201–224). New York: Wiley.

Humphreys, L. G., & Montanelli, R. G., Jr. (1975). An investigation of the parallel analysis criterion for determining the number of common factors. *Multivariate Behavioral Research, 10,* 193–205.

Lazarsfeld, P. F. (1959). Latent structure analysis. In S. Koch (Ed.), *Psychology: A study of a science. Vol. 3* (pp. 476–542). New York: McGraw-Hill.

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

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

McDonald, R. P. (1981). The dimensionality of tests

and items. *British Journal of Mathematical and Statistical Psychology, 34,* 100–117.

McDonald. R. P. (1985). *Factor analysis and related methods.* Hillsdale NJ: Erlbaum.

Montanelli, R. G., Jr., & Humphreys, L. G. (1976). Latent roots of random data correlation matrices with squared multiple correlations on the diagonal: A Monte Carlo study. *Psychometrika, 41,* 341–348.

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

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

Stout, W. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika, 52,* 589–618.

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

Thurstone, L. L. (1947). *Multiple factor analysis.* Chicago IL: University of Chicago Press.

Tucker, L. R (1983). Searching for structure in binary data. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement: A Festschrift for Frederic M. Lord* (pp. 215–235). Hillsdale NJ: Erlbaum.

Tucker, L. R, Humphreys, L. G., & Roznowski, M. (1986). *Comparative accuracy of five indices of dimensionality of binary items* (Tech. Rep. No. 1). Champaign IL: University of Illinois, Department of Psychology.

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.

van den Wollenberg, A. L. (1982). Two new test statistics for the Rasch model. *Psychometrika, 47,* 123–140.

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.

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

Acknowledgments

This research was sponsored by Personnel and Training Research Programs, Psychological Sciences Division, Office of Naval Research under Contract No. N00014-84-K-0186. The authors thank Tim Davey for his help and many suggestions on this research.

Author's Address

Send requests for reprints or further information to Mary Roznowski, Department of Psychology, Ohio State University, 1885 Neil Avenue, Columbus OH 43210-1222, U.S.A.