An Assessment of Stout's Index of Essential Unidimensionality

John Hattie, The University of North Carolina at Greensboro
Krzysztof Krakowski, The University of Western Australia
H. Jane Rogers, Teachers College, Columbia University
Hariharan Swaminathan, University of Massachusetts

A simulation study was conducted to evaluate the dependability of Stout's *T* index of unidimensionality as used in his DIMTEST procedure. DIMTEST was found to dependably provide indications of unidimensionality, to be reasonably robust, and to allow for a practical demarcation between one and many dimensions. The procedure was not affected by the method used to identify the initial subset of unidimensional items. It was, however, found to be sensitive to whether the multidimensional data arose from a compensatory model or a partially compensatory model. DIMTEST failed when the matrix of tetrachoric correlations was non-Gramian and hence is not appropriate in such cases. *Index terms: DIMTEST, essential unidimensionality, factor analysis, item response models, Stout's test of unidimensionality, tetrachoric correlations, unidimensionality.*

A fundamental assumption of test theory is that a score can only have meaning if the set of items measures only one attribute or dimension. If the measuring instrument is composed of items that measure different dimensions, then it is difficult to interpret the total score from a set of items, to make psychological sense when relating variables, or to interpret individual differences. Despite the importance of this assumption to all testing models, there have been few systematic attempts to investigate this assumption and, until recently, little success at providing a defensible procedure to assess the claim of unidimensionality. Hattie (1984, 1985) theoretically and empirically assessed over 30 indexes of unidimensionality and noted the inadequacies of most of these indexes. Hattie suggested that procedures be based on defensible theory and that unidimensionality be examined in the framework of local independence (Lord & Novick, 1968; McDonald, 1981).

Local independence requires that for fixed trait level $\boldsymbol{\theta}$ (i.e., conditional on a vector of traits, $\boldsymbol{\theta}$), the responses of an individual to different items are statistically independent. Lord & 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 person's value on the trait is determined, the behavior is random, in the sense of statistical independence. (p. 538)

This principle of local independence provides a mathematical definition of latent traits; $\boldsymbol{\theta}$ can be interpreted as a set of traits that the items measure in common. Once these trait values are fixed at a given value (i.e., conditioned on), the responses to items become statistically independent. Thus, in order to determine the dimensionality of a set of items it is necessary and sufficient to identify the minimal set of traits such that at all fixed levels of these traits the item responses are independent. This principle applies to linear as well as nonlinear re-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 20, No. 1, March 1996, pp. 1–14

0146-6216/96/010001-14$1.95

gression functions of observed item responses on the trait values.

The requirement that item responses be statistically independent for fixed values of the traits is very stringent, because it requires that for fixed values of the traits, not only the covariances be 0, but that all higher-order moments be products of the univariate moments. McDonald (1979) has suggested that this "strong" principle of local independence can be replaced with a "weak" principle of local independence, by requiring that only the covariances among the items be 0 for fixed values of the traits. Note that when the item responses (conditional on the trait values) have a multivariate normal density, the weak principle implies the stronger principle; hence, in this case the two principles are equivalent.

Assessing Essential Unidimensionality With DIMTEST

Stout (1987, 1990) used this weaker form of local independence to develop his arguments for "essential unidimensionality." He devised a statistical index, embodied in his DIMTEST procedure, based on the fundamental principle that local independence should hold approximately when sampling from a subpopulation of examinees of approximately equal θ level. According to Stout (1987), a test $(U_1, ..., U_N)$ of length N is said to be essentially unidimensional if there exists a latent variable θ such that for all values of θ,

$$\frac{1}{N(N-1)} \sum_{1<i\neq j<N} \left|\mathrm{Cov}\left(U_i, U_j|\theta\right)\right| \approx 0. \quad (1)$$

That is, on average, the conditional covariances over all item pairs must be small in magnitude (for more details regarding the theoretical developments see Junker, 1990, 1991, 1992; Nandakumar, 1987, 1991; Nandakumar & Stout, 1993; Stout, 1990). Essential unidimensionality can therefore be thought of as an empirical operationalization of the weak principle of local independence.

Stout (1990) then developed an empirical notion of unidimensionality to match his definition of essential unidimensionality. Either a subjective analysis of item content or an exploratory factor analysis is used to develop a core set of items, which is termed the *assessment subtest*. The remaining set of items, termed the *partitioning subtest*, is used to partition examinees into groups for a stratified analysis. When the total set of items is unidimensional, then the assessment and partitioning tests are both unidimensional, but when the dimensionality is greater than 1, then "the partitioning subtest will contain many items that load heavily on at least one other dimension not measured by the assessment subtest" (Stout, 1987, pp. 591–592).

The Four Steps in DIMTEST

Step 1. A core set of M items is selected from the test so that these items are as unidimensional as possible. This is called Assessment Subtest 1 (AT1). There are three suggested procedures for identifying these M items: (1) an expert or judgmental analysis can be used to define the unidimensional set, (2) a principal components solution of the tetrachoric correlation matrix over all N items can be calculated, or (3) a cluster analysis (Roussos, Stout, & Marden, 1993) can be used. The M items loading most highly on the second unrotated factor are selected.

Step 2. A second set of M items is selected from the remaining items such that they are similar in difficulty and dimensionality to the items in AT1. This is called Assessment Subtest 2 (AT2). AT2 is later used to correct the T statistic for bias, because the mean shifts in the positive direction for all short tests as a consequence of selecting items that are overly homogeneous with respect to difficulty.

Step 3. The remaining $n = N - 2M$ items comprise the set by which the examinees are scored and then partitioned into subgroups, called the Partitioning Subtest (PT). For the strong principle of local independence, the statistics in Step 4 should be calculated on the basis of the number of examinees who have the same subtest score; however, because the number within each subtest-score group typically is too small, Stout (1987) recommended that respondents be assigned to groups on the basis of total score such that a large number (approximately 20) are in each group.

Step 4. For each of these subgroups, the variance estimates ($\hat{\sigma}_k^2$ and $\hat{\sigma}_{U,k}^2$) and the standard error of estimate (S_k) are computed using the AT1 items (see

Nandakumar & Stout, 1993, Equation 3). The usual variance estimate, $\hat{\sigma}_k^2$, is the observed variance of the AT1 subtest and is sensitive to departures of unidimensionality. The "unidimensional" variance estimate, $\hat{\sigma}_{U,k}^2$, is the summed variances across the k groups and remains the same regardless of the dimensionality of the dataset. These estimates are then summed across K subgroups to obtain

$$T_L = \frac{1}{\sqrt{K}} \sum_{k=1}^{K} \frac{\left|\hat{\sigma}_k^2 - \hat{\sigma}_{U,k}^2\right|}{S_k}. \quad (2)$$

A similar statistic, T_B, is calculated from the items in AT2. T_L is a measure of the amount of multidimensionality present locally for subgroup k. If unidimensionality holds for the subgroup k examinees, then T_L should equal 0 except for statistical error. This is analogous to subdividing examinees into their subtest score groups and then asking whether the principle of local independence holds. T_L is computed from the M items in AT1 and is sensitive to dimensionality and sources of bias; T_B is based on the maximally similar set in AT2 and is sensitive to sources of bias but not dimensionality. Thus, T_B is used to correct for bias. The final statistic is

$$T = \frac{T_L - T_B}{\sqrt{2}}. \quad (3)$$

The basic principle underlying DIMTEST is that if unidimensionality holds then the basic item response theory (IRT) model assumption of local (conditional) independence holds approximately within each examinee subgroup; hence, the two within-subgroup variance estimates should be approximately equal. Stout (1987) demonstrated that T is asymptotically normally distributed when unidimensionality holds. Stout's T can be used to test the hypothesis H_0: $d_E = 1$ versus the alternative hypothesis H_1: $d_E > 1$ where d_E is essential dimensionality. The null hypothesis is rejected if T is greater than or equal to z with some specified α level.

Stout recommends modifications to T when the sample size is moderate (i.e., > 2,000) or large (i.e., > 40,000). For moderate samples, for example, examinees are divided into Q subsets (at least 20) and the Q value is calculated using the above four steps separately for each of these subsets. T is then the sum of these separate Q values divided by the square of the number of subsets (Stout, 1987, p. 596). In the present simulation, the formula for small samples was used because the sample size was 1,000.

Performance of the DIMTEST T Statistic

Stout (1987) conducted a simulation to assess how well the nominal level of significance is approximated by the actual level of significance and how much power the procedure displayed when $d_E = 1$ and $d_E = 2$. He varied four factors: the general form of the item response functions, examinee population size ($J = 750$ or $J = 20{,}000$), assessment subtest size, and the number of dimensions. T was sufficiently powerful to detect or reject essential unidimensionality, although the rejection rates were less desirable when there was guessing.

Nandakumar (1991) compared the Holland (1981; Holland & Rosenbaum, 1986) procedure based on assessing whether the items are conditionally positively associated (and noted the close similarities to DIMTEST; see also Stout, 1987), linear factor analysis, nonlinear factor analysis (NLFA), and DIMTEST. She reported that NLFA methods (with a one-factor quadratic) accurately recovered the dimensionality when the correlations among the θs were low (< .5), but both the nonlinear and Holland and Rosenbaum procedures were not as effective when the correlation between θs was high ($\rho > .5$). The linear factor methods were not adequate for assessing dimensionality. DIMTEST always correctly confirmed the dimensionality of the simulated datasets (when $d_E = 1$ or 2).

Nandakumar & Stout (1993) found that T was a poor indicator for testing H_0: $d_E = 1$ when there was high discrimination ($a > 1.1$) and guessing. They found that in such situations, the easiest items (including those with high guessing parameters and low discrimination) tended to be selected for AT1 and, because the respondents are grouped on the basis of the items in the PT, there was much misclassification of low θ examinees. After rejecting NLFA as a method for diminishing the influence of difficulty on the second factor loadings (in the one real dataset they used they still found a "diffi-

culty factor"), they suggested a correction procedure to ensure a greater match of the factor loadings (and thus discriminations) across the set of items chosen in AT1 and in PT. Given this new statistic (T corrected for bias using AT2), their simulations indicated that there was much improvement in rejection rates when $d_E = 1$ or 2.

A simulation study by de Champlain & Gessaroli (1991) found that the accuracy of T was affected by both sample size and test length. T performed best on tests with more than 25 items (as expected by Stout) and with sample sizes greater than 500. They also found, as did Hattie (1984), that the fit statistics based on the incremental fit of the proportion of the sum of squares of the residuals after fitting a one-factor quadratic (using NLFA) was also very effective and less sensitive to the number of items and sample size.

There are a number of issues that arise with respect to DIMTEST. They include the use of tetrachoric correlations, the identification of the items to comprise AT1, and the methods for constructing multidimensional data.

Calculation of Tetrachoric Correlations

The selection of items for AT1 in DIMTEST depends on the correct calculation of the tetrachoric correlations, and it is well-known that sample-based estimates of the tetrachoric correlation matrix are often not positive definite (see Lord & Novick, 1968, p. 349). The DIMTEST program (Stout, Nandakumar, Junker, Chang, & Steidinger, 1991) does not indicate the number of nonpositive definite matrices nor the effects of these matrices on the subsequent statistics. The problem of nonpositive definite matrices typically occurs when one of the correct or incorrect cells of the two-by-two item response tables contains a 0 or a value near 0 (see Lord, 1980; Pearson, 1901). Many formulas have been suggested for calculating tetrachorics to overcome this problem, and the major issue relates to the number of terms to be used in the infinite series: McNemar (1955) used the first four terms, Elderton (1906) the first seven, and Christoffersson (1975) and Muthén (1978) the first ten terms.

A further problem with the use of tetrachorics is that they are inappropriate when θ distributions are not normal (Lord, 1980). This is likely to occur when there is guessing. Carroll (1945) demonstrated a procedure for adjusting the proportions to more correctly estimate the tetrachoric correlations, but this correction is not used in DIMTEST.

Identification of the Core Set

A second major problem relates to the core set of items identified for AT1, because it is critical that it is accurately determined. The identification procedure suggested by Stout uses the principal components method, and the choice of items is based on second factor loadings because linear factor models typically lead to the first factor being defined as a "difficulty factor." McDonald & Ahlawat (1974) convincingly demonstrated that this patterning of the first factor is an artifact of using the incorrect linear factor model instead of the more correct nonlinear model. They demonstrated that data generated by the normal ogive model should yield spurious factors due to nonlinearity, but these will tend to be "... negligible unless the items vary widely in difficulty level, and/or we have sharply discriminating items that approximate a perfect scale" (p. 98). In general, they argued that the term "factors due to difficulty" should not be used but should be replaced by the notion of factors due to nonlinearity. Stout claims to minimize the effect of these nonlinearities by using AT2, which selects items using a somewhat similar difficulty grouping and using tetrachorics.

A procedure that seems more appropriate in the context of IRT models would be to use NLFA (Etezadi-Amoli & McDonald, 1983). Nandakumar & Stout (1993) used NLFA on a set of real data that, when using linear factor analysis, led to factors due to nonlinearity. To their "... surprise, the difficulty factor reappeared even with the nonlinear factor analysis" (Nandakumar & Stout, 1993, p. 50). Therefore, they did not implement or recommend the use of NLFA in DIMTEST. However, in their particular dataset there were many high discriminating items ($a > 1.5$) and much guessing ($c > .15$), which can lead to the presence of many nonlinearities (Gourlay, 1951; Hulin, Drasgow, & Parsons, 1983). NLFA models can re-

duce or eliminate the effects of "difficulty" factors and have many other advantages (Etezadi-Amoli & McDonald, 1983; McDonald, 1982).

Constructing Multidimensional Data

In all cases in which DIMTEST has been used, the data have been simulated by a compensatory model (CM) (de Champlain & Gessaroli, 1991; Nandakumar, 1991; Nandakumar & Stout, 1993; Roussos et al., 1993). There are many other constructions of multidimensional data as Coombs (1954), following Johnson (1935), demonstrated. Coombs defined three methods of constructing multidimensional data: the conjunctive model, in which an excess of one trait, no matter how large, does not compensate for lower trait levels in other dimensions; the disjunctive model, in which an examinee will pass an item if he/she is dominant over the item in any one dimension and will fail only if the item dominates him/her in all dimensions; and the CM, in which an examinee's response to an item is a function of a weighted sum of underlying abilities. The conjunctive and the disjunctive models are psychologically distinct but are isomorphic to each other mathematically. A multidimensional, three-parameter CM was outlined in Hattie (1984) in which the probability of a correct response is specified by

$$P\left(x_{ij}=1\middle|a_i,b_i,\theta_j\right) = c_i + \frac{1-c_i}{1+\exp\left[-d\sum_d\left(\mathbf{a}_{id}\boldsymbol{\theta}_{jd}-\mathbf{b}_{id}\right)\right]}, \quad (4)$$

where

$P(x_{ij})$ is the probability of a correct response to item i by person j;

$\mathbf{a}_{id}$ is a vector of discrimination parameters for item i on dimension d;

$\mathbf{b}_{id}$ is the difficulty parameter for item i on dimension d (although strictly there is only one b parameter, see Reckase, Ackerman, & Carlson, 1988);

c_i is the guessing parameter for item i; and

$\boldsymbol{\theta}_{jd}$ is a vector of trait parameters for person j on dimension d.

Other researchers (notably Ackerman, 1987, 1992; Ansley & Forsyth, 1985; Reckase et al., 1988; Wang, 1987, 1988) have demonstrated that with multidimensional CMs the univariate calibration of two-dimensional response data can be explained in terms of the interaction between the multidimensional test information and the distributions of the two traits. Various interpretations of the multidimensional item difficulty, multidimensional discrimination, and multidimensional item information have been suggested.

Sympson (1978) proposed a partially compensatory model (PCM) in which a decrease in one trait could only be offset by a large increase in the other trait (this model has often been erroneously called a noncompensatory model). Outside of a relatively narrow trait range, the probability of correctly answering the item reduces to 0 regardless of the value of the stronger trait (see Lord, 1984). The multidimensional PCM can be represented as:

$$P\left(x_{ij}=1\middle|a_i,b_i,\theta_j\right) = c_i + \left(1-c_i\right)\prod_d \frac{1}{1+\exp\left[-d\left(\mathbf{a}_{id}\boldsymbol{\theta}_{jd}-\mathbf{b}_{id}\right)\right]}. \quad (5)$$

The probability of a correct response is simply the product of probabilities for each dimension (but see Coughlan, 1974; Jannarone, 1986). It is not known what effect the choice of underlying multidimensional model has on DIMTEST's T statistic.

Purpose

This study was concerned with five questions about DIMTEST:

1. Does DIMTEST satisfactorily identify a unidimensional versus a multidimensional model for varying values of discrimination, correlation between dimensions, and guessing or the spread of difficulty?
2. Can T distinguish between dimensionality when the dimensions are related using CMs or PCMs?
3. Although not required or argued by Stout, the third question concerned whether T was monotonically related to the number of dimensions.
4. The study evaluated the effects of a different method of calculating tetrachoric correlations, and the effects on T when T is based on non-

Gramian matrices.

5. The effects of using NLFA to determine AT1 and thus to calculate the T statistic were investigated.

Method

The program by which simulated data were created provided control over the choice of model, the number of dimensions, the number of items, the difficulty range, the discrimination, the correlation between dimensions, and the amount of guessing. The first dataset included 35 items from a unidimensional domain ($d = 1$), with discriminations (a) all equal to 1.0, and two levels of guessing ($c = 0$ or $c = .15$). There were two difficulty (b) ranges—[−2, −1, 0, 1, 2] and [−1, −.5, 0, .5, 1]. For each of the four data combinations (2 difficulty × 2 guessing), 15 sets of data were generated each based on 1,000 examinees with θ normally distributed. A three-parameter IRT model was used to simulate performance on these items.

The second dataset, a two-dimensional case ($d = 2$), included 18 items from the first dimension and 17 items from a second dimension. The third dataset, a three-dimensional case ($d = 3$), included 12 items from the first dimension, 11 items from a second dimension, and 11 items from a third dimension. For both the $d = 2$ and $d = 3$ datasets, discrimination values of $a = 1$ were formed into either a two- or three-factor simple structure pattern and multiplied by a triangular decomposition matrix based on intercorrelations of .1, .3, or .5 between the two factors. These values were selected to reflect cases in which this was an almost orthogonal relationship between the dimensions ($\rho = .1$) to a case in which there was much overlap ($\rho = .5$). This procedure and the relation between discrimination values and factor loadings are explained in Hattie (1984).

For both the second and third datasets, two b ranges were selected [−1, −.5, 0, .5, 1] or [−2, −1, 0, 1, 2] and two levels of c ($c = 0$ or $c = .15$). $\rho = .1$, .3, or .5 between the dimensions (which relates to the discrimination of the items; see Hattie, 1984; Hattie & Krakowski, 1994) and the type of model (compensatory using Equation 4 or partially compensatory using Equation 5) were varied. For each permutation of the two- and three-dimension case, 15 sets of data were generated. Thus, there were 2 dimensions × 2 difficulty ranges × 2 levels of guessing × 3 levels of ρ × 2 methods of generating data × 15 datasets = 720 datasets. The responses of 1,000 examinees were simulated for each of these 720 combinations.

Data were generated using the DIMENSION program (see Hattie & Krakowski, 1994). The method for calculating tetrachorics outlined by Kirk (1973) was used, which is based on a Gaussian (8-point) quadrature supplemented by Newton-Raphson iteration. This is a more refined method for estimating tetrachorics compared to the method used in the DIMTEST program (Stout et al., 1991).

Three methods—DIMTEST, refined tetrachorics (RT), and NLFA—were compared to investigate their accuracy to assess H_0: $d_E = 1$. This comparison was undertaken in five steps. First, the nature of the items selected by the three methods for AT1 was investigated to determine whether there were any patterns in how the methods selected items. Second, the frequency and the effect of having nonpositive definite matrices were assessed. Third, the relationships among the T indexes were examined. Fourth, the factors affecting T were investigated using an analysis of variance (ANOVA) design. Fifth, rejection rates for the various indexes were examined.

Results

Choice of Items

Table 1 presents the number of items selected for AT1 when $d = 1$ and the percentages of these selected items within the two b ranges [−1, 1] or [−2, 2] and for the two levels of c (0, .15) across the three methods of analysis (DIMTEST, RT, and NLFA). These percentages are based on the total number of items selected for each permutation divided by the total number of items that the procedure used to classify the items.

There were 2 (b) × 2 (c) × 15 (replications) × 35 (items) = 2,100 possible items; 611 of these items were selected in AT1. When, for example, b was the most negative (either −1 for the datasets with a b range from −1 to 1, or −2 for the datasets with a b range from −2 to 2), 205 of these items were selected

Table 1
Number of Items Selected (No.) by DIMTEST, RT, and NLFA When $d = 1$ and Percentages for Each Item Type for $b = [-1,1]$ and $[-2,2]$, and $c = 0.0$ and .15

		DIMTEST					RT					NLFA				
			[−1, 1]		[−2, 2]			[−1, 1]		[−2, 2]			[−1, 1]		[−2, 2]	
Simulated Difficulty	No.	%	0	.15	0	.15	%	0	.15	0	.15	%	0	.15	0	.15
−2 or −1	205	34	32	22	44	39	31	29	22	31	41	7	7	14	1	5
−1 or −.5	47	8	10	13	5	4	7	8	14	3	3	19	18	27	10	23
0 or 0	76	7	16	24	1	4	8	22	24	2	4	56	49	40	76	65
1 or .5	84	14	19	13	10	13	15	21	13	13	12	13	18	13	13	8
2 or 1	199	32	23	28	39	38	34	20	28	50	37	4	9	8	1	0

for AT1. Thus, 34% of the 611 items selected for AT1 had an extremely negative b. When $b = [-1, 1]$ and $c = 0$, 32% of the items were $b = -1$, 10% were $b = -.5$, 16% were $b = 0$, 19% were $b = .5$, and 23% were $b = 1$. It was expected, when the data were generated with $d = 1$, that each item would have an equal probability of being selected for AT1. This was not the case, as evidenced in Table 1.

DIMTEST and RT selected items with more extreme bs (e.g., of the 66% for DIMTEST, 34% were the most negative bs and 32% were the most positive bs) and tended not to select items from the middle of the b distribution. By contrast, for NLFA, 56% of the selected items had $b = 0$.

Thus, DIMTEST, which is based on a principal components analysis, tended to select items that led to maximizing the variance (i.e., the more extreme items); the NLFA procedure, which is based on maximum likelihood, tended to select items that provided best fit. The RT procedure differed little from DIMTEST, although it was slightly less affected by extreme items.

When there were two or three underlying dimensions, it was expected that the various procedures would select the AT1 items from the 15 items that defined the dominant dimension. Table 2 presents the summary statistics for $d = 2$ and $d = 3$. For $d = 2$, DIMTEST (98%) and RT (97%) were more likely than NLFA (89%) to select items from a single dimension. The varying levels of b, c, and model (CM or PCM) made little difference in the percentage of occasions that AT1 items with a single dimension were selected across all three procedures.

Similarly, for $d = 3$, there were few differences across the methods; however, the percentage of selecting all AT1 items from the same dimension decreased markedly for $d = 3$ compared to $d = 2$. Also, the effects of the varying levels of b, c, and model were more apparent than when $d = 2$. The percentage of selecting the correct AT1 items diminished with an increased range of b (e.g., for DIMTEST there was a decrease from 82% for b in the range $[-1, 1]$ to 73% for b in the range $[-2, 2]$), higher intercorrelations between the dimensions (81% for $\rho = .1$, 75%

Table 2
Percentage of Times That AT1 Items Were Selected From Within a Single Dimension for DIMTEST, NLFA, and RT for Levels of ρ and c, Ranges of b, and Models (PCM or CM) With $d = 1$ and $d = 2$

Number of Factors and Method	ρ			b		c		Model	
	.1	.3	.5	[−1, 1]	[−2, 2]	0	.15	PCM	CM
$d = 2$									
DIMTEST	98	97	92	99	92	97	94	92	99
RT	97	96	93	99	92	96	95	91	99
NLFA	89	85	78	83	85	82	85	79	88
$d = 3$									
DIMTEST	81	75	76	82	73	78	77	71	84
RT	78	73	74	80	70	73	76	68	82
NLFA	72	72	70	71	71	73	70	68	74

for ρ = .3, and 76% for ρ = .5), and when the CM was used rather than the PCM (84% compared to 71%).

Positive Definiteness

Using tetrachoric calculations as computed by DIMTEST, 13% of all the matrices were not positive definite. As found in many other studies (Carroll, 1945, 1961; Hattie, 1984, 1985; Roznowski, Tucker, & Humphreys, 1991), the majority of these cases occurred when there was much guessing, with the PCM, and with $b = [-1,1]$ (see Table 3). In such cases, it is more likely that an item will occur that is very easy (i.e., that every examinee will answer correctly), and this can lead to major difficulties in computing the tetrachoric correlation. Consequently, the tetrachorics are poorly estimated. Although not apparent in the present simulation, the opposite case (i.e., every examinee will answer the item incorrectly) can also lead to a high incidence of non-Gramian matrices (see Hattie, 1984).

Table 3
Percentage of Matrices That Were Nonpositive Definite

Source	Percent
$d = 1$	3
$d = 2$	
ρ = .1	21
ρ = .3	24
ρ = .5	23
$d = 3$	
ρ = .1	8
ρ = .3	8
ρ = .5	13
Difficulty	
[−1, 1]	66
[−2, 2]	34
Guessing	
0	0
.15	100
Model	
PCM	96
CM	4

The effect of nonpositive definiteness on T was dramatic. The average T was .52 when the matrices were not positive definite, and 2.64 when positive definite. The breakdown by number of dimensions is presented in Table 4. Mean T values were less than 2 for $d = 2$ and $d = 3$ (across all possible ρs) when the tetrachoric correlation matrix was not positive definite. These values indicate that the datasets were essentially unidimensional. Mean Ts were always greater than 2 (indicating more than one dimension) for $d = 2$ and $d = 3$ when the matrix was positive definite. Thus, when the matrix is not positive definite, T should not be used. In the DIMTEST program (Stout et al., 1991), there is a message indicating the presence of very small frequencies and these values are replaced with .005. The program then continues through the subsequent steps (this procedure was followed here). Clearly, this correction is inadequate. Thus, in the following analyses only the indexes based on the positive definite matrices were used for the tetrachoric methods.

Relationships Among the T Indexes

Across all simulations, the correlation between T based on DIMTEST's tetrachorics and the RT method was .92, DIMTEST Ts and the NLFA Ts $r = .62$, and for RT and NLFA $r = .67$. There was much variability, with rs as low as .10, between DIMTEST and RT and DIMTEST and NLFA. These results suggest that the three methods produce values of T that might lead to different conclusions.

Factors Affecting T

Table 5 presents mean Ts for the different methods for two ranges of b and two levels of c, as well as levels of d, ρ, and the PCM and CM. As ρ increased from ρ = .1 to ρ = .3, T was less able to detect the correct dimensionality for all methods (e.g., for DIMTEST mean T was 2.58 for ρ = .1 and 3.04 for ρ = .3). When ρ = .5, mean T typically decreased. Mean T for the PCM was considerably smaller than mean T for the CM for all three methods (e.g., for DIMTEST, .50 vs. 4.29). For the PCM, mean T always indicated that the data were unidimensional. The hypothesis of only one dominant dimension was less likely to be rejected for DIMTEST when $b = [-1, 1]$ (mean $T = 1.97$) rather than when $b = [-2, 2]$ (mean $T = 3.39$) or when $c = 0$ (mean $T = 2.45$) versus when $c = .15$ (mean $T = 2.79$). All three estimation procedures were appropriately sensitive to the number of dimensions (for DIMTEST mean $T = -.06$, 4.48, and

Table 4
Mean T and Percentage of Matrices That Were Nonpositive Definite For Varying Levels of d and ρ

Type of Matrix, Mean T, and Percentage	$d = 1$	$d = 2$			$d = 3$		
		$\rho = .1$	$\rho = .3$	$\rho = .5$	$\rho = .1$	$\rho = .3$	$\rho = .5$
Not Positive Definite							
Mean T	−.13	1.48	.63	.08	.42	.07	−.03
Percent	3.30	19.20	20.80	20.00	6.70	7.50	10.80
Positive Definite							
Mean T	−.06	4.48	3.41	2.53	3.67	2.73	2.19
Percent	96.70	80.80	79.20	80.00	93.30	92.50	89.20

3.41 for $d = 1$, 2, and 3, respectively).

An ANOVA was used to assess the relative effects of the various parameters on T. Because ρ was nested within the number of dimensions, these were entered as a single effect with 7 levels: $d = 1$; $d = 2, \rho = 1$; $d = 2, \rho = .3$; $d = 2, \rho = .5$; $d = 3, \rho = 1$; $d = 3, \rho = .3$; and $d = 3, \rho = .5$.

The ANOVA in Table 6 shows that the pattern of mean squares was similar for all three procedures. Most of the variance was accounted for by the data generation model (CM or PCM) and by the number of underlying dimensions. Thus, T was sensitive to whether the data were compensatory or partially compensatory and to the number of dimensions, and much less sensitive to variations in b, correlation between the dimensions, and c. Unlike many competing indexes of unidimensionality, T was most sensitive to dimensionality.

Table 5
Mean T for DIMTEST, RT, and NLFA for Levels of ρ, c, d, Ranges of b, and Models

Total and Parameter	DIMTEST	RT	NLFA
Total	2.64	2.20	2.15
Correlation			
$\rho = .1$	2.58	2.32	2.18
$\rho = .3$	3.04	2.60	2.53
$\rho = .5$	2.35	1.64	1.72
Difficulty			
$b = [-1,1]$	3.39	2.91	2.28
$b = [-2,2]$	1.97	1.50	2.02
Guessing			
$c = 0$	2.79	2.46	2.37
$c = .15$	2.45	1.95	1.92
Number of Dimensions			
$d = 1$	−.06	−.14	.31
$d = 2$	4.48	2.85	2.28
$d = 3$	3.41	2.34	2.63
Model			
PCM	.50	.48	.52
CM	4.29	3.93	3.78

Rejection Rates

Table 7 presents the rejection rates for testing the null hypothesis that there is one underlying dimension (i.e, that $d_E = 1$). For $d = 1$, the null case was likely to be rejected at more than the expected 5% level. For multiple dimensions ($d = 2$ or 3), the procedures were likely to reject the one-dimensional case between 76% to 89% of the time.

Table 8 presents these rejection rates for the levels of ρ and c, and for the two ranges of b. For CM, the hypothesis that $d_E = 1$ was appropriately rejected most of the time. For the PCM, the hypothesis that $d = 1$ was rejected far less than expected ($\alpha = .05$) under most conditions, especially when $d = 3$ (between 0% to 20%). This finding was not surprising given that the increased number of dimensions increases the chance that a small probability correct on one of the three dimensions would cause the multiplying effect to lead to a reduced probability correct (regardless of the ability on the other dimensions).

For the CM, T from DIMTEST was more likely to reject this hypothesis more frequently than T from RT or NLFA. For example, when $d = 2$, $\rho = .1$, $c = 0$, and $b = [-2, 2]$, DIMTEST rejected 100% of the datasets, compared with 93% for RT, and 80% for NLFA. The differences between the three methods were most marked when there was a high correlation between the dimensions ($\rho = .5$). In such cases,

Table 6
Results of ANOVA on *T* for DIMTEST, RT, and NLFA

		DIMTEST			RT			NLFA		
Source of Variation	*df*	MS	*F*	*p*	MS	*F*	*p*	MS	*F*	*p*
d	6	125.66	63.11	<.001	129.75	60.30	<.001	122.62	42.16	<.001
b	1	20.29	10.19	<.001	41.53	19.30	<.001	14.60	5.02	.025
c	1	29.52	14.83	<.001	11.27	5.24	.022	42.25	14.53	<.001
Model (PCM/CM)	1	417.99	209.91	<.001	340.62	158.30	<.002	229.95	766.63	<.001
d × *b*	6	3.87	1.95	.071	9.44	4.39	<.001	8.78	3.02	.006
d × *c*	6	7.05	3.54	.002	4.87	2.27	.036	3.37	1.16	.327
d × Model	6	110.76	55.62	<.001	111.38	51.76	<.001	126.69	43.55	<.001
b × *c*	1	4.77	2.40	.122	10.25	4.76	.029	.44	.15	.699
b × Model	1	27.31	13.72	<.001	41.08	19.09	<.001	47.45	16.31	<.001
c × Model	1	.86	.43	.511	.21	.10	.757	2.46	.84	.358
d × *b* × *c*	6	5.58	2.80	.011	6.81	3.16	.005	1.63	.56	.761
d × *b* × Model	6	8.22	4.13	<.001	20.08	9.33	<.001	3.12	1.07	.377
d × *c* × Model	6	6.72	3.38	.003	4.12	1.91	.076	4.80	1.65	.130
b × *c* × Model	1	1.82	.91	.340	.01	0.00	.953	11.52	3.96	.047
d × *b* × *c* × Model	6	4.80	2.41	.066	4.01	1.87	.134	1.41	.49	.819
Within		1.99 (*df* = 681)			2.15 (*df* = 681)			2.91 (*df* = 784)		

T from DIMTEST more correctly rejected the hypothesis that $d_E = 1$ both when $d = 2$ and when $d = 3$. For example, when $d = 2$, $\rho = .5$, $c = .15$, and $b = [-2, 2]$, DIMTEST rejected 73% of the datasets, compared with 27% for RT, and 47% for NLFA.

Discussion

DIMTEST is based on the weaker principle of local independence and is designed not to identify whether a set of items is or is not unidimensional, but whether there is a sufficiently dominant dimension such that the test user can proceed to meaningfully interpret a single total score across the set of items. This study, however, identified some concerns with DIMTEST. The most important is the nature of the data; that is, whether the multidimensional data conform to a compensatory or partially compensatory model. DIMTEST is only applicable for identifying compensatory multidimensional data. *T* did not discriminate between the various dimensions in the partially compensatory case, probably because of problems in estimating the tetrachorics.

Table 7
Rejection Rates (%) for Testing H_o: $d_E = 1$ for DIMTEST, RT, and NLFA

	d		
Approach	1	2	3
DIMTEST	15	89	78
RT	15	88	79
NLFA	15	85	76

Various methods (e.g., the size of the interaction factor using NLFA, or the size of the correlations between dimensions under the different models) were examined to determine whether a dataset was compensatory or partially compensatory, but little success was achieved; thus, a careful judgmental analysis of the nature of success on the items, a cognitive processing analysis of the competencies required to correctly answer the items, or more attention to partially compensatory estimation procedures are warranted. Given that the majority of instances of positive-definiteness came from partially compensatory data, DIMTEST should not be used or interpreted for this type of data.

Furthermore, DIMTEST assesses only essential unidimensionality and does not claim to identify the resulting dimension(s). It may be that because of the choice of the items for ATI the method only identifies a "bloated specific" [i.e., a factor indexed by a series of items that are slight variants of each other (Cattell, 1964, 1978)]. Given Humphreys' (1986; Roznowski et al., 1991) admonition that useful tests are rarely unidimensional at the lower-order factor

Table 8
Percent Rejection Rates (%) for Testing H_o: $d_E = 1$ Based on DIMTEST, RT, and NLFA

ρ, *c*, and Method	*d* = 2				*d* = 3			
	PCM		CM		PCM		CM	
	[−1,1]	[−2,2]	[−1,1]	[−2,2]	[−1,1]	[−2,2]	[−1,1]	[−2,2]
ρ = .1, *c* = 0								
DIMTEST	80	80	100	100	20	0	100	100
RT	80	100	100	93	0	0	100	100
NLFA	46	66	77	80	7	20	100	93
ρ = .1, *c* = .15								
DIMTEST	73	13	100	93	0	0	100	100
RT	73	20	100	93	0	7	100	93
NLFA	46	33	80	93	10	0	100	93
ρ = .3, *c* = 0								
DIMTEST	60	20	100	77	7	0	93	87
RT	60	20	100	93	7	7	93	80
NLFA	20	60	73	80	0	0	100	77
ρ = .3, *c* = .15								
DIMTEST	20	20	100	100	0	0	100	60
RT	20	27	100	100	0	0	100	53
NLFA	20	7	80	87	0	0	100	93
ρ = .5, *c* = 0								
DIMTEST	47	13	93	93	0	7	100	93
RT	47	13	93	100	0	13	100	27
NLFA	13	27	53	53	0	0	100	80
ρ = .5, *c* = .15								
DIMTEST	7	27	93	73	0	0	100	80
RT	7	7	93	27	0	0	100	27
NLFA	0	7	53	47	7	0	87	27

level, it is important that the test user goes beyond the statistical methods and attempts to clearly identify and defend the interpretation of the set of items. Moreover, because many tests claiming to be unidimensional include few items, a major improvement would be to adapt DIMTEST for shorter tests (< 25 items). At minimum, more simulations of the performance of DIMTEST with shorter tests would be of much value.

An improvement to DIMTEST would be to improve the estimation of the tetrachoric correlations. The problems of accurately estimating the tetrachorics are primarily a function of the existence of cells with 0 values and the number of terms in the estimation series. Methods programmed by Christoffersson (1975) and Muthén (1978) were most successful in earlier assessments of indexes of unidimensionality (Hattie, 1984, 1985). Muthén's methods are most similar to those (using a different algorithm and set of principles) in NOHARMII (Fraser, 1988); however, in this study this refined tetrachoric method was not as effective at detecting essential unidimensionality as DIMTEST.

A matrix of sample tetrachorics is often non-Gramian. The RT method used here as an alternative to DIMTEST did not appreciably improve the performance of *T* or reduce the number of nonpositive definite matrices. One possibility is to include an indicator as to the number of tetrachorics that approach 0 (e.g., the number of tetrachorics less than .15), but earlier work using this method was not encouraging. Hulin et al. (1983) and Carroll (1945) suggested modifications to address these problems, and they may be valuable to include in future simulations. Whatever method is used to estimate the tetrachorics, it is critical that programs to calculate *T* include a warning if the matrix is not positive definite and proceed no further.

The incorporation of a nonlinear method in the first step did not lead to an improvement over

DIMTEST. The nonlinear methods tended to select a different subset of items for AT1, compared to the tetrachoric-based methods. DIMTEST selected more discriminating items for AT1, and it can be claimed that such items are more discriminating because of the possible presence of dimensions other than the "essential" dimension. The nonlinear methods were not as effective in discriminating between unidimensional and multidimensional datasets, although the mean levels of T under the different methods were similar. DIMTEST was less affected by other parameters than the nonlinear method.

A surprising finding was that, except for the nonlinear method, T was not monotonically related to the underlying dimensionality. This lack of monotonicity resulted primarily from the effects of the partially compensatory models in which T decreased most dramatically for three compared to two dimensions (and in most cases the mean was lower than when there was one dimension). This, again, highlights the problems of using T for partially compensatory data. The differences between T for two and three dimensions for the compensatory model were minimal (4.80 vs. 5.14, respectively) although the indexes were much greater than in the one-dimensional case. Thus, T can only be used to assess essential unidimensionality and should not be used as a general index of dimensionality.

It is difficult for any index to detect unidimensionality given the myriad of possibilities for other deviations to affect an index. For example, many fit statistics based on the Rasch model attempt to detect deviations from unidimensionality as well as the presence of guessing and the deviation from a common discrimination (Rogers & Hattie, 1987; Traub & Wolfe, 1981). DIMTEST is robust to deviations from most sources and at the same time detects deviations from unidimensionality.

Future Directions

There are two competing directions in which research on assessing unidimensionality can proceed. First, multidimensional IRT models can be developed to better estimate the parameters of multidimensional data. It is already known that using unidimensional IRT models to estimate parameters when the data are truly multidimensional is problematic. For example, Ackerman (1987) reported negative correlations between estimated item difficulty and item discrimination estimates with their true values when multidimensional data parameters were estimated using unidimensional models. Others have attempted to provide heuristics to interpret multidimensional data using unidimensional calibration methods (Luecht & Miller, 1991), whereas others have provided defensible statistics (Ackerman, 1992; Carlson, 1987; Junker, 1991, 1992; McKinley & Reckase, 1983; Reckase et al., 1988). If such multidimensional models were to be developed, it is difficult to imagine how they could be useful, given the myriad of ways that a person's performance on items could be weighted to attain the item response (but see Tam, 1992).

The second approach is to develop procedures for detecting unidimensionality so that better essentially unidimensional tests can be developed. Such procedures could include developing multidimensional estimation programs and using DIMTEST. By appropriate use of (nonlinear) factor analysis (to reduce the problem, as much as possible, to a one- vs. two-dimension problem), and the use of DIMTEST, more defensible essentially unidimensional sets can be developed. Methods are being developed for creating essentially unidimensional tests, by beginning with a core set of items that can be demonstrated to be essentially unidimensional using expert judgment, NLFA, and other methods of validity assessment (see Maguire, Hattie, & Haig, 1993; McDonald & Mulaik, 1979). Items then are added sequentially such that at each step the new augmented set is evaluated for essential unidimensionality using DIMTEST. The resulting tests are more likely to be unidimensional and relate to meaningful and dependable constructs.

References

Ackerman, T. (1987, April). *The robustness of LOGIST and BILOG IRT estimation programs to violations of local independence.* Paper presented at the annual meeting of the American Educational Research Association, Washington DC.

Ackerman, T. A. (1992). A didactic explanation of item

bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29,* 67–91.

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

Carlson, J. E. (1987). *Multidimensional item response theory estimation: A computer program.* Iowa City IA: American College Testing Program.

Carroll, J. B. (1945). The nature of the data, or how to choose a correlation coefficient between items or between tests. *Psychometrika, 10,* 1–19.

Carroll, J. B. (1961). The nature of data, or how to choose a correlation coefficient. *Psychometrika, 26,* 347–372.

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.

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

Coombs, C. H. (1954). *A theory of data.* New York: Wiley.

Coughlan, J. R. (1974). *A multidimensional extension of the normal ogive, logistic and linear latent trait model.* Unpublished master's thesis, University of Toronto, Toronto, Canada.

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

Elderton, W. R. (1906). *Frequency curves and correlation.* Washington DC: Harren.

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

Fraser, C. (1988). *NOHARMII: A Fortran program for fitting unidimensional and multidimensional normal ogive models of latent trait theory.* Armidale, Australia: The University of New England.

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

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

Hattie, J. A. (1984). Decision criteria for assessing unidimensionality: An empirical study. *Multivariate Behavioral Research, 19,* 49–78.

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

Hattie, J. A., & Krakowski, K. (1994). DIMENSION: A program to generate unidimensional and multidimensional item data. *Applied Psychological Measurement, 17,* 252.

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

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

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

Humphreys, L. G. (1986). An analysis and evaluation of test and item bias in the prediction context. *Journal of Applied Psychology, 71,* 327–333.

Jannarone, R. J. (1986). Conjunctive item response theory kernels. *Psychometrika, 51,* 357–373.

Johnson, H. M. (1935). Some neglected principles in aptitude testing. *American Journal of Psychology, 47,* 159–165.

Junker, B. W. (1990). *Progress in characterizing strictly unidimensional IRT representations* (Technical Rep. No. 498). Pittsburgh PA: Carnegie Mellon University, Department of Statistics.

Junker, B. W. (1991). Essential independence and likelihood-based ability estimation for polytomous items. *Psychometrika, 56,* 255–278.

Junker, B. (1992, April). *Ability estimation in unidimensional models when more than one trait is present.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

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

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

Lord, F. M. (1984). *Conjunctive and disjunctive item response functions* (ETS Report No. 150-520). Princeton NJ: Educational Testing Service.

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

Luecht, R. M., & Miller, T. R. (1991, April). *Unidimensional calibrations and interpretations of multidimensional tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Maguire, T. O., Hattie, J. A., & Haig, B. (1993). Construct validity and achievement assessment. *The Alberta Journal of Educational Research, 40,* 109–126.

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 Statisti-*

cal 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., & Mulaik, S. A. (1979). Determinacy of common factors: A nontechnical review. *Psychological Bulletin, 43,* 289–374.

McKinley, R. L., & Reckase, M. D. (1983). MAXLOG: A computer program for estimating the parameters of a multidimensional extension of the two-parameter logistic model. *Behavior Research Methods and Instrumentation, 15,* 389–390.

McNemar, Q. (1955). Opinion-attitude methodology. *Psychological Bulletin, 43,* 289–374.

Muthén, B. (1978). Contributions to factor analysis of dichotomous items. *Psychometrika, 30,* 419–440.

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

Nandakumar, R. (1991, April). *Assessing dimensionality of a set of items—comparison of different approaches.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Nandakumar, R., & Stout, W. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics, 18,* 41–68.

Pearson, K. (1901). On the correlation of characters not quantitatively measurable. *Royal Society Philosophical Transactions, 195,* (Series A), 1–47.

Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building unidimensional tests using multidimensional items. *Journal of Educational Measurement, 25,* 193–203.

Rogers, H. J., & Hattie J. A. (1987). A monte carlo evaluation of several person and item fit statistics in latent trait models. *Applied Psychological Measurement, 2,* 47–58.

Roussos, L. A., Stout, W. F., & Marden, J. I. (1993). *Analysis of the multidimensional structure of standardized tests using DIMTEST with hierarchical cluster analysis.* Unpublished manuscript, University of Illinois, Department of Statistics, Champaign.

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

Stout, W. F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52,* 589–617.

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

Stout, W. F., Nandakumar, R., Junker, B., Chang, H. H., & Steidinger, D. (1991). *DIMTEST and TESTSIM* [Computer program]. Champaign: University of Illinois, Department of Statistics.

Sympson, J. B. (1978). A model for testing with multidimensional items. In D. J. Weiss (Ed.), *Proceedings of the 1977 computerized adaptive testing conference* (pp. 82–98). Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

Tam, S. S. (1992). *A comparison of methods for adaptive estimation of a multidimensional trait.* Unpublished doctoral dissertation, Graduate School of Arts and Sciences, Columbia NY.

Traub, R. E., & Wolfe, R. G. (1981). Latent trait theories and the assessment of educational achievement. In D. C. Berliner (Ed.), *Review of research in education* (Vol. 9) (pp. 377–435). Washington DC: American Educational Research Association.

Wang, M. M. (1987, April). *Estimation of ability parameters from response data to items that are precalibrated with a unidimensional model.* Paper presented at the annual meeting of the American Educational Research Association, Washington DC.

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

Acknowledgments

This research was sponsored by an Australian Research Grant. The authors thank W. Stout for his helpful consultations in the design and improvements to this article, and two anonymous reviewers for their suggestions.

Author's Address

Send requests for reprints or further information to John Hattie, University of North Carolina, Greensboro NC 27412-5001, U.S.A. Internet: hattiej@dewey.uncg.edu.