Assessing Essential Unidimensionality of Real Data

Ratna Nandakumar

University of Delaware

The capability of DIMTEST in assessing essential unidimensionality of item responses to real tests was investigated. DIMTEST found that some test data fit an essentially unidimensional model and other data did not. Essentially unidimensional test data identified by DIMTEST then were combined to form two-dimensional test data. The power of Stout's statistic T was examined for these two-dimensional data. DIMTEST results on real tests replicated findings from simulated tests—T discriminated well between essentially unidimensional and multidimensional tests. T was also highly sensitive to major traits and insensitive to relatively minor traits that influenced item responses. *Index terms: DIMTEST, essential unidimensionality, essential independence, multidimensionality, unidimensionality.*

Most of the currently used item response theory (IRT) models assume unidimensionality. From the strict IRT perspective, unidimensionality refers to one, and only one, trait underlying test items. Yet, item responses are multiply determined (Hambleton & Swaminathan, 1985, chap. 2; Humphreys, 1981, 1985, 1986; Reckase, 1979, 1985; Stout, 1987; Traub, 1983). Therefore, from a substantive viewpoint, the assumption of unidimensionality requires that the test items measure one dominant trait. Stout (1987) coined the term *essential unidimensionality* to refer to a particular mathematical formulation of a test having exactly one dominant trait. Dimensionality is, however, determined by the joint influence of test items and examinees taking the test (Reckase, 1990). In addition, extraneous factors such as teaching methods and anxiety level of examinees also may influence the dimensionality of item response data. Thus, dimensionality has to be assessed each time a test is administered to a new group of examinees.

Traditionally, factor analysis has been the most popular approach for assessing dimensionality (Hambleton & Traub, 1973; Lumsden 1961). Despite its serious limitations in analyzing dichotomous data (e.g., see Hulin, Drasgow, & Parsons, 1983, chap. 8), factor analysis has been the primary method used to study the robustness of the unidimensionality assumption (Drasgow & Parsons 1983; Harrison, 1986; Reckase, 1979). A number of other promising methods have been proposed and used in varying degrees to assess dimensionality, such as full information factor analysis based on the principle of marginal maximum likelihood (Bock, Gibbons, & Muraki, 1988; Wilson, Wood, & Gibbons, 1983); nonlinear factor analysis (Etazadi-Amoli & McDonald, 1983; McDonald, 1962; McDonald & Ahlawat, 1974); Holland & Rosenbaum's (1986) test of unidimensionality, monotonicity, and conditional independence based on contingency tables; Roznowski, Tucker, & Humphreys' (1991) methods based on the principle of local independence and second factor loadings; and Stout's (1987) statistical procedure based on essential independence and essential dimensionality. Hattie (1984, 1985) provided a comprehensive review of traditional approaches used to assess dimensionality, and Zwick (1987) applied some of these procedures to assess dimensionality of the National Assessment of Educational Progress data. Despite having several procedures available that assess dimensionality, there is no preferred method among researchers and there is often dissatisfaction in assessing dimensionality (Berger & Knol, 1990; Hambleton & Rovinelli, 1986; Hattie, 1985).

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 17, No. 1, March 1993, pp. 29-38

0146-6216/93/010029-10$1.75

Stout (1987) proposed a statistical test (DIMTEST) to assess essential unidimensionality of the latent space underlying a set of items. Nandakumar (1987) and Nandakumar & Stout (1993) have modified, refined, and validated DIMTEST further for assessing essential unidimensionality on a variety of simulated tests. This study demonstrates the validity and usefulness of Stout's procedure on real, as opposed to simulated, tests. Essentially unidimensional data were combined to form two-dimensional data. The power of Stout's statistic T was examined for these two-dimensional data.

DIMTEST for Assessing Essential Unidimensionality

DIMTEST, a statistical test for assessing unidimensionality, is based on the theory of essential dimensionality and essential independence (Stout, 1987, 1990). An item pool is essentially independent with respect to the latent trait vector Θ if, for a given initial segment of the item pool, the average absolute conditional (on Θ) covariances of item pairs approach 0 as the length of the segment increases. When only one dominant θ meets the essential independence assumption, the item pool is considered essentially unidimensional.

In contrast, the assumption of local independence requires the conditional covariances to be 0 for all item pairs. The number of traits required to satisfy the local independence assumption is the dimensionality of the test. To satisfy the assumption of local independence in the traditional definition of dimensionality (Lord & Novick, 1968), all traits required to respond to test items correctly are counted; however, essential dimensionality counts only dominant traits required to satisfy the assumption of essential independence (as opposed to local independence). Using this definition, DIMTEST assesses the fit of the model generating the given item responses to the essentially unidimensional model. Nandakumar (1991) described the theoretical differences between traditional dimensionality and essential dimensionality and established, using monte carlo studies, the usefulness of DIMTEST for assessing essential unidimensionality in the possible presence of several secondary dimensions.

DIMTEST assumes that a group of N examinees takes an L-item test. Each examinee produces a response vector that can be scored as 1s and 0s—1 denotes a correct response, and 0 denotes an incorrect response. It is assumed that essential independence with respect to some dominant trait θ holds and that the item response functions are monotonic with respect to the same trait θ. The hypothesis is stated as follows:

$$H_0: d_E = 1 \text{ versus } H_1: d_E > 1 , \quad (1)$$

where d_E denotes the essential dimensionality of the latent space underlying a set of items.

To assess essential unidimensionality of a given set of test data, DIMTEST follows several steps (for details see Nandakumar & Stout, 1993; Stout 1987).

1. Split the test items into three subtests: AT1, AT2, and PT. The L items are split into two short assessment subtests, AT1 and AT2, and a long partitioning subtest, PT. AT1 items are selected so that they all measure the same dominant trait. The items can be selected using either factor analysis or expert opinion. If factor analysis is used, part of the sample is used (a sample of 500 is recommended) for this purpose. M items with the highest loadings of the same sign on the second factor are selected for AT1. M is automatically determined by DIMTEST as a function of the sample size and the test length. If expert opinion is used to select AT1 items, at most one quarter of the total items that tap the same dominant trait should be selected.

Once the AT1 items are selected, M AT2 items are selected so that they have a difficulty distribution similar to the AT1 items (for details see Stout, 1987). The remaining items ($n = L - 2M$) form the partitioning subtest, PT.

2. Split examinees into subgroups. Examinees are assigned to K different subgroups based on their score on PT. All examinees obtaining the same PT total score are assigned to the same subgroup. When PT is "long" (e.g., 80 items or

more) and the test is essentially unidimensional, within each subgroup k examinees are assumed to be of approximately similar trait levels. When PT is not long, AT2 compensates for the bias in AT1 caused by PT being short. Also, AT2 compensates for the bias in AT1 caused by the presence of guessing or the difficulty factor that is often found by factor analysis.

3. Compute the statistic T. Within each subgroup k, the variance estimates $\hat{\sigma}_k^2$ and $\hat{\sigma}_{Ud,k}^2$ (the unidimensional variance estimate) and the standard error of estimate S_k are computed using item responses of AT1. These estimates are then summed across K subgroups to obtain

$$T_L = \frac{1}{(k)^{1/2}} \sum_{k=1}^{K} \left(\frac{\hat{\sigma}_k^2 - \hat{\sigma}_{Ud,k}^2}{S_k} \right) , \tag{2}$$

where the subscript "L" indicates a long test. The usual variance estimate for subgroup k is given by

$$\hat{\sigma}_k^2 = \sum_{j=1}^{J_k} [Y_j^{(k)} - \bar{Y}^{(k)}]^2 / J_k , \tag{3}$$

where

$$Y_j^{(k)} = \sum_{i=1}^{M} U_{ijk} / M \tag{4}$$

and

$$\bar{Y}^{(k)} = \sum_{j=1}^{J_k} Y_j^{(k)} / J_k . \tag{5}$$

U_{ijk} (1 or 0) denotes the response for item i by examinee j in subgroup k, and J_k denotes the total number of examinees in subgroup k. The "unidimensional" variance estimate for subgroup k is given by

$$\hat{\sigma}_{Ud,k}^2 = \sum_{i=1}^{M} \hat{p}_i^{(k)} [1 - \hat{p}_i^{(k)}] / M^2 , \tag{6}$$

where

$$\hat{p}_i^{(k)} = \sum_{j=1}^{J_k} U_{ijk} / J_k . \tag{7}$$

The standard error of estimate for subgroup k is given by

$$S_k = \{ [(\hat{\mu}_{4,k} - \hat{\sigma}_k^4) + \hat{\delta}_{4,k} / M^4] / J_k \}^{1/2} , \tag{8}$$

where

$$\hat{\mu}_{4,k} = \sum_{j=1}^{J_k} [Y_j^{(k)} - \bar{Y}^{(k)}]^4 / J_k \tag{9}$$

and

$$\hat{\delta}_{4,k} = \sum_{i=1}^{M} \hat{p}_i^{(k)} [1 - \hat{p}_i^{(k)}] [1 - 2\hat{p}_i^{(k)}]^2 . \tag{10}$$

Similarly, the statistic T_B is computed using items of subtest AT2.

Stout's statistic T is then given by

$$T = (T_L - T_B) / (2)^{1/2} . \tag{11}$$

The decision rule is to reject H_o if $T \geq Z_\alpha$, where Z_α is the upper $100(1 - \alpha)$ percentile of the standard normal distribution, and α is the desired level of significance.

The basic underlying principle in assessing dimensionality using DIMTEST is that, when the assumption of essential independence holds within subgroups, the value of T will be "small" and the null hypothesis will be accepted. When the assumption of essential independence fails within subgroups, T will be "large," and the null hypothesis will be rejected. When the given test data are well-modeled by an essentially unidimensional model, items of AT1, AT2, and PT will all be measuring the same dominant dimension. Examinees placed into subgroups based on their PT scores will be of approximately similar trait levels; therefore, their responses to AT1 items will be independent, which leads to the equality of the variance estimates $\hat{\sigma}_k^2$ and $\hat{\sigma}_{Ud,k}^2$. Hence, T will be "small," suggesting the tenability of H_o. On the other hand, when the test data are not well-modeled by an essentially unidimensional model, PT will have items of more than one dominant trait and the assumption of local independence within subgroups will not be tenable. This will result in a large within-group variability with respect to AT1 items. Therefore, the variance estimate $\hat{\sigma}_k^2$ will be much larger than $\hat{\sigma}_{Ud,k}^2$; thus, T will be "large," and H_o will be rejected.

Simulation studies (Nandakumar, 1987; Nandakumar & Stout, 1993; Stout, 1987) on a wide variety of tests have demonstrated the utility of DIMTEST in discriminating between one- and two-dimensional tests. Nandakumar (1991) demonstrated the usefulness of DIMTEST with the aid of a rough index of deviation from essential unidimensionality. The tests in Nandakumar (1991) were modeled by two- and higher-dimensional IRT models, and the test items were influenced by major and secondary traits to varying degrees. For some tests, the secondary trait(s) influenced a high proportion of items, and for others the secondary trait(s) influenced only a small proportion of items. DIMTEST reliably accepted the hypothesis of essential unidimensionality, provided the model generating the test was close to the essentially unidimensional model; this was established when each of the secondary traits influenced relatively few items, or when secondary traits influenced many items and the degree of influence on each item was small. The Type I error in these cases was within tolerance of the nominal error level. As the degree of influence of the secondary traits increases, however, the approximation to an essentially unidimensional model degenerates, inflating the observed Type I error of the hypothesis of essential unidimensionality. The simulation results of Nandakumar and Stout have demonstrated the power of T when the model generating the item responses is two-dimensional (two major traits), with correlations between traits as high as .7 and items jointly influenced by both traits.

Method

Data

The characteristics of all datasets and tests are described in Table 1. The tests were analyzed for essential unidimensionality using DIMTEST. Some of the datasets were assessed to be essentially unidimensional, and others were multidimensional. The nature of the multidimensionality was further explored and essentially unidimensional subtests were derived from multidimensional tests whenever possible. In addition, two-dimensional

Table 1
Number of Items (L), Number of Examinees (N), and Description of the Tests

Test	L	N	Description
1	36	2,428	U.S. History Test
2	31	2,428	Subtest—Test 1 Geography Items Deleted
3	30	2,439	Literature Test
4	30	1,984	Grade 10 ASVAB Arithmetic Reasoning
5	30	1,961	Grade 12 ASVAB Arithmetic Reasoning
6	25	1,981	Grade 10 ASVAB Auto Shop
7	25	1,974	Grade 12 ASVAB Auto Shop
8	25	1,990	Grade 10 ASVAB General Science
9	25	1,988	Grade 12 ASVAB General Science
10	40	2,491	ACT Mathematics Usage
11	40	5,000	ACT Reading
12	30	5,000	Subtest—First 30 Items of Test 11
13	40	5,000	ACT Science
14	28	5,000	Subtest—First 28 Items of Test 13
15	30	750	Subsample of Examinees from Test 12 (Reading)
16	30	1,000	Subsample of Examinees from Test 12 (Reading)
17	30	1,250	Subsample of Examinees from Test 12 (Reading)
18	30	2,000	Subsample of Examinees from Test 12 (Reading)
19	28	750	Subsample of Examinees from Test 14 (Science)
20	28	1,000	Subsample of Examinees from Test 14 (Science)
21	28	1,250	Subsample of Examinees from Test 14 (Science)
22	28	2,000	Subsample of Examinees from Test 14 (Science)

test data were created using essentially unidimensional test data to study the power of DIMTEST.

The datasets used in the present study came from different sources. The U.S. history and literature data were obtained from the test data for examinees in Grade 11 of the 1986 National Assessment of Educational Progress (National Assessment of Educational Progress, 1988) from Educational Testing Service.

The U.S. history test required knowledge of U.S. history (colonization, the Revolutionary War, the Civil War, World Wars I and II, post-World War II) and world geography. The literature test examined knowledge of novels, short stories, and plays; myths, epics, and Biblical characters and stories; poetry; and nonfiction. The Armed Services Vocational Aptitude Battery (ASVAB) data for the Arithmetic Reasoning, Auto Shop, and General Science subtests for Grades 10 and 12 were obtained from Linn, Hastings, Hu, & Ryan (1987). The Arithmetic Reasoning subtest consisted of arithmetic word problems; the Auto Shop subtest examined knowledge of automobiles, tools, and shop terminology and practices; and the General Science test consisted of items that required knowledge in solving high school level physical, life, and earth science problems.

The mathematics usage test data, the reading test data, and the science test data were obtained from the American College Testing (ACT) program. The mathematics usage test required knowledge in solving different types of mathematics problems, such as arithmetic and algebra operations, geometry, numeration, story problems, and advanced topics. The reading test consisted of 4 passages, each followed by 10 questions. The first three passages were taken from books about the humanities, and the last passage was taken from a book about psychology. The science test consisted of 7 passages, each followed by 5 to 7 questions. Each passage was based on a different science topic, such as the characteristics of dinosaurs or the periods of a pendulum and the relationship to string length and ball mass.

To examine the effect of sample size on DIMTEST, both Test 12 and Test 14 were randomly split into four mutually exclusive subsamples. Test 12 was split into Tests 15, 16, 17, and 18; Test 14 was split into Tests 19, 20, 21, and 22.

Two-dimensional test data were created by combining responses from test data that were assessed as essentially unidimensional by DIMTEST. These tests were two-dimensional from a content perspective. Characteristics of the tests, and the associated sample sizes, are summarized in Table 2. For all these tests, item subsets were randomly selected from the indicated sources.

Table 2
Two-Dimensional Tests: Number of Items (L) Used From Each Subtest in Forming the Corresponding Two-Dimensional Test and the Number of Examinees (N)

	Subtest 1		Subtest 2		
Test	Test Number	L	Test Number	L	N
23	15	30	19	6	750
24	16	30	20	6	1,000
25	17	30	21	6	1,250
26	18	30	22	6	2,000
27	12	30	14	6	5,000
28	4	30	8	5	1,853
29	4	30	8	10	1,853
30	9	25	5	5	1,811
31	9	25	5	10	1,811
32	2	31	3	5	2,428
33	2	31	3	8	2,428
34	2	31	3	10	2,428

Results

Unidimensional Studies

Tests 1 through 14 (except Tests 2, 12, and 14, which were derived subtests of Tests 1, 11, and 13, respectively), were tested initially for essential unidimensionality using DIMTEST. In each case, 500 examinees were selected randomly for use in factor analysis to select AT1 and AT2 items. The remainder of the sample was used to compute Stout's statistic T. The size of AT1 (M) also was determined by DIMTEST. For each test, the T value, the p value, and M are reported in Table 3. Table 3 shows that the p values associated with

Tests 3, 4, 5, 8, and 9 were all well above the nominal level of significance ($\alpha = .05$), which indicates the essentially unidimensional nature of these tests. However, the p values associated with Tests 1, 6, 7, 10, 11, and 13, were all well below $\alpha = .05$, which indicates the multidimensional nature of these test data. For these tests, the nature of the multidimensionality was explored.

Table 3
Results of H_o: $d_E = 1$ For Tests, Subtests, and Subsamples (Factor Analysis Was Used to Select AT1 Items)

Test and Subtest	T	p	M
Test			
11	8.7	0.00	10
1	6.2	0.00	6
7	3.6	0.00	5
13	3.2	0.00	12
14	3.0	0.00	5
10	2.8	0.00	10
6	2.3	.01	5
2	1.3	.09	5
8	1.0	.17	5
3	.7	.23	6
5	.6	.26	4
12	.5	.32	7
9	−.3	.60	6
4	−.7	.73	6
Subsample of Test 12 Examinees			
15	.1	.48	5
16	.5	.32	7
17	−.1	.52	7
18	1.0	.16	5
Subsample of Test 14 Examinees			
19	1.8	.03	7
20	3.2	.01	6
21	1.4	.08	7
22	2.9	0.00	7

When test data are essentially unidimensional, the AT1 items are of the same dominant dimension as the rest of the items; therefore, DIMTEST does not reject the null hypothesis. When the test data are not essentially unidimensional, however, the AT1 items are dimensionally different from the rest of the items, and DIMTEST rejects the null hypothesis of essential unidimensionality. Following this reasoning, for tests with very low p values, the content of AT1 items was examined.

For Test 1, Items 12 through 16 and Item 6 were selected for AT1. Items 12 through 16 were homogeneous and differed dimensionally from the rest of the items in Test 1. These 5 items required world geography knowledge, and the rest of the items concerned U.S. history. It was also possible that these items were selected for AT1 due to chance alone. In order to test for this, DIMTEST was applied 100 times; each time the 2,428 examinees were randomly split into two groups of 500 and 1,928 examinees. That is, AT1 items were selected repeatedly on different random samples of 500 examinees each. The resampling results showed that Items 12 through 16 were consistently selected for AT1. In addition to these items, one or two more items, which varied from run to run, were selected from the rest of the items. Hence, it was concluded that the geography items were dimensionally different from the other items. A subtest, Test 2, was formed consisting of all items of Test 1 except for the five geography items. The p value associated with Test 2 ($p = .09$) shows evidence of essential unidimensionality. Furthermore, from the content perspective, items of AT1 did not form a subset that was dimensionally different from the rest of the items of Test 2.

A similar phenomenon was observed with the Test 11 and Test 13 data. For Test 11, the last 10 items (items followed by the last passage) formed part of subtest AT1. Again, these same 10 items formed part of AT1 in repeated sampling applications of DIMTEST. These 10 items tapped the "psychology" content area, which is different from the "literature" content area tapped by the first three passages. Another possibility for the second dimension is that because these were the last 10 items of the reading test, speededness could have had an effect. Based on these observations, it was concluded that these items were dimensionally different from the rest, and a subtest, Test 12, was formed consisting of the first 30 items of Test 11. The p value associated with Test 12 ($p = .32$) shows evidence of

an essentially unidimensional model underlying the data (see Table 3). In addition, items of ATI came from all the passages of Test 11.

For Test 13, the 12 items following the last two passages formed part of ATI. As with Test 1 and Test 11, after repeated sampling applications of DIMTEST, these items were removed. The resulting subtest, Test 14, which consisted of the first 28 items of Test 13, was still found to be multidimensional ($p = 0.00$). Thus, a unidimensional subtest could not be formed. Unlike the reading test items in Test 12, the science test items in Test 14 came from distinctly different content areas, with a moderate correlation among content areas, and required a higher level of abstract reasoning. Thus, in addition to different content areas, difficulty or speededness could have caused major secondary dimensions in Test 14.

For Tests 6, 7, and 10, the p values were low, but the items of ATI did not form a subgroup tapping a secondary trait as found in Tests 1, 11, and 13. In addition, it was found that items in Tests 1, 11, and 13 tapped multiple major content areas. Therefore, these test data were treated as multidimensional.

Table 3 shows dimensionality results for the subsamples of the unidimensional reading test (Test 12) and the multidimensional science test (Test 14). The p values associated with Tests 15 through 18 show evidence of a high degree of essential unidimensionality underlying these tests. These results are consistent with the results of Test 12. The selection of ATI items for Tests 15 through 18 was highly varied, and yet essential unidimensionality was affirmed consistently. The results of Tests 19 through 22, consistent with the results reported for Test 14, affirmed multidimensionality of these test data. Items of ATI varied highly for Tests 19 through 22, and yet the null hypothesis of essential unidimensionality was rejected consistently, except for Test 21.

Two-Dimensional Studies

Results for the two-dimensional tests are reported in Table 4. Because items that tap a distinct second dimension, from the content perspective, were clearly known, these items were forced to be selected in ATI (that is, items from Subtest 2 in Table 2 were selected for ATI). In the two-dimensional case, expert opinion was used to select ATI items. The T and p values for the ACT reading and science tests—Tests 23 through 27—confirmed the two-dimensional nature of these test data. As expected, the power of T increased with sample size (see Table 2 for sample sizes).

Table 4
Results of H_o: $d_E = 1$ For Two-Dimensional Tests (Expert Opinion Was Used for Selecting AT1 Items)

Test	T	p	M
23	1.9	0.00	6
24	2.7	0.00	6
25	3.7	0.00	6
26	3.3	0.00	6
27	6.8	0.00	6
28	2.8	0.00	5
29	6.2	0.00	10
30	4.3	0.00	5
31	4.1	0.00	10
32	3.0	.04	5
33	3.4	0.00	8
34	2.0	.02	10

The T and p values of the two-dimensional ASVAB Arithmetic and General Science tests—Tests 28 through 31—confirmed the multidimensional nature of these test data. For Tests 28 and 29, there was a sharp increase in T as the degree of contamination, as measured by the number of item responses contaminated, increased from 5 to 10 (see Table 4). As with the other two-dimensional tests, the T and p values for the literature and history tests—Tests 32, 33, and 34—also confirmed the multidimensional nature of these data.

DIMTEST was applied again to a sample of test data selected from two-dimensional tests. This time factor analysis was used as the method of selection for ATI items. The purpose of this analysis was to evaluate whether the factor analysis method of selection of ATI items would

lead to p values similar to those obtained using expert opinion. For these tests, factor analysis could not always identify items that were purely unidimensional from a content perspective. Subtest ATI had a mixture of items tapping both dimensions, and DIMTEST was then able to correctly assess dimensionality only when there were 1,000 or more examinees for computing the statistic.

Discussion and Conclusions

None of the tests examined in the present study was strictly unidimensional—that is, none measured only one trait. Instead, the items in every test were influenced by several secondary traits in addition to the major trait intended to be measured. Using DIMTEST, some test data were assessed as fitting an essentially unidimensional model, but others were not. This was dependent on whether the secondary traits were major or minor.

The unidimensionality analysis of Tests 1, 11, and 13 showed that T was robust against relatively minor correlated traits influencing test items and was sensitive to major traits. Although both Test 12 and Test 14 were paragraph comprehension tests, they differed widely in the degree of approximation to essential unidimensionality. Test 12 had 3 passages each followed by 10 items dealing with humanities. Although these passages came from different sources, the model underlying the item responses approximated an essentially unidimensional model. This is an example where a few secondary traits (possibly highly correlated) each influenced a large group of items. In contrast, Test 14 had 5 passages each followed by 5 or 6 items. These passages, although they concerned science in general, came from widely different and conceptually difficult topics, and the model underlying the item responses did not approximate an essentially unidimensional model. This is an example in which many secondary traits each influence a small group of items, but the strength of the influence of these secondary traits is such that item responses cannot be well modeled by an essentially unidimensional model. These results are consistent with simulation results obtained by Nandakumar (1991) in that the number of items influenced by secondary traits and the strength of the secondary traits present determine the degree to which the assumption of essential unidimensionality is violated.

The results for two-dimensional tests demonstrated very good power for T. T has the capability to ignore minor secondary traits, which should be largely discounted, from the major dominant traits. This was evidenced in several cases. For example, there was inherent multidimensionality in Test 2 because it covered a range of time periods in history. However, the p value was above the nominal level of significance, suggesting the acceptance of unidimensionality. By contrast, with the additional contamination of only 5 literature items (Test 32) or 5 map items (Test 2), the p value fell below the nominal level, indicating essential multidimensionality of the data. This sensitivity of T to major dimensions illustrates its power.

These results have illustrated both the factor analysis and expert opinion approaches to ATI item selection. Table 3 presents results based on factor analysis, and Table 4 presents results from expert opinion. It is evident that factor analysis serves as an exploratory tool and expert opinion serves as a confirmatory tool in selecting items for ATI to assess essential unidimensionality.

The simulations and real data studies applying DIMTEST used test lengths of 25 or more items and 700 or more examinees. These findings may not replicate for test lengths of fewer than 25 items or with smaller sample sizes. De Champlain & Gessaroli (1991) demonstrated that DIMTEST loses power when both the test length and sample sizes are small (for example, a test of 25 items with 500 examinees). More research is needed to assess the effectiveness of DIMTEST for small samples.

The dimensionality of a given set of item responses is a continuum—it cannot be determined whether the responses of a group of examinees to a set of items is truly essentially

unidimensional or truly multidimensional; rather, the dimensionality can only be approximated. Although the exact number of dimensions in an IRT model is rigorously defined for a finite length test, the number of dominant dimensions—whether determined by Stout's essential dimensionality conceptualization or by some other conceptualization—is only rigorously definable for an infinitely long test. In other words, for a finite test (that is, for any real test data) a judgment must be made as to whether a particular IRT model is seen as having one, or more than one, dominant dimension, based on where on the continuum the degree of multidimensionality falls. One consequence of this is that the performance of trait estimation procedures such as LOGIST or BILOG needs to be addressed in the context of the assessment of the amount of lack of unidimensionality. In this regard, indices of lack of essential unidimensionality developed by Junker & Stout (1991) will be useful. These indices can be used to decide when it is safe to use unidimensional estimation procedures such as LOGIST and BILOG to arrive at accurate estimates of trait level.

In cases in which the approximation of an essentially unidimensional model to the data is in question, there are various alternatives. The test items can be split into essentially unidimensional subtests (e.g., Test 1 and Test 11). Another possible approach is to investigate the applicability of "testlets" to the data (Rosenbaum, 1988; Thissen, Steinberg, & Mooney, 1989). If the assumption of local independence is violated within the passages but maintained among the passages, the theory of testlets promises unidimensional scoring for such tests. For example, Tests 13 and 14 could fall into this category. Multidimensional modeling can be applied if either of the above procedures cannot be applied (Reckase, 1989).

References

Berger, M. P., & Knol, D. L. (1990, April). *On the assessment of dimensionality in multidimensional item response theory models.* Paper presented at the annual meeting of the American Educational Research Association, Boston.

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

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.

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

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

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

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

Hambleton, R. K., & Traub, R. E. (1973). Analysis of empirical data using two logistic latent trait models. *British Journal of Mathematical and Statistical Psychology, 24,* 273–281.

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

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

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

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

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

Humphreys, L. G. (1981). The primary mental ability. In M. P. Friedman, J. P. Das, & N. O'Connor (Eds.), *Intelligence and learning* (pp. 87–102). New York: Plenum Press.

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

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

Junker, B., & Stout, W. (1991, July). *Structural robustness of ability estimates in item response theory.* Paper presented at the 7th European Meeting of the Psychometric Society, Trier, Germany.

Linn, R. L., Hastings, N. C., Hu, G., & Ryan, K. E. (1987). *Armed Services Vocational Aptitude Battery: Differential item functioning on the high school form* (Tech. Rep. No. F41689-84-D- 0002). Dayton OH: USAF Human Resources Laboratory.

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

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

McDonald, R. P. (1962). A general approach to nonlinear factor analysis. *Psychometrika, 4,* 397–415.

McDonald, R. P., & Ahlawat, K. S. (1974). Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology, 27,* 82–89.

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

Nandakumar, R. (1991). Traditional dimensionality vs. essential dimensionality. *Journal of Educational Measurement, 28,* 1–19.

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

National Assessment of Educational Progress. (1988). *User guide: 1985–86 public-use data tapes.* Princeton NJ.

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

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

Reckase, M. (1989). *The interpretation and application of multidimensional item response theory models; and computerized testing in the instructional environment* (Technical Rep. No. N00014-85-C-0241). Iowa City IA: The American College Testing Program.

Reckase, M. D. (1990, April). *Unidimensional data from multidimensional tests and multidimensional data from unidimensional tests.* Paper presented at the annual meeting of the American Educational Research Association, Boston.

Rosenbaum, P. R. (1988). Item bundles. *Psychometrika, 53,* 349–359.

Roznowski, M. A., Tucker, L. R., & Humphreys, L. G. (1991). Three approaches to determining the dimensionality of binary data. *Applied Psychological Measurement, 15,* 109–128.

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

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensional assessment and ability estimation. *Psychometrika, 55,* 293–326.

Thissen, D., Steinberg, L., & Mooney, J. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement, 26,* 247–260.

Traub, R. E. (1983). A priori considerations in choosing an item response model. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 57–70). British Columbia: Educational Research Institute of British Columbia.

Wilson, D., Wood, R. L., & Gibbons, R. (1983). *TESTFACT: Test scoring and item factor analysis* [Computer program]. Chicago: Scientific Software.

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

Acknowledgments

The author thanks William F. Stout and Brian Junker for their helpful comments and many suggestions on this research, and Mark Reckase and Tim Miller for providing the ACT data. This research was partially supported by Office of Naval Research Grant ONR- N00014-91-J-1208.

Author's Address

Send requests for reprints or further information to Ratna Nandakumar, Department of Educational Studies, 213 Willard Hall, University of Delaware, Newark DE 19716, U.S.A.