Assessing Unidimensionality of Polytomous Data

Ratna Nandakumar and Feng Yu, University of Delaware
Hsin-Hung Li, National Changhua University of Education
William Stout, University of Illinois at Urbana-Champaign

This study investigated the performance of Poly-DIMTEST (PD) to assess unidimensionality of test data produced by polytomous items. Two types of polytomous data were considered: (1) tests in which all items had the same number of response categories, and (2) tests in which items had a mixed number of response categories. Test length, sample size, and the type of correlation matrix (used in factor analysis for selecting ATI subset items) were varied in Type I error analyses. For the power study, the correlation between θs and the item-θ loadings were also varied. The results showed that PD was able to confirm unidimensionality for unidimensional simulated test data, with the average observed level of significance slightly below the nominal level. PD was also highly effective in detecting lack of unidimensionality in various two-dimensional tests. As expected, power increased as the sample size and test length increased, and the correlation between the θs decreased. The results also demonstrated that use of Pearson correlations to select ATI items led to equally good or better performance than using polychoric correlations; therefore Pearson correlations are recommended for future use. *Index terms: dimensionality; factor analysis; item response theory, dimensionality; Poly-DIMTEST, polytomous item data.*

Item response theory (IRT) is a major measurement modeling paradigm for measurement data. It describes how to draw inferences about examinee trait levels (θ) from item responses. Three assumptions of commonly used IRT-based models are monotonicity, local independence, and unidimensionality. Monotonicity states that an examinee's probability of responding to an item answered correctly increases as the examinee's θ increases; local independence assumes that the responses to different items in a test are independent of each other for a given θ; and unidimensionality assumes that all items in a test measure a single θ. Violating the unidimensionality assumption could seriously bias item and θ parameter estimation (Ackerman, 1989; Kirisci & Hsu, 1995). Therefore, unidimensionality assessment should be a prerequisite before applying most commonly used IRT models.

Many methods have been developed for assessing the unidimensionality of dichotomously scored items. Some commonly used procedures are linear factor analysis (Hambleton & Traub, 1973; Hattie, 1985; Reckase, 1979), nonlinear factor analysis (Gessaroli & De Champlain, 1996), and DIMTEST (Nandakumar & Stout, 1993; Stout, 1987). Linear factor analysis can be conducted based on a phi (Pearson) correlation matrix or a tetrachoric correlation matrix. The use of phi correlation matrices has been found to overestimate the number of true underlying dimensions in a test (Hambleton & Rovinelli, 1986; McDonald & Ahlawat, 1974). Tetrachoric correlation matrices have an advantage: One common factor in a tetrachoric correlation matrix is a sufficient condition for the latent unidimensionality of a set of items (Lord & Novick, 1968, p. 382). However, this is not a necessary condition (Hambleton & Swaminathan, 1985, p. 22). Moreover, tetrachoric

Applied Psychological Measurement, Vol. 22 No. 2, June 1998, 99–115
©1998 Sage Publications, Inc.

from the SAGE Social Science Collections. All Rights Reserved.

(r_t) correlation matrices are not always positive definite (De Ayala & Hertzog, 1989; Hattie, 1984; Mislevy, 1986). Thus, traditional factor analysis methods do not provide satisfactory solutions for the dimensionality assessment of items.

Nonlinear factor analysis, using the NOHARM computer program, is another promising method for assessing unidimensionality (Etazadi-Amoli & McDonald, 1983). The incremental fit index developed by Gessaroli & De Champlain (1996) is based on the sum of the squares of the residual covariances (after fitting a specified nonlinear factor model), and is used as an index of dimensionality. Results indicate that this procedure is especially good for small sample sizes and test lengths (e.g., a 20-item test with 500 examinees).

Stout's DIMTEST is a nonparametric statistical procedure based on the concept of essential unidimensionality (Stout, 1987, 1990). The concept of essential unidimensionality attempts to model a dataset governed by one dominant dimension with the possible presence of several minor dimensions. Essential unidimensionality is a realistic concept because items frequently are multiply determined and it is not uncommon to find transient θs, each influencing a few items to a small degree. In assessing dimensionality, however, only the dominant dimensions are relevant. Studies based on simulated data and real data have shown that DIMTEST is a potentially powerful procedure for assessing the undimensionality of dichotomously scored item responses (Hattie, Krakowski, Rogers, & Swaminathan, 1996; Nandakumar, 1993, 1994; Nandakumar & Stout, 1993; Roussos, Stout, & Marden, 1993).

Recently, polytomously scored items have been used frequently to measure examinee θ. One attractive feature of a polytomous item is that it provides more information about examinee θ than a dichotomous item (Donoghue, 1994; Samejima, 1976; Thissen, 1976). Different kinds of IRT models have been developed to model polytomous data with different scoring formats (Drasgow, 1995). These include the graded response model (Samejima, 1969, 1976), the rating scale model (Andrich, 1978), the partial-credit model (Masters, 1982), and the generalized partial-credit model (Muraki, 1992). Most polytomous IRT models also require that items measure a unidimensional θ. One disadvantage of polytomous items, however, is that they may be time-consuming and expensive to score. Thus, large-scale testing instruments, such as the NAEP mathematics and reading tests, contain both dichotomous and polytomous items (Johnson & Carlson, 1994).

Most dimensionality assessment methods use item-pair correlations as a starting point in the dimensionality analysis. Therefore, the procedure used in the computation of item-pair correlations has a significant role in the dimensionality assessment. For dichotomous items, either phi or tetrachoric correlations can be computed. Similarly, for polytomous items either Pearson or polychoric correlations can be computed. A polychoric correlation (r_p) between two ordinal items is not simply a correlation between two sets of item scores, but rather an estimate of the correlation between their corresponding examinee-generated latent variables, which are assumed to be bivariate normal. In comparison to the Pearson correlation (r), r_p is a more consistent and better estimator of the true correlation between two ordinal variables (Jöreskog & Sörbom, 1988). Carlson (1993) applied factor analysis to r_ps to investigate the dimensionality of 1992 NAEP mathematics and reading data. His results showed that the type of items (dichotomous vs. polytomous) did not impose multidimensionality; that is, the item type did not impose structure on the data. An important drawback of r_p matrices is that they are not always positive definite.

Poly-DIMTEST (PD; Li & Stout, 1995) is another method that can assess unidimensionality of polytomous item data. It is an extended version of DIMTEST used to assess unidimensionality of data resulting either from polytomous items of the same number of response categories (e.g., all items having three response categories), or a combination of items having a different number of response categories (e.g., dichotomous, three response categories, and four response categories).

This study was designed to investigate the performance of PD to assess the unidimensionality of test data produced by administering polytomous item formats, as opposed to purely dichotomous item formats.

The Poly-DIMTEST Procedure

PD is a statistical procedure with an associated computer program that can be used to assess unidimensionality of ordered polytomous item data. It is assumed that a group of examinees takes an N-item test. Each examinee produces a vector of item responses. Each item response can be scored from 0 to m_i, where m_i is the maximum possible score for the ith item. In other words, the ith item is categorized into $m_i + 1$ ordered response categories $(0, 1, \ldots, m_i)$. The hypothesis for assessing unidimensionality can be stated as

$$H_0: d_E = 1 \text{ vs. } H_1: d_E > 1 , \tag{1}$$

where d_E denotes the essential dimensionality (the number of dominant dimensions) of the latent space underlying the given set of item responses. Implementation of PD to assess unidimensionality (similar to DIMTEST; Nandakumar & Stout, 1993), involves a number of steps.

1. N test items are split into three subtests: AT1, AT2, and PT. A relatively small set of M items is first selected for the subtest AT1 in such a way that the items appear to measure the same dominant θ. This can be achieved either through expert opinion or exploratory factor analysis (see Nandakumar & Stout, 1993, for details). In this study, factor analysis was used to select AT1 subtest items. Similar to the DIMTEST factor analysis approach, items with the highest loadings on the second factor were selected. After selecting AT1 items, a second set of M items is selected for the AT2 subtest so that the items in AT2 match the item difficulty distribution of AT1 items. PT, the partitioning subtest, contains the remaining items $(n = N - 2M)$ and is used to partition examinees into subgroups. When factor analysis is used for selecting AT1 items, a part of the examinee sample is set aside to avoid overfitting the data. The remaining examinees are then used to compute the DIMTEST statistic as described in step 3.
2. Examinees are assigned to K different subgroups, according to their scores on the PT subtest. Each subgroup k $(k = 1, \ldots, K)$ contains J_k examinees.
3. Two variance estimates $\hat{\sigma}_k^2$ and $\hat{\sigma}_{u,k}^2$ are calculated within each subgroup k, using AT1 item responses. Then, the difference between these two variance estimates is normalized within each subgroup k and added across all K subgroups to obtain the statistic T_L (for details see Li & Stout, 1995):

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

where the usual variance estimate $\hat{\sigma}_k^2$ is given by

$$\hat{\sigma}_k^2 = \sum_{j=1}^{J_k} \frac{\left(Y_j^{(k)} - \overline{Y}^{(k)} \right)^2}{J_k} , \tag{3}$$

the "unidimensional variance" estimate $\hat{\sigma}^2_{u,k}$ is given by

$$\hat{\sigma}^2_{u,k} = \frac{1}{M^2}\sum_{i=1}^{M}\left[\frac{\sum_{j=1}^{J_k} U^2_{ijk}}{J_k m_i^2} - \left(\frac{\sum_{j=1}^{J_k} U_{ijk}}{J_k m_i}\right)^2\right], \tag{4}$$

and the standard error S_k is given by

$$S^2_k = \frac{\hat{\mu}_{4,k} - \hat{\sigma}^4_k + \frac{\hat{\lambda}_{4,k}}{M^4} + 2\sqrt{\frac{(\hat{\mu}_{4,k} - \hat{\sigma}^4_k)\hat{\lambda}_{4,k}}{M^4}}}{J_k}. \tag{5}$$

In Equations 2 to 5, $0 \leq U_{ijk} \leq m_i$ denotes the ith ordered item response by the jth examinee in the kth subgroup, and

$$Y^{(k)}_j = \frac{\sum_{i=1}^{M}\left(\frac{u_{ijk}}{m_i}\right)}{M} \tag{6}$$

denotes the average proportion correct on the AT1 subtest obtained by the jth examinee in the kth subgroup (the score on each item i is rescaled by dividing by m_i, so that all items have the same score range of 0 to 1).

$$\bar{Y}^{(k)} = \frac{\sum_{j=1}^{J_k} Y^{(k)}_j}{J_k} \tag{7}$$

denotes the average proportion correct on the AT1 subtest for the examinees in the kth subgroup, and

$$\hat{\mu}_{4,k} = \sum_{j=1}^{J_k}\frac{\left(Y^{(k)}_j - \bar{Y}^{(k)}\right)^4}{J_k}, \tag{8}$$

and

$$\hat{\lambda}_{4,k} = \sum_{i=1}^{M}\frac{1}{m_i^4}\left(\hat{\alpha}_{i,4} - 4\hat{\alpha}_{i,3}\hat{\alpha}_{i,1} - \hat{\alpha}^2_{i,2} + 8\hat{\alpha}^2_{i,1}\hat{\alpha}_{i,2} - 4\hat{\alpha}^4_{i,1}\right), \tag{9}$$

where

$$\hat{\alpha}_{i,1} = \frac{\sum_{j=1}^{J_k} U_{ijk}}{J_k}, \quad \hat{\alpha}_{i,2} = \frac{\sum_{j=1}^{J_k} U^2_{ijk}}{J_k}, \quad \hat{\alpha}_{i,3} = \frac{\sum_{j=1}^{J_k} U^3_{ijk}}{J_k}, \quad \hat{\alpha}_{i,4} = \frac{\sum_{j=1}^{J_k} U^4_{ijk}}{J_k}. \tag{10}$$

The statistic T_B is similarly computed using items in subtest AT2. Finally, Stout's statistic T to test for unidimensionality is given by

$$T = \frac{T_L - \sqrt{\frac{V_1}{V_2}}T_B}{\sqrt{1 + \frac{V_1}{V_2}}}. \tag{11}$$

The ratio V_1/V_2 in Equation 11 is an empirical weight used to adjust for the influence of possibly different response category items in the AT1 and the AT2 subtests. This weight was discovered by Li & Stout (1995) through simulation studies. Although subtests AT1 and AT2 contained the same number of items, the variance of subtest scores $\left[Y_j^{(k)}\right]$ may have differed depending on the number of response categories each item had in these subsets. This adjustment allows the statistic T to approximately follow the standard normal distribution when the unidimensionality assumption holds. V_1 is the variance of the set of integers from 0 to the possible total score of the AT1 subtest $\left(\sum_{i=1}^{M} m_i\right)$ before rescaling and, similarly, V_2 is the variance of AT2. When AT1 and AT2 have the same possible total scores, the ratio V_1/V_2 equals 1.0. For example, consider a case in which AT1 contained one dichotomous item, one three-category item, and three five-category items with a possible total score of 15, and AT2 contained five five-category items with a possible total score of 20. Then V_1 is the sample variance of (0, 1, 2, ..., 15), which is 20; V_2 is the sample variance of (0,1, 2, ..., 20), which is 35; and the ratio V_1/V_2 is .571. The decision rule is to reject H_0 when $T \geq Z_\alpha$, where Z_α is the upper 100(1- α) percentile of the standard normal distribution, and α is the desired level of significance.

Method

A monte carlo simulation study was conducted to investigate the performance of PD in assessing the unidimensionality of polytomous test data. This study had two objectives: (1) to investigate the Type I error performance of PD in true unidimensional situations ($d = 1$), with the nominal level of significance specified at $\alpha = .05$; and (2) to investigate the power of PD for various $d = 2$ settings.

The Type I Error Study

To make the simulated data as realistic as possible, estimated item parameters (a_i, b_i, c_i and d_{iv}) were obtained from the National Assessment of Educational Progress test data (Johnson & Carlson, 1994). For the unidimensional simulation study, all items were influenced by only one θ, which was assumed to have a standard normal distribution. Each examinee's θ was randomly selected from this distribution.

Two different test lengths were used ($n = 20$ and 40). For each test length, there were two different structure patterns: (1) all items with the same number of response categories (dichotomous, three-, four-, and five-response category items), and (2) a mixture of dichotomous items and one type of polytomous items (dichotomous and three-category items, dichotomous and four-category items, and dichotomous and five-category items). In the mixture case, the proportion of dichotomous items was approximately 75% (14 for a 20-item test and 30 for a 40-item test), and the remaining items were polytomous items with the same number of response categories. These proportions of items were modeled to match the item structure of the NAEP tests. Two sample sizes were studied: 1,000 and 1,500.

Dichotomous item responses were generated from the three-parameter logistic model,

$$P_i(\theta_j) = c_i + \frac{1 - c_i}{1 + \exp\left[-1.7a_i(\theta_j - b_i)\right]}, \tag{12}$$

where

$P_i(\theta_j)$ is the probability of correctly responding to dichotomous item i by examinee j,
a_i is the discrimination parameter of dichotomous item i,
b_i is the difficulty parameter of dichotomous item i, and
c_i is the guessing parameter of dichotomous item i.

Polytomous item responses were generated using the generalized partial-credit model (Muraki, 1992),

$$P_{ig}(\theta_j) = \frac{\exp\left[\sum_{v=0}^{g} 1.7a_i(\theta_j - b_i + d_{iv})\right]}{\sum_{g=0}^{m_i} \exp\left[\sum_{v=0}^{g} 1.7a_i(\theta_j - b_i + d_{iv})\right]}, \quad g = 0, 1, \ldots, m_i, \tag{13}$$

where

$P_{ig}(\theta_j)$ is the probability of obtaining a score g by examinee j with θ_j to polytomous item i,
a_i is the slope parameter of polytomous item i,
b_i is the location parameter of polytomous item i,
m_i is the maximum score of polytomous item i,
d_{iv}s are a set of threshold parameters of polytomous item i, with associated constraints $d_{i0} = 0$, and

$$\sum_{v=1}^{m_i} d_{iv} = 0 . \tag{14}$$

For each simulated examinee, each item response probability was computed using either Equation 12 or Equation 13, depending on whether the item was dichotomously or polytomously scored. For a dichotomously scored item, when the computed probability was greater than a uniform random variable generated in the interval (0,1), the item was considered answered correctly and a score of 1 was assigned; otherwise, a score of 0 was assigned. For a polytomous item with $m_i + 1$ response categories, probabilities $P_g(g = 0, 1, \ldots, m_i)$ for each category were computed using Equation 13. Then a set of cumulative probabilities was obtained as follows:

$$P'_g = \frac{\sum_{i=0}^{g} P_i}{\sum_{g=0}^{m_i} P_g}, \quad g = 0, \ldots, m_i . \tag{15}$$

When a randomly generated uniform variable from the interval (0,1) was equal to or larger than P'_g, but smaller than P'_{g+1}, a score of g was assigned.

For a given combination of test length, test type, and sample size, two sets of examinee responses were generated, with 500 examinees in the first set and the remaining portion in the second set. Because factor analysis was used to select items for the dimensionally homogeneous AT1 subtest, the first set of responses was used for the factor analysis. The second set of responses was used to compute the PD T statistic. In performing the factor analysis, the type of correlation matrix used could impact the factor loadings. For example, with dichotomous items, because of the nonlinearity

of IRT models, the use of r (as contrasted with r_t) for factor analysis often leads to a pronounced, artifactual "difficulty factor" when the underlying model is unidimensional (Hambleton & Swaminathan, 1985, chap. 2; McDonald & Ahlawat, 1974). Use of r_t can partially alleviate this problem. Similarly, for polytomous item data, r_p is appropriate for use in factor analysis. However, when there are many response categories (such as five) it may make little difference in the value of the computed correlation whether r or r_p is used. Therefore, both types of correlations were used for comparison purposes. In this study, r_p was computed using the PRELIS program (Jöreskog & Sörbom, 1988) and factor loadings were computed using the principal factor analysis program that is part of the PD procedure.

Thus, the factors varied were test length (20, 40), sample size (1,000, 1,500), correlation type (r, r_p), and test type (4 types of same-category items and 3 types of mixture items). All factors were completely crossed, resulting in a total of 56 different combinations. For each of these combinations, PD was applied to assess unidimensionality. PD was replicated 100 times for each of the 56 cells, and the number of times H_0 was rejected was noted.

The Power Study

The design of the power study used the same 56 combinations of conditions as the Type I error study, plus two additional factors: correlation between θs and item loading patterns. For each type of test, two possible item loading patterns were considered: in Type 1, half the items loaded on θ_1 and half the items loaded on θ_2; in Type 2, the first third of the items loaded on θ_1, the second third loaded on θ_2, and the last third loaded equally on both θ_1 and θ_2.

θ_1 and θ_2 were generated from a bivariate normal distribution with both means of 0 and variances of 1. The correlation between θ_1 and θ_2 was either $\rho = .7$ or .3; these values were considered to correspond approximately to the upper- and lower-bound values in real tests. Estimated item parameters from 1992 NAEP unidimensional test data were used in simulating responses (Johnson & Carlson, 1994). For items that loaded on both θ_1 and θ_2, two sets of estimated unidimensional item parameters were independently selected from the 1992 NAEP unidimensional data (Johnson & Carlson, 1994).

Responses for the two-dimensional dichotomous items were generated using the three-parameter logistic model with compensatory θs (Reckase & McKinley, 1983),

$$P_i(\theta_1, \theta_2) = c_i + \frac{1 - c_i}{1 + \exp\{-1.7[a_{1i}(\theta_1 - b_{1i}) + a_{2i}(\theta_2 - b_{2i})]\}}, \tag{16}$$

where a_{1i} and a_{2i} denote discriminations of item i on θ_1 and θ_2, respectively. Similarly, b_{1i} and b_{2i} denote difficulties on θ_1 and θ_2, respectively.

Item responses for the polytomous items were generated using a two-dimensional extension of the unidimensional generalized partial-credit model,

$$P_{ig}(\theta_1, \theta_2) = \frac{\exp\left\{\sum_{v=0}^{g} 1.7\,[a_{1i}(\theta_1 - b_{1i} + d_{1iv}) + a_{2i}(\theta_2 - b_{2i} + d_{2iv})]\right\}}{\sum_{g=0}^{m_i} \exp\left\{\sum_{v=0}^{g} 1.7\,[a_{1i}(\theta_1 - b_{1i} + d_{1iv}) + a_{2i}(\theta_2 - b_{2i} + d_{2iv})]\right\}}, \tag{17}$$

$$g = 0, 1, \ldots, m_i,$$

where

a_{1i} is the slope corresponding to θ_1,

b_{1i} is the location corresponding to θ_1,

d_{1iv} represents the threshold parameters of item i corresponding to θ_1,
a_{2i} is the slope corresponding to θ_2,
b_{2i} is the location corresponding to θ_2, and
d_{2iv} represents the threshold parameters of item i corresponding to θ_2.

For a given combination of test length, test type, sample size, level of correlation, type of item loading, and type of correlation matrix, two sets of examinee samples were generated as in the unidimensional study. The first sample, consisting of 500 examinees, was used to select AT1 items through factor analysis, and the second sample (the remaining examinees), was used to compute the statistic T. All factors were completely crossed, producing a total of 224 different combinations. For each of these combinations, PD was replicated 100 times and the number of times H_0 was rejected was noted.

Results

Type I Error Study

Table 1 shows the performance of PD for tests in which all items had the same number of response categories, and results for the mixture case. In Table 1, each panel (a four-row set) shows the results for tests containing the same type of items. For example, the top panel shows results for tests with dichotomously scored items (n_2), the second panel shows the results for tests with three-category items (n_3), and so forth. The last two columns show the observed Type I error rates over 100 replications based on r and r_p matrices, respectively.

The observed Type I error rates were all close to .05, the nominal level. The average Type I error rate was .038, leading to the conclusion that PD showed good performance with regard to the nominal level of significance. The average Type I error rate (.047) for 40-item tests was somewhat higher than the corresponding error rate (.030) for 20-item tests. Comparison of r and r_p showed that, on the average, they both yielded approximately the same Type I error rate, .039 for r and .037 for r_p. The highest observed error rate for the 56 cases was .07. These results also demonstrated that the Type I error performance of PD in tests in which all items had the same number of response categories (mean = .038) was similar to the mixture case (mean = .039).

Power Study

Results for the two-dimensional power study are reported in Tables 2–5. Tables 2 and 3 show results for item loading Type 1 (items loaded either on θ_1 or θ_2); Tables 4 and 5 show results for item loading Type 2 (some items loaded simultaneously on both θ_1 and θ_2). The results in Tables 2–5 indicate that power increased as the sample size increased, as the number of items increased, and as the correlation between the θs decreased. These results show that PD had good power in the two-dimensional cases with either pure polytomously scored items or a mixture of dichotomously scored and polytomously scored items.

For tests in which all items had the same number of response categories (Table 2), power increased as the number of item response categories increased and the number of items increased. This was true for both r and r_p correlation matrices. The average rejection rates for r were 98 (ρ = .3, 40 items), 89 (ρ = .3, 20 items), 81 (ρ = .7, 40 items), and 64 (ρ = .7, 20 items). The corresponding averages for r_p were very similar: 93, 89, 81, and 64, respectively. The results in Table 3 show that power decreased slightly for tests containing mixed category items. For example, the average rejection rates for r were 89 (ρ = .3, 40 items), 80 (ρ = .3, 20 items), 61 (ρ = .7, 40 items), and 46 (ρ = .7, 20 items). The corresponding averages for r_p were 89, 82, 63, and 42. These results also show that r led to equal, or sometimes better, performance than r_p. However, this difference was small and within the range of chance variation.

Table 1
Type I Error Study: Rejection Frequencies per 100 Trials for Tests With Same Number of Categories per Item and Tests With a Mixture of Dichotomous and Polytomous Items (n_i = the Number of Items With i Response Categories)

Sample Size	Test Length	n_2	n_3	n_4	n_5	Rejection Frequency r	Rejection Frequency r_p
Tests With the Same Number of Categories							
1000	20	20				4	5
1500	20	20				1	2
1000	40	40				5	6
1500	40	40				4	3
1000	20		20			3	4
1500	20		20			3	3
1000	40		40			5	6
1500	40		40			4	3
1000	20			20		3	4
1500	20			20		2	2
1000	40			40		6	5
1500	40			40		5	3
1000	20				20	5	3
1500	20				20	1	2
1000	40				40	6	5
1500	40				40	4	3
Tests With Dichotomous and Polytomous Items							
1000	20	14	6			5	4
1500	20	14	6			3	1
1000	40	30	10			4	6
1500	40	30	10			3	4
1000	20	14		6		3	3
1500	20	14		6		4	2
1000	40	30		10		4	6
1500	40	30		10		5	3
1000	20	14			6	3	3
1500	20	14			6	3	1
1000	40	30			10	5	7
1500	40	30			10	6	5

Tables 4 and 5 show similar results for tests containing items that loaded on both θs. In Table 4, the average rejection rates for r were 98 (ρ = .3, 40 items), 80 (ρ = .3, 20 items), 67 (ρ = .7, 40 items), and 46 (ρ = .7, 20 items). The corresponding results for r_p were 85, 73, 63, and 37. Similarly, the average rejection rates for r in Table 5 were 88 (ρ = .3, 40 items), 67 (ρ = .3, 20 items), 44 (ρ = .7, 40 items), and 28 (ρ = .7, 20 items); the corresponding results for r_p were 78, 51, 48, and 18. The rejection rates in Tables 4 and 5 indicate that when tests contained mixed-θ items, the power decreased. Moreover, mixed categories coupled with mixed θs further reduced power. In this case, r produced somewhat greater power.

For tests with all dichotomously scored items (n_2), results were compared for three types of correlations—r, r_p, r_t. Power was highest when AT1 items were selected using r_t, and lowest using r_p. Theoretically, in the dichotomous case, r_t is the same as r_p. It provides the true underlying relationship between any two dichotomously scored items, provided there is an underlying normal distribution of θ. However, estimated correlations for dichotomous items may not yield the

Table 2
Power Study: Rejection Frequencies per 100 Trials for Tests With the Same Number of Categories, and Item Loading Type 1 (n_i = Number of Items With i Response Categories; n_m = Number of Items Loading on θ_1; n_v = Number of Items Loading on θ_2)

Sample		Test	n_2		n_3		n_4		n_5		Rejection Frequency		
Size	$\rho(\theta_1, \theta_2)$	Length	n_m	n_v	n_m	n_v	n_m	n_v	n_m	n_v	r_t	r	r_p
1000	.3	20	10	10							91	88	83
1500	.3	20	10	10							94	91	84
1000	.7	20	10	10							50	40	29
1500	.7	20	10	10							65	57	34
1000	.3	40	20	20							100	94	70
1500	.3	40	20	20							100	96	75
1000	.7	40	20	20							77	73	60
1500	.7	40	20	20							90	84	64
1000	.3	20			10	10						85	80
1500	.3	20			10	10						88	87
1000	.7	20			10	10						54	42
1500	.7	20			10	10						69	58
1000	.3	40			20	20						97	100
1500	.3	40			20	20						100	100
1000	.7	40			20	20						88	85
1500	.7	40			20	20						86	92
1000	.3	20					10	10				83	90
1500	.3	20					10	10				93	96
1000	.7	20					10	10				63	60
1500	.7	20					10	10				79	76
1000	.3	40					20	20				93	96
1500	.3	40					20	20				99	100
1000	.7	40					20	20				75	87
1500	.7	40					20	20				83	92
1000	.3	20							10	10		87	93
1500	.3	20							10	10		95	96
1000	.7	20							10	10		65	63
1500	.7	20							10	10		83	82
1000	.3	40							20	20		99	100
1500	.3	40							20	20		100	100
1000	.7	40							20	20		69	64
1500	.7	40							20	20		85	88

same results because of the different procedures used for estimating r_p and r_t. For example, r_t is estimated from the item's two-by-two contingency table (Lord & Novick, p. 345, 1968), whereas r_p is estimated using maximum likelihood with a complex iterative algorithm that requires large samples for accurate estimation (Jöreskog, 1994). Therefore, it is not surprising to observe better performance based on r_t, especially when the sample size is not large.

In the current study, because there were only 500 examinees with as many as 40 items, estimates of r_p contained more error and, thus, led to the selection of AT1 items that were not as dimensionally homogeneous as in the case of r_t. For example, in panel 1 of Table 2 with $\rho = .3$ and using r_p, the power was higher for 20-item tests (83 and 84 rejections) than for 40-item tests (70 and 75 rejections). This was not the case using r_t, where the corresponding rejections were 91 and 94 for 20 items, and 100 and 100 for 40 items.

Table 3
Power Study: Rejection Frequencies per 100 Trials for Tests With a Mixture of Dichotomous and Polytomous Items and Item Loading Type 1 (n_i = Number of Items With i Response Categories; n_m = Number of Items Loading on θ_1; n_v = Number of Items Loading on θ_2)

Sample		Test	n_2		n_3		n_4		n_5		Rejection Frequency	
Size	$\rho(\theta_1, \theta_2)$	Length	n_m	n_v	n_m	n_v	n_m	n_v	n_m	n_v	r	r_p
1000	.3	20	7	7	3	3					80	75
1500	.3	20	7	7	3	3					86	83
1000	.7	20	7	7	3	3					46	34
1500	.7	20	7	7	3	3					57	45
1000	.3	40	15	15	5	5					90	80
1500	.3	40	15	15	5	5					93	86
1000	.7	40	15	15	5	5					55	51
1500	.7	40	15	15	5	5					65	58
1000	.3	20	7	7			3	3			77	71
1500	.3	20	7	7			3	3			83	84
1000	.7	20	7	7			3	3			41	37
1500	.7	20	7	7			3	3			52	47
1000	.3	40	15	15			5	5			84	92
1500	.3	40	15	15			5	5			92	95
1000	.7	40	15	15			5	5			57	62
1500	.7	40	15	15			5	5			69	73
1000	.3	20	7	7					3	3	74	75
1500	.3	20	7	7					3	3	83	83
1000	.7	20	7	7					3	3	32	38
1500	.7	20	7	7					3	3	43	47
1000	.3	40	15	15					5	5	82	89
1500	.3	40	15	15					5	5	89	92
1000	.7	40	15	15					5	5	52	55
1500	.7	40	15	15					5	5	65	74

Moreover, the precision of r_p estimates depended on the number of categories. The higher the number of categories, the more the number of thresholds and the better the estimation of the underlying bivariate normal correlation. Two categories were not sufficient to obtain good estimation. For example, contrast the performance of r_p in panel 1 versus panel 4 in Table 2, where the performance for the five categories per item case was distinctly better than for the two categories per item case.

Analysis of Variance

Because many factors were manipulated in this study, an analysis of variance (ANOVA) was conducted for the results in each table in order to examine whether there were significant main effects or interactions. The dependent variable was the rejection frequency and independent variables were the sample size, test length, item type, and the correlation type for the Type I error study. The power study had an additional independent variable—the correlation between θs. The ANOVA results for the data in Tables 1, 3, and 5 are presented in Table 6. It can be seen that for all three datasets, the main effects of test length and sample size were significant; longer tests and larger sample sizes led to higher Type I error rates (close to the nominal level). The main effects of item type and correlation type were not significant for the Type I error study nor were any of the interactions. For the power study, on the other hand, the main effects of test length, sample size, item type, and correlation between θs were significant. That is, larger sample sizes, longer tests, more response categories, and low correlation between θs led to higher power. There was no systematic pattern among significant interactions across tables.

Table 4

Power Study: Rejection Frequencies per 100 Trials for Tests With Items of the Same Number of Categories and Item Loading Type 2 (n_i = Number of Items With i Response Categories; n_m = Number of Items Loading on θ_1; n_v = Number of Items Loading on θ_2; n_x = Number of Items Loading on Both θ_1 and θ_2)

Sample		Test	n_2			n_3			n_4			n_5			Rejection Frequency	
Size	$\rho(\theta_1, \theta_2)$	Length	n_m	n_v	n_x	n_m	n_v	n_x	n_m	n_v	n_x	n_m	n_v	n_x	r	r_p
1500	.3	40	13	13	14										96	77
1000	.3	40	13	13	14										90	68
1500	.7	40	13	13	14										63	65
1000	.7	40	13	13	14										44	44
1500	.3	20	7	7	6										67	54
1000	.3	20	7	7	6										47	46
1500	.7	20	7	7	6										33	28
1000	.7	20	7	7	6										25	22
1500	.3	40				13	13	14							99	76
1000	.3	40				13	13	14							97	62
1500	.7	40				13	13	14							80	41
1000	.7	40				13	13	14							52	33
1500	.3	20				7	7	6							90	74
1000	.3	20				7	7	6							79	67
1500	.7	20				7	7	6							43	33
1000	.7	20				7	7	6							31	22
1500	.3	40							13	13	14				100	99
1000	.3	40							13	13	14				100	95
1500	.7	40							13	13	14				83	80
1000	.7	40							13	13	14				75	78
1500	.3	20							7	7	6				97	91
1000	.3	20							7	7	6				88	90
1500	.7	20							7	7	6				68	72
1000	.7	20							7	7	6				49	49
1500	.3	40										13	13	14	100	100
1000	.3	40										13	13	14	100	100
1500	.7	40										13	13	14	81	89
1000	.7	40										13	13	14	56	68
1500	.3	20										7	7	6	93	88
1000	.3	20										7	7	6	78	74
1500	.7	20										7	7	6	38	39
1000	.7	20										7	7	6	33	24

Table 5
Power Study: Rejection Frequencies per 100 Trials for Tests With a Mixture of Dichotomous and Polytomous Items and Item Loading Type 2 (n_i = Number of Items With i Response Categories; n_m = Number of Items Loading on θ_1; n_v = Number of Items Loading on θ_2; n_x = Number of Items Loading on Both θ_1 and θ_2)

Sample		Test	n_2			n_3			n_4			n_5			Rejection Frequency	
Size	$\rho(\theta_1, \theta_2)$	Length	n_m	n_v	n_x	n_m	n_v	n_x	n_m	n_v	n_x	n_m	n_v	n_x	r	r_p
1500	.3	40	10	10	10	4	4	2							94	85
1000	.3	40	10	10	10	4	4	2							88	70
1500	.7	40	10	10	10	4	4	2							53	56
1000	.7	40	10	10	10	4	4	2							36	41
1500	.3	20	5	5	4	2	2	2							78	48
1000	.3	20	5	5	4	2	2	2							66	40
1500	.7	20	5	5	4	2	2	2							30	19
1000	.7	20	5	5	4	2	2	2							26	17
1500	.3	40	10	10	10				4	4	2				88	82
1000	.3	40	10	10	10				4	4	2				86	78
1500	.7	40	10	10	10				4	4	2				53	59
1000	.7	40	10	10	10				4	4	2				33	45
1500	.3	20	5	5	4				2	2	2				76	68
1000	.3	20	5	5	4				2	2	2				55	55
1500	.7	20	5	5	4				2	2	2				32	23
1000	.7	20	5	5	4				2	2	2				21	18
1500	.3	40	10	10	10							4	4	2	89	76
1000	.3	40	10	10	10							4	4	2	81	74
1500	.7	40	10	10	10							4	4	2	55	53
1000	.7	40	10	10	10							4	4	2	32	35
1500	.3	20	5	5	4							2	2	2	72	53
1000	.3	20	5	5	4							2	2	2	52	38
1500	.7	20	5	5	4							2	2	2	31	17
1000	.7	20	5	5	4							2	2	2	25	13

Table 6
ANOVA Summaries for a Sample of Simulation Results

Effect	DF	SS	MS	*F*	*P*
Results for Table 1					
Sample size (S)	1	28.13	28.13	52.2	0.00
Test length (T)	1	21.13	21.13	39.23	0.00
Item type (I)	3	.25	.08	.15	.92
Correlation (C)	—	—	—	—	—
Correlation type (CT)	1	0.13	0.13	.23	.64
S × T	1	0.00	0.00	0.00	1.00
S × I	3	2.13	.71	1.32	.31
S × C	—	—	—	—	—
T × I	3	.63	.21	.39	.76
T × C	—	—	—	—	—
I × C	—	—	—	—	—
CT × S	1	.50	.50	.93	.35
CT × T	1	2.00	2.00	3.71	.08
CT × I	3	2.13	.71	1.32	.31
CT × C					
Results for Table 3					
Sample Size (S)	1	963.02	963.02	142.82	0.00
Test Length (T)	1	2,227.70	2,227.70	330.37	0.00
Item Type (I)	2	62.38	31.19	4.63	.02
Correlation (C)	1	12,128	12,129	1,798.70	0.00
Correlation Type (CT)	1	9.19	9.19	1.36	.25
S × T	1	2.52	2.52	.10	.76
S × I	2	9.54	4.77	.71	.50
S × C	1	63.02	63.02	9.35	.01
T × I	2	130.88	65.44	9.70	0.00
T × C	1	238.50	238.50	35.37	0.00
I × C	2	19.29	9.65	1.43	.26
CT × S	1	.02	.02	0.00	.96
CT × T	1	50.02	50.02	7.42	.01
CT × I	2	286.13	143.06	21.22	0.00
CT × C	1	.52	.52	.08	.78
Results for Table 5					
Sample Size (S)	1	1,463.00	1,463.00	55.20	0.00
Test Length (T)	1	6,745.00	6,745.00	254.51	0.00
Item Type (I)	2	187.54	93.77	3.54	.04
Correlation (C)	1	15,732	15,732	593.63	0.00
Correlation Type (CT)	1	744.19	744.19	28.08	0.00
S × T	1	11.02	11.02	.42	.52
S × I	2	9.29	4.65	.18	.84
S × C	1	3.52	3.52	.13	.72
T × I	2	18.29	9.15	.35	.71
T × C	1	2.52	2.52	.10	.76
I × C	2	28.29	14.14	.53	.59
CT × S	1	25.52	25.52	.96	.34
CT × T	1	305.02	305.02	11.51	0.00
CT × I	2	216.13	108.06	4.08	.03
CT × C	1	336.02	336.02	12.68	0.00

Discussion

Results indicated that PD was able to confirm unidimensionality for unidimensional simulated test data with the average observed level of significance slightly below the nominal level. PD was also able to effectively detect lack of unidimensionality in various two-dimensional tests. As expected, power increased as the sample size increased, as test size increased, and as the correlation between θ_1 and θ_2 decreased. In particular, the average number of rejections using Pearson correlations for selection of AT1 items was 93 for $\rho = .3$ and $N = 40$, 79 for $\rho = .3$ and $N = 20$, 63 for $\rho = .7$ and $N = 40$, and 46 for $\rho = .7$ and $N = 20$. The corresponding averages using polychoric correlations were 86, 74, 64, and 40. These results demonstrated that the use of Pearson correlations to select AT1 items led to almost as good, or better, power than polychoric correlations.

The use of factor analysis to select items for AT1 was a major focus of this study. Findings showed that, as expected, the power of PD depended heavily on the selection of AT1 items. For example, in two-dimensional tests, when the AT1 items were dimensionally homogeneous, the null hypothesis of unidimensionality was more likely to be rejected than when AT1 was contaminated by multidimensional items. Selection of AT1 items through factor analyses depended, in turn, on the type of correlation matrix used.

This study showed that the use of Pearson correlations for factor analyses provided equally good, or sometimes better, results than using polychorics. Because polychoric correlation estimates the underlying linear relationship between two latent variables using thresholds (categories), it could be that larger numbers of thresholds result in better estimates. As a result, estimation of the polychoric correlation would be more accurate for four-category items than for dichotomously scored items; this result is shown in Tables 2 and 4. In addition, the estimation of polychoric correlation, as implemented in the PRELIS program, uses a maximum likelihood method and requires large samples for accurate estimation. When sample sizes are not large, polychoric correlations could lead to inaccurate results. Because selection of AT1 subtest items using Pearson correlations led to Type I error rates similar to those obtained with polychoric correlations, and the power with Pearson almost equaled and some times exceeded that of the polychoric, Pearson correlations should be used in factor analytic selection of AT1 items because they are easier and quicker to compute than polychoric correlations.

The use of polytomous items is becoming increasingly common as new types of assessments, such as performance and authentic assessments, become more widely used. When unidimensional IRT models are used for scoring large-scale performance assessments, the key assumption of unidimensionality must be evaluated. That is, it is important to verify that the test is strictly unidimensional or at least essentially unidimensional. Therefore, techniques to accurately assess unidimensionality are very important.

PD appears to be a promising method for assessing unidimensionality. However, this study, although fairly comprehensive in terms of factors involved and the number of replications performed, was limited in scope. In realistic applications, many assumptions may not strictly hold and many more factors will influence test scores. Future simulation studies should include additional factors or larger variation within factors, such as including small sample sizes (e.g., 300 to 500 examinees), more varied combinations of numbers of item response categories in the same test, and skewed distributions of θs. Poly-DIMTEST also needs to be validated on real datasets before it can be recommended for large scale use in dimensionality assessment.

References

Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement, 13*, 113–127.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561–573.

Carlson, J. E. (1993, April). *Dimensionality of NAEP instruments that incorporate polytomously scored items.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta GA.

De Ayala, R. J., & Hertzog, M. A. (1989, March). *A comparison of methods for assessing dimensionality for use in item response theory.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Donoghue, J. R. (1994). An empirical examination of the IRT information of polytomous scored reading items under the generalized partial-credit model. *Journal of Educational Measurement, 4*, 295–311.

Drasgow, F. (Ed.). (1995). Polytomous item response theory [Special Issue]. *Applied Psychological Measurement, 19 (1).*

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

Gessaroli, M. E., & De Champlain, A. (1996). Using an approximate chi-square statistic to test the number of dimensions underlying the responses to a set of items. *Journal of Educational Measurement, 2*, 157–179.

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.

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.

Hattie, J., Krakowski, K., Rogers, J., & Swaminathan, H. (1996). An assessment of Stout's index of essential dimensionality. *Applied Psychological Measurement, 20*, 1–14.

Johnson, E. G., & Carlson, J. E. (1994). *The NAEP 1992 technical report.* National Center for Educational Statistics, Report #23-TR20.

Jöreskog, K. G. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika, 59*, 381–389.

Jöreskog, K. G., & Sörbom, D. (1988). *PRELIS: A preprocessor for LISREL* (2nd ed.). Chicago: Scientific Software.

Kirisci, L., & Hsu, T. (1995, April). *The robustness of BILOG to violations of the assumption of unidimensionality of test items and normality of ability distribution.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Li, H.-H., & Stout, W. F. (1995, April). *Assessment of unidimensionality for mixed polytomous and dichotomous item data: Refinements of Poly-DIMTEST.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores.* Reading MA: Addison-Wesley.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174.

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

Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics, 11*, 3–31.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Psychometrika, 16*, 159–176.

Nandakumar, R. (1993). Assessing essential unidimensionality of real data. *Applied Psychological Measurement, 17*, 29–38.

Nandakumar, R. (1994). Assessing latent trait unidimensionality of a set of items—comparison of different approaches. *Journal of Educational Measurement, 31*, 1–18.

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

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., & McKinley, R. L. (1983, April). *The definition of difficulty and discrimination for multidimensional item response theory models.* Paper presented at the annual meeting of

the American Educational Research Association, Montreal.

Roussos, L. A., Stout, W. F., & Marden, J. I. (1993, April). *Dimensional and structural analysis of standardized tests using DIMTEST with hierarchical cluster analysis.* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta GA.

Samejima, F. (1969). A general model for free-response data. *Psychometric Monograph*, No. 18.

Samejima, F. (1976). Graded response model of the latent trait theory and tailored testing. In C. K. Clark (Ed.), *Proceedings of the First Conference on Computerized Adaptive Testing* (pp. 5–17). Washington DC: U.S. Government Printing Office.

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 modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55*, 293–325.

Thissen, D. (1976). Information in wrong responses to Raven progressive matrices. *Journal of Educational Measurement, 13*, 201–214.

Acknowledgments

The authors thank the editor and two anonymous reviewers for helpful comments and suggestions, which led to numerous improvements.

Author's Address

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