The performance of the following four methodologies for assessing unidimensionality was examined: (1) DIMTEST; (2) the approach of P. W. Holland and P. R. Rosenbaum; (3) linear factor analysis; and (4) non-linear factor analysis. Each method is examined and compared with other methods using simulated data sets and real data sets. Seven data sets, all with 2,000 examinees, were generated with 3 unidimensional and 4 2-dimensional data sets. Two levels of correlation between abilities were considered: \( p=0.3 \) and \( p=0.7 \). Eight real data sets were used; four were expected to be unidimensional, and the other four were expected to be two-dimensional. Findings suggest that, while the linear factor analysis often overestimated the number of underlying dimensions, the other three methods correctly confirmed unidimensionality, but differed in their ability to detect the lack of unidimensionality. DIMTEST showed excellent power in detecting the lack of unidimensionality. Holland and Rosenbaum's approach and non-linear factor analysis approaches showed good power, provided the correlation between abilities was low. Four tables present study data, and there is a 46-item list of references.

(Author/SLD)
Assessing Dimensionality of a Set of Items — Comparison of Different Approaches

Ratna Nandakumar
Department of Educational Studies
University of Delaware

August 10, 1992


1 The author would like to convey special thanks to Brian Junker and William Stout for their time and insightful suggestions on this research.
### Assessing Dimensionality of a Set of Items - Comparison of Different Approaches

**Author(s):** Ratna Nandakumar

**Funding Numbers:**
- N00014-90-J-1940

**Performing Organization:**
- University of Illinois
  - Department of Statistics
  - 725 South Wright Street
  - Champaign, IL 61820

**Sponsoring/Monitoring Agency:**
- Cognitive Sciences Program
  - Office of Naval Research
  - 800 N. Quincy
  - Arlington, VA 22217-5000

**Supplementary Notes:**
- To be published in Journal of Educational Measurement

**Abstract:**
See reverse

**Subject Terms:**
See reverse

**Distribution/Availability Statement:**
Approved for public release; distribution unlimited

**Number of Pages:** 36

**Price Code:** UL

---

**Report Date:** 10 August 1992

**Report Type and Dates Covered:** Technical: 1990-93

**Funding Numbers:**
- N00014-90-J-1940

**Performing Organization Report Number:**
- 1992 - No. 3

**Sponsoring/Monitoring Agency Report Number:**
- 4421-548
Assessing Dimensionality of a Set of Items—Comparison of Different Approaches

Abstract
This study examines the performance of the following four methodologies for assessing unidimensionality: DIMTEST, Holland and Rosenbaum's approach, linear factor analysis, and nonlinear factor analysis. Each method is examined and compared with other methods on simulated data sets and on real data sets. Seven data sets, all with 2000 examinees, were generated: three unidimensional, and four two-dimensional data sets. Two levels of correlation between abilities were considered: $\rho=.3$ and $\rho=.7$. Eight different real data sets were used: four of them were expected to be unidimensional, and the other four were expected to be two-dimensional. Findings suggest that, while the linear factor analysis often overestimated the number of underlying dimensions, the other three methods correctly confirmed unidimensionality but differed in their ability to detect lack of unidimensionality. DIMTEST showed excellent power in detecting lack of unidimensionality; Holland and Rosenbaum's and nonlinear factor analysis approaches showed good power, provided the correlation between abilities was low.

Subject terms: DIMTEST, unidimensionality, essential dimensionality, non-linear factor analysis, item response theory.
It is well known that most item response theory (IRT) models require the assumption of unidimensionality. According to Lord and Novick (1968), dimensionality is defined as the total number of abilities required to satisfy the assumption of local independence. If there is only one ability affecting the responses of a set of items to meet the assumption of local independence, then that set is referred to as a unidimensional set. It has also been long argued that responses to test items are multiply determined (Humphreys, 1981, 1985, 1986; Hambleton & Swaminathan, 1985, chap. 2; Reckase, 1979, 1985; Stout, 1987; Traub, 1983; Yen, 1985), and several abilities unique to items or common to relatively few items are inevitable. The ability which the test is intended to measure (i.e., the ability common to all items) will be referred to as the dominant ability, and abilities unique to or influencing responses to few items will be referred to as minor abilities. Given that item responses are multiply determined, it is intuitively clear that, in order to satisfy the assumption of unidimensionality, it is required that a given test measure a single dominant ability. A number of simulation studies have demonstrated that a dominant ability can be recovered well, using computer programs such as LOGIST, in the presence of several minor factors (Reckase, 1979; Drasgow & Parsons, 1983; Harrison, 1986). Although counting only dominant dimensions violates Lord and Novick's (1968) definition of dimensionality, it is commonly accepted that, in order to apply unidimensional item response theory models, it is sufficient to show that there is one dominant ability underlying the responses to a set of items.

Stout (1987, 1990) provided a mathematically rigorous definition of dominant dimensionality referred to as essential dimensionality and provided a statistical test (DIMTEST) to assess whether a set of items met the requirement for essential unidimensionality. Junker (1988, 1991) further explored essential dimensionality for dichotomous and polytomous items and established consistency results for the maximum likelihood ability estimates of \( \theta \) under essential unidimensionality. Essential dimensionality is the total number of abilities required to satisfy the assumption of essential independence.
Assessing Dimensionality—Comparison

An item pool is said to be essentially independent (EI) with respect to the latent variable vector \( \Theta \) if, for a given subset of items, the average absolute conditional (on \( \Theta \)) covariances of responses to item pairs approaches zero as the length of the subset increases. When conditional covariances based on only one dominant ability meet the assumption of essential independence, the response data is said to be essentially unidimensional \( (d_E=1) \). In contrast, the assumption of local independence requires that the conditional covariances be zero for responses to any item pair, and the number of abilities required to those conditional covariances is the dimensionality. According to this definition of dimensionality, all major and minor abilities influencing item responses have to be considered when assessing the local independence assumption; whereas, according to the essential dimensionality, it is sufficient to consider only the influence of dominant abilities. Hence, essential independence and essential dimensionality are weaker forms of local independence and traditional dimensionality respectively.

Stout's definition of essential dimensionality is conceptually based on an infinite item pool. An infinite item pool can be conceptualized in two ways: 1. as a consequence of continuing the test construction process beyond the \( N \) items of the test being studied where the \( N \) items become a subset of the item pool; 2. as a consequence of a sequence of finite tests where each finite test is optimally constructed. For example, a 20-item test is constructed with the knowledge that the test is going to be only 20 items long and that it is not necessarily a subset of an optimal 40-item test. In this way, an item pool is a collection of optimal finite test length tests (for details see Junker, 1991; Junker & Stout, 1991).

In assessing essential unidimensionality of given item responses, DIMTEST assesses the likelihood that the given set of item responses come from an essentially unidimensional item pool. That is, DIMTEST assesses whether or not the model generating the given item responses is close to the EI, \( d_E=1 \) model. The major focus in assessing essential unidimensionality of a given set of item responses is to determine how "minor" the influence of minor abilities is and whether the influence of these minor abilities can be
ignored when assessing essential unidimensionality.

Historically speaking, linear factor analysis has been used to assess the dimensionality of the latent space underlying the responses to a set of items. If the results indicate a one-factor solution, then it can be inferred that one dominant ability is influencing item responses. There are, however, a number of technical as well as methodological problems associated with using linear factor analyses to assess dimensionality. For example, difficulty levels of items and guessing levels of multiple-choice items can each play a major role in affecting the factor structure of item responses (for details see Carroll, 1945; Hulin, Drasgow, & Parsons, 1983, chap. 8; Zwick, 1987). Consequently, many attempts have been made by researchers in recent years to develop new methods to assess dimensionality. Some of the recently developed methods include nonlinear factor analysis (McDonald & Aihlawat, 1974); Bejar’s procedure (Bejar, 1980); order analysis (Wise, 1981); modified parallel analysis (Hulin, Drasgow, & Parsons, 1983, p. 255); residual analysis (Hambleton & Swaminathan, 1985, p. 163); Bock’s full information factor analysis (Bock, Gibbons, & Muraki, 1985); Holland and Rosenbaum’s test of unidimensionality, monotonicity, and conditional independence (Rosenbaum, 1984; Holland & Rosenbaum, 1986); Roznowski, Tucker, and Humphreys’ procedures (1991); and Stout’s unidimensionality procedure DIMTEST (Stout, 1987).

Hattie (1985), Hambleton and Rovinelli (1986), and Berger and Knol (1990) have reviewed several procedures for assessing dimensionality, including some of the above mentioned procedures. The main focus of this paper is to study and compare some of the procedures to assess dimensionality that are most recent, seem promising, and are little studied. Four procedures are considered and compared in this paper: DIMTEST, Holland and Rosenbaum’s procedure, nonlinear factor analysis, and linear factor analysis. Linear factor analysis was used, because of its historical importance, as a benchmark to compare other procedures. Several sets of unidimensional and multidimensional test data were simulated and used to study the performance of all four procedures for assessing
Assessing Dimensionality—Comparison

dimensionality. The same procedures were then repeated with real test data.

Description of Procedures

Linear Factor Analysis

Linear factor analysis is the most commonly used approach to assess dimensionality. With linear factor analysis, each extracted factor is presumed to represent a dimension, and items that load heavily on a given factor are considered good measures of that dimension. There are a number of fundamental problems associated with applying linear factor analysis to binary data. First, linear factor analysis assumes that the relationship between the observed variables and the underlying factors is linear and that the variables are continuous in nature. But it is clear for dichotomous data that the relationship between the performance and the underlying latent variable is not linear. Hence, applying factor analysis to phi or tetrachoric correlations of binary item responses produces difficulty factors (Hulin, Drasgow, & Parsons, 1983, chap. 8). Second, in computing tetrachoric correlations, the cell entries of the fourfold table for a pair of dichotomous items sometimes equal zero, making it difficult to determine an appropriate value for the correlation. Third, determination of the number of significant factors could be problematic.

In this study the statistical package LISCOMP was used to perform exploratory linear factor analysis using tetrachoric correlations. Three different approaches were used to determine the number of significant factors: parallel analysis, the chi-square test of goodness of fit, and goodness of fit statistics (the means and standard deviations of the squares of residual correlations and absolute residuals).

According to parallel analysis (Humphreys & Montanelli, 1975), the eigenvalues of the given correlation matrix are compared with the eigenvalues of random data. The random data consist of binary responses generated with the same number of items and examinees as that of the given data. The largest eigenvalue from the random data is used
as the cutoff point for eigenvalues from the actual data to determine the number of significant factors. That is, the number of eigenvalues of the actual data greater than the largest eigenvalue of the random data is taken as the significant number of factors underlying the given data.

The second method used to determine the number of factors was the chi-square test of goodness of fit from LISCOMP. The third method involves comparisons of means and standard deviations of squares of residuals and absolute values of residuals after fit of an m-factor model with the corresponding values from the random data. If the residuals are sufficiently "small," then one can regard the fit of the model as "reasonably satisfactory" (McDonald, 1981; Hattie, 1985, Hambleton & Rovinelli, 1986; and Berger & Knol, 1990).

Nonlinear Factor Analysis

McDonald (1967, 1980, 1982) and McDonald and Ahlawat (1974) have demonstrated that applying linear factor analysis to unidimensional binary data yields "nonlinear factors" rather than "difficulty factors." Nonlinear factors account for nonlinear relationships among the variables by using higher order polynomials in the factor model (for example, quadratic and cubic terms). McDonald developed the method of nonlinear factor analysis (NLFA) to account for the nonlinearity of the data as an improvement over linear factor analysis. The variables in the model can be expressed as polynomial functions of latent traits or factors. For example, a two-factor model with linear and quadratic terms would be of the following form:

\[ Y_i = b_{i0} + b_{i1} \theta_1 + b_{i2} \theta_2 + \theta_1^2 + \theta_2^2 + d_i u_i \]  

where \( Y_i \) denotes the examinee's score on item \( i \), \( \theta_1 \) and \( \theta_2 \) denote latent traits, \( b_{ijk} \) denotes the factor loading of the \( i \)-th item on the \( j \)-th common factor for the \( k \)-th degree
element in the polynomial; \( u_i \) denotes the unique factor and \( d_i \) denotes the unique factor loading for item \( i \). Hambleton and Rovinelli (1986) have demonstrated the use of NLFA to assess dimensionality and found it to be a promising method. They, however, caution about the criterion for the adequacy of the fit of the model.

In the present study, NLFA embodied in the computer program NOFA, developed by Etazadi-Amoli and McDonald (1983), was used. The fit of the model is studied just as in the case of the linear factor analyses, by comparing the means and standard deviations of squared residuals and absolute residuals with the corresponding values of random data and linear factor analyses. The chi-square statistic values are not available from NOFA.

Holland and Rosenbaum's Test of Lack of Fit of a Unidimensional, Monotone, and Conditional Independent Model

Rosenbaum (1984) and Holland and Rosenbaum (1986) have proved theorems concerning conditional association that can be applied to assess dimensionality. The basic notion in Holland and Rosenbaum's (H&R) theorems is that if the items are locally independent, unidimensional, and the item characteristic curves are monotone, then the items are conditionally positively associated. Specifically, the conditional covariances between any pair of item response functions of a set of unidimensional dichotomous item responses given any function of the remaining item responses will be nonnegative. The test of this relationship can be specified as

\[ H_0: \text{Cov}(X_i, X_j | \sum_k X_k) \geq 0 \quad \text{vs.} \quad H_1: \text{Cov}(X_i, X_j | \sum_k X_k) < 0 \]

Conditional associations for each pair of items is tested, given the number-right score on the remaining items. The Mantel-Haenszel test (M-H) (Mantel & Haenszel, 1959)
is used to test this hypothesis. To perform the M-H test on a given pair of items, a 2x2 contingency table is constructed for the pair for each of the possible number-right scores on the remaining items. The cell values of a 2x2 table for item pair i and j for examinees with total score \( k (k=1,2,\ldots,K) \) on the remaining items can be denoted as the following: the number of examinees who got both item i and item j correct \( (n_{11k}) \), the number of examinees who got both item i and item j incorrect \( (n_{00k}) \), the number of examinees who got item i correct and item j incorrect \( (n_{10k}) \), and the number of examinees who got item i incorrect and item j correct \( (n_{01k}) \). The M-H statistic is then given by

\[
Z = \frac{n_{11+} - E(n_{11+}) + 1/2}{\sqrt{V(n_{11+})}}
\]  

(1)

where \( n_{11+} = \sum_{k=1}^{K} n_{11k} \) and \( E(n_{11+}) \) and \( V(n_{11+}) \) are the expectation and the variance of \( n_{11+} \) given by

\[
E(n_{11+}) = \sum_{k=1}^{K} \frac{n+0+k n+1+k}{n++k} \sum_{k=1}^{K} \frac{n+0+k n+1+k}{n++k} 
\]  

(2)

and

\[
V(n_{11+}) = \sum_{k=1}^{K} \frac{n+0+k n+1+k}{n++k} \sum_{k=1}^{K} \frac{n+0+k n+1+k}{n++k} \frac{n+0+k n+1+k - 1}{n++k} \frac{n+0+k n+1+k - 1}{n++k-1} 
\]  

(3)

The plus subscript in Equations 2 and 3 denotes the summation over that subscript. The computed Z-value is compared to the lower tail of the standard normal distribution. A statistically significant Z implies that the pair of items in question are not conditionally associated, given the sum of the remaining items and are thus inconsistent with the unidimensional model. In this manner, the M-H statistic is computed for all \( N(N-1)/2 \)
Assessing Dimensionality—Comparison

If a "large" number of pairs are shown not to be conditionally associated, then the unidimensional assumption is inappropriate.

Since H&R approach tests each item pair with significance level \( \alpha \), the simultaneous inference for all item pairs can be based on Bonferroni bounds (Holland & Rosenbaum, 1986, Junker, 1990, and Zwick, 1987). According to Bonferroni bounds, one would accept \( H_0 \) if the number of rejections at level \( \alpha \) is around \( t \alpha \), where \( t \) is the number of tests performed, which is equal to \( N(N-1)/2 \); one would reject \( H_0 \) if at least one test is rejected at level \( \alpha/t \).

Rosenbaum (1984), Zwick (1987), and Ben-Simon and Cohen (1990) have demonstrated the application of H&R approach to assess dimensionality. Ben-Simon and Cohen found the H&R approach to be conservative and erroneously misclassified nearly half of the multidimensional item pools they analyzed as unidimensional. Zwick found H&R approach to be consistent with other procedures investigated in assessing unidimensionality of NAEP reading data.

**DIMTEST**

Stout (1987) developed DIMTEST to test the hypothesis of essential unidimensionality: the existence of one dominant dimension. Nandakumar and Stout (in press) further modified and improved the performance of DIMTEST. The improvements have lead to the following: a robust procedure against presence of guessing in item responses; a better control of the observed level of significance, and greater power; and automation of the size of assessment subtests, as described below. The hypothesis to test unidimensionality can be stated as
Assessing Dimensionality—Comparison

\[ H_0: d_E = 1 \quad \text{vs.} \quad H_1: d_E > 1 \]

where \( d_E \) denotes the essential dimensionality of the item pool of which the given test items are a part.

In order to apply DIMTEST, it is assumed that a group of \( J \) examinees take an \( N \)-item test. Each examinee produces a vector of responses of 1s and 0s with 1 denoting a correct response and 0 denoting an incorrect response. It is also assumed that essential independence with respect to some dominant ability \( \Theta \) holds and that the item response functions are monotone with respect to the same dominant ability \( \Theta \). DIMTEST has several steps. These are briefly described here (for details see Stout, 1987; Nandakumar and Stout, in press).

Step 1: The \( N \) items of the test are split into three subtests: \( \text{AT}1 \), \( \text{AT}2 \), and \( \text{PT} \). First, \( \text{AT}1 \) items are selected so that these items all measure the same dominant ability. This can be achieved either through factor analysis (FA) or through expert opinion (EO). If FA method is chosen, \( M \) items with highest loadings on the second factor (before rotation) are selected. In this case, the program automatically determines the size \( M \) of \( \text{AT}1 \) as a function of the test length and the sample size. If EO is sought, on the other hand, it is recommended that, at most, one-quarter of the total items should be selected that tap the same ability. After selecting items of \( \text{AT}1 \), items of \( \text{AT}2 \) are selected, also of the same size \( M \), so that items of \( \text{AT}1 \) and \( \text{AT}2 \) have the same difficulty distribution (for details see Stout, 1987). The remaining items \((n=\text{N}-2M)\) form the partition subtest \( \text{PT} \). In the present study, FA is chosen to select \( \text{AT}1 \) items. For examples where EO is used to select \( \text{AT}1 \) items, see Nandakumar (in press).

When FA is used to select \( \text{AT}1 \) items, the given sample of \( J \) examinee responses are partitioned into two groups. One group of examinee responses (500 examinees recommended) is used for exploratory factor analysis to select \( \text{AT}1 \) and \( \text{AT}2 \) items, and the other group of examinee responses is used to compute the Stout’s statistic \( T \).
Assessing Dimensionality—Comparison

Step 2: The second group of examinees (if the first group of examinees is used for FA) are partitioned into K subgroups based on their PT score. That is, all examinees obtaining the same total score on PT are assigned to the same subgroup \( k = 1, 2, \ldots, K \).

Step 3: Within each subgroup \( k \), examinee responses to subtest items AT1 and AT2 are used to compute the unidimensional statistic \( T \) given by

\[
T = (T_1 - T_2)/\sqrt{2},
\]

where

\[
T_i = \frac{1}{K^{1/2}} \sum_{k=1}^{K} \left[ \frac{\hat{\sigma}^2_k - \hat{\sigma}_{U,k}^2}{S_k} \right]
\]

is computed using items of ATi. The \( \hat{\sigma}^2_k \) and \( \hat{\sigma}_{U,k}^2 \) and \( S_k \) are given as follows.

The usual variance estimate for subgroup \( k \) is given by

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

where

\[
Y_{j}^{(k)} = \sum_{i=1}^{M} U_{ijk}/M, \text{ and } \bar{Y}^{(k)} = \sum_{j=1}^{J_k} Y_{j}^{(k)}/J_k
\]

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

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

where

\[
\hat{p}_i^{(k)} = \sum_{j=1}^{J_k} U_{ijk}/J_k
\]
And the standard error of estimate for subgroup $k$ is given by

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

where

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

and

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

The computed $T$-value is referred to the upper tail of the standard normal distribution to obtain the significance level. The significant values associated with unidimensional tests are expected to be large while the significant values associated with multidimensional tests are expected to be within the margin of the specified level of significance.

DIMTEST assesses the degree of closeness of an essentially unidimensional model to the model generating the observed data. This is done by splitting the test items into three subtests—AT1, AT2, and PT—as described above. When the model underlying the test item responses is close to essentially unidimensional, items of AT1, AT2, and PT would all be of the same dominant dimension; therefore, the value of the statistic $T$ computed based on AT1, AT2 would be "small," leading to the tenability of $H_0$. When the model underlying the test responses is not essentially unidimensional, however, items of AT1 would be dimensionally different from items of AT2 and PT and the value of the statistic $T$ will be "large" leading to the rejection of $H_0$.

DIMTEST has been found to discriminate between unidimensional and two-dimensional tests for a variety of simulated test data when the correlation between abilities is as high as .7 (Stout, 1987; Nandakumar & Stout, in press). Nandakumar (1991)
Assessing Dimensionality—Comparison

has shown the usefulness of DIMTEST to assess essential unidimensionality in the possible presence of several minor abilities. The findings indicate that essential unidimensionality is established when each of the minor abilities influence relatively few items, or, if minor abilities are influencing many items, the strength of the influence of the minor abilities is low. As the strength of the minor abilities increases, the approximation to an essentially unidimensional model degenerates, inflating the type—I error of the test of hypothesis of essential unidimensionality. Nandakumar (in press) has further replicated these findings on a wide variety of real test data. This study also demonstrates the sensitivity of DIMTEST to major and minor abilities influencing item responses.

Description of Test Data

The Simulated Test Data

Seven data sets, DATA1—DATA7, were generated. Of the seven, three data sets, DATA1—DATA3, are strictly unidimensional, consisting of 25, 40, and 50 items, respectively. The other four data sets, DATA4—DATA7, are two-dimensional with length $N=25$ and correlation between abilities $\rho=.3$, $N=25$ and $\rho=.7$, $N=50$ and $\rho=.3$, and $N=50$ and $\rho=.7$, respectively. All 7 data sets have 2000 examinees. These data set characteristics are summarized in Table 1.

Table 1 about here

The unidimensional data sets were generated using the three—parameter logistic model given by

\[ 16 \]
Assessing Dimensionality—Comparison

\[ P_i(\theta) = c_i + \frac{1-c_i}{1 + e^{x p \{ -1.7 \left[ a_i (\theta_b - b_i) \right] \} }}. \]  

(5)

The abilities \((\theta)\) were independently generated from the standard normal distribution, and the item parameters \((a_i, b_i, c_i)\) of real tests as described in Nandakumar (1991) were used in generating item responses. For example, items of DATA 1 have a larger variability in discrimination power \((a_i)\), ranging from 1.22 to 2.82; items of DATA 2 have a smaller variability of \(a_i\), ranging from 1.07 to 2.00. For each simulated examinee, the probability of correctly answering each item, \(P_i(\theta)\), was computed using the three-parameter logistic model. For each item \(i\), a random number between 0 and 1 was generated from a uniform distribution. If the computed probability, \(P_i(\theta)\), was greater than or equal to the random number generated, the examinee was said to have answered the item correctly and was given a score of 1; otherwise the examinee was given a score of 0. The two-dimensional test data were generated according to the multidimensional compensatory model (Reckase & McKinley, 1983) given by

\[ P_i(\theta_1, \theta_2) = c_i + \frac{1-c_i}{1 + e^{x p \{ -1.7 \left[ a_{i1} (\theta_1 - b_{i1}) + a_{i2} (\theta_2 - b_{i2}) \right] \} }}. \]  

(6)

The abilities \(\theta = (\theta_1, \theta_2)\) were sampled from a bivariate normal distribution with both means zero and both variances one. Two levels of correlation coefficients between the abilities were used: .3 and .7. The guessing level was taken to be .20 for all tests. The discrimination parameters \((a_{i1}, a_{i2})\) for each item were independently generated as follows:

\[ a_{i1} \sim N \left[ \mu, \frac{\sigma}{2} \right], \quad a_{i2} \sim N \left[ \mu, \frac{\sigma}{2} \right], \]

where \(\mu\) and \(\sigma\) are the mean and standard deviation of the distribution of discrimination.
parameters of the respective unidimensional tests with the same number of items. Similarly $b_{1i}$ and $b_{2i}$ were assumed to be independent of each other for each item and were generated as follows:

$$b_{1i} \sim N(\mu, \sigma), \quad b_{2i} \sim N(\mu, \sigma),$$

where $\mu$ and $\sigma$ are the mean and standard deviation of the distribution of difficulty parameters of the respective unidimensional test with the same number of items. For example to generate test data DATA4 with $N=25$ and $\rho=.3$, the means and standard deviations of $a_i$'s and $b_i$'s of item parameters used for DATA1 were used. The item responses (0,1) were generated exactly as described for unidimensional case by using $P_1(\theta)$ of (6).

The Real Test Data

The real test data used in this study came from two different sources. The National Assessment of Educational Progress (NAEP, 1988) data for the 1986 US History (HIST) and Literature (LIT) for grade 11/age 17 were obtained from Educational Testing Service. The Armed Services Vocational Aptitude Battery (ASVAB) data for Arithmetic Reasoning (AR) and General Science (GS) for grade 10 were obtained from Linn, Hastings, Hu, and Ryan (1987). For all data sets, examinees who missed one or more items were deleted from the analyses. Test sizes and sample sizes for all real tests are given in bottom half of Table 1. Since all four test data were assessed as unidimensional by the methods employed in this article (details are provided in Results section), they were combined to form two-dimensional tests. Four two-dimensional tests were formed as follows. The test data HSTLIT1 was formed by combining the data of 31 items of HIST with the data of 5 items of LIT randomly selected from 30 items. Similarly HSTLIT2 was formed by combining the responses of 31 items of HIST with the responses of 10 items of LIT, and the test data GS
was formed by combining responses of 30 items of AR with the responses of 10 items of GS. The two-dimensional test HSTGEO contains 31 history items spanning US history from the colonization period to modern times (HIST) and in addition contains 5 map items requiring the knowledge of geographical location of different countries in the world. This is the actual history test according to NAEP. But it was shown using DIMTEST that the 5 map items formed a separate dimension significantly different from history items (Nandakumar, in press). Hence the data on these 5 map items were removed from the history test to form HIST with 31 items, and the original history data were treated as a natural two-dimensional test.

Results

The results of DIMTEST and the H&R approach will be studied together and compared because of the similarity in the underlying theory and because both of them are statistical tests. Likewise the results of linear and nonlinear factor analysis will be studied and compared together.

The Simulated Test Data

DIMTEST and H&R Procedure

The results of DIMTEST and the H&R approach for simulated data are presented at the top of Table 2. For all data sets, the significance levels associated with DIMTEST indicate that DIMTEST is able to correctly confirm unidimensionality and detect lack of unidimensionality for both correlation (between abilities) levels $\rho=.3$ and $\rho=.7$. For example, all three unidimensional data sets, DATA1–DATA3, have small $T$–values and large significant values, implying the acceptance of the null hypothesis of essential unidimensionality (here the data were simulated as strictly unidimensional). Two-dimensional data, DATA4–DATA7, on the other hand, have large $T$–values, strongly
rejecting the null hypothesis of essential unidimensionality.

Table 2 about here

The results of the H&R approach indicate that for unidimensional tests, the number of significant negative partial associations at level $\alpha (\alpha=.05)$ are far below the expected number ($t\alpha$), strongly confirming the unidimensional nature of these data sets. Among the two-dimensional data sets, DATA4 and DATA6 ($\rho=.3$) were correctly assessed as multidimensional. For these data, the number of significant negative partial associations at level $\alpha$ were beyond $t\alpha$ level, and the number of significant negative partial associations beyond level $\alpha/t$ were 15 and 1, respectively, identifying them as multidimensional. The test data DATA5 and DATA7 ($\rho=.7$), on the other hand, were assessed as unidimensional. For DATA5 and DATA7, the number of significant negative partial associations at level $\alpha$ were within $t\alpha$ level, and the number of significant negative partial associations beyond level $\alpha/t$ was zero, making them unidimensional tests. It was disappointing to note that for many of the item pairs measuring different traits, in two-dimensional tests, the covariance did not approach significance. One reason for this could be the noise in the conditional score. More research is necessary to draw definite conclusions.

Linear and Nonlinear Factor Analysis

The computer programs used to do the analyses, LISCOMP and NOFA, are heavily computationally intensive and consume enormous CPU time. In addition, LISCOMP cannot handle more than about 40 variables. For these reasons, not all data sets were included in the linear factor analyses, but all data sets were included in the nonlinear factor analyses. The results of linear and nonlinear factor analyses are presented in Table 3.
Based on parallel analyses, one factor would be retained for DATA1, DATA2, and DATA5; two factors would be retained for DATA4. Whereas, according to the significance levels associated with a chi-square test of goodness of fit, in Table 3, a two-factor model fits DATA1, a four-factor model fits DATA2 and DATA4, and a three-factor model fits DATA5. Similar chi-square values are not available for nonlinear models.

The goodness of fit statistics—the means and standard deviations of squared residuals and absolute residuals—are reported for all data sets in Table 3. The top entry in Table 3 refers to random data (RANDOM) with 25 variables and 2000 examinees. Because of the cost of computations, only one random data set was used to compare the goodness of fit statistics. Comparing goodness of fit statistics of RANDOM with DATA1, it appears that both one-factor quadratic and one-factor cubic models fit as well as the four-factor linear model. However, since the differences in the magnitude of residuals among models are small, one could argue that four-factor linear and one-factor quadratic or cubic models are over fit and that one should go with a more parsimonious model. Observance of the significance values of the chi-square test of goodness of fit indicates that the two-factor model fits the data. If one strictly applies the criterion of using random data residuals as a guide to determine the number of factors, however, a one-factor model with a quadratic term seems to be the right choice. Similar observations can be made for DATA2.

Comparing goodness of fit statistics for linear and nonlinear factor analysis, it can be seen that for DATA4 and DATA5, the two-factor quadratic model fits better than the three-factor linear model, confirming the two-dimensional nature of data. Here again one could argue, based on the absolute residuals, that the differences in the residuals are small and that the quadratic models or three-factor and four-factor linear models are an over fit.
The significant values associated with the chi-square test indicate overestimation of factors for DATA4. As expected, the means and the standard deviations of squared residuals and absolute residuals are much larger for DATA4 ($\rho=.3$) than for DATA5 ($\rho=.7$), reflecting more deviation from unidimensionality for DATA4. For DATA5, the goodness of fit analyses support a one-factor quadratic model. Likewise the two-factor quadratic model fits DATA6, and one-factor quadratic model fits DATA7.

In summary, there are many criteria that can be used to assess dimensionality by linear factor analysis approach. The different criteria may give rise to different conclusions regarding the dimensionality of the data set in consideration. In the present study it is shown that the significant values associated with the chi-square test overestimated the number of factors in most cases. Parallel analyses correctly identified the dimensionality in some cases. Nonlinear factor analyses exhibited a better fit than the linear factor analyses. DIMTEST and H&R procedures were excellent in confirming unidimensionality. DIMTEST demonstrated greater power in detecting multidimensionality for correlations between abilities as high as .7. H&R and nonlinear factor analysis methods demonstrated good power provided the correlation between abilities was low ($\rho=.3$).

The Real Test Data

DIMTEST and H&R Procedure

The results of DIMTEST and H&R for real data sets are presented at the bottom of Table 2. For data sets LIT, HIST, AR, and GS, the $T$-values associated with DIMTEST indicate that these data can be approximated by an essentially unidimensional model. The results of H&R approach for these data are also consistent with DIMTEST results in that the number of significant negative partial associations, for each one of the tests, is less than the nominal level $\alpha$. While both approaches strongly support that HIST, AR, and GS are essentially unidimensional, the decision is not clear for LIT because there is one negative
partial association that is significant beyond level $\alpha/t$, and the $T$–value of DIMTEST is in
the border line region, indicating presence of violations to the unidimensionality
hypothesis.

For two–dimensional data HSTLIT1, HSTLIT2, ARGS, and HSTGEO, the
$T$–values associated with DIMTEST strongly indicate the multidimensional nature of these
data. Relatively large $T$–values associated with ARGS and HSTGEO indicate that abilities
within these tests are more orthogonal than abilities in HSTLIT1 and HSTLIT2. The
results based on H&R approach, however, indicate that all four data sets are
unidimensional. For each one of the two–dimensional data sets, the number of significant
negative partial associations is well below the nominal level $\alpha t$, and none of the partial
associations are significant beyond level $\alpha/t$. Even with a liberal $\alpha = .10$, the number of
negative partial associations did not rise above the nominal level for any of the tests. These
results suggest that the H&R approach lacks power.

On further examination of H&R results, it was found that the $M–H Z$–values for
many of the item pairs, where items were supposed to be measuring different traits, did not
reach significance level. One explanation for this could be that for these item pairs, the
conditional score ($\Sigma X_k$), on the basis of which the examinees are classified into different
groups, may be contaminated with items tapping different abilities. This could be
especially true for HSTLIT2 and ARGS where one quarter of the test items are from the
second dominant dimension. Because of the noise in the conditional score distribution, the
covariance of item pairs measuring different abilities may not be exhibiting significant
negative covariance. A proper conditional score may considerably increase the power of the
H&R approach.

Linear and Nonlinear Factor Analysis

The results of linear and nonlinear factor analysis for a selection of real data sets are
reported in Table 4. The results are consistent with the simulated test data in that for all
cases nonlinear factor models fit better than linear factor models. According to the chi-square test of goodness of fit, the four-factor model was best fitting for all data sets where linear factor analysis was performed. Based on goodness of fit statistics, a one-factor quadratic model fits LIT, AR, and HSTLIT1 better than three- or four-factor linear models. Since a one-factor quadratic model fits as well as a two-factor quadratic model, a more parsimonious model is strongly recommended in these cases. For HSTLIT2 and ARGS, again it appears that a one-factor quadratic model is appropriate. If chi-square statistics were available along with the goodness of fit statistics for nonlinear factor analyses, it would have aided in the interpretation.

Table 4 about here

In summary, for real data sets, the results are somewhat consistent with simulated data sets. For data sets assessed as unidimensional by DIMTEST and H&R, the chi-square tests based on the linear factor analysis indicated a four-factor model for the same data. Although we do not know the true dimensionality of real data, these results suggest that linear factor analysis is overestimating the underlying dimensionality. Whereas, the other three methodologies were excellent in identifying essential unidimensionality but differed in identifying lack of unidimensionality. DIMTEST demonstrated greater power than either the H&R or the nonlinear factor analysis methods. It appears that with the appropriate conditional score the power of the H&R approach could be improved, and with some type of fit statistics and the associated significance levels, the power of nonlinear factor analysis could be improved.
Assessing Dimensionality—Comparison

Discussion

Based on this limited study, findings demonstrate that the linear factor analysis approach to assessing essential unidimensionality is not satisfactory. This finding is consistent with the previous research and theory (see for example, Hambleton & Rovinelli, 1986; Hattie, 1984). In contrast to linear factor analysis, DIMTEST, H&R, and nonlinear factor analysis were each shown to be promising methodologies to assess dimensionality.

In this study, all three methodologies exhibited sensitivity to discriminate between one— and two—dimensional test data. For simulated unidimensional test data, all three procedures were able to confirm unidimensionality. For the real data, all three procedures were consistent in identifying unidimensionality of HIST, AR, and GS. For two—dimensional test data, however, the three procedures differed in their ability to detect the lack of unidimensionality. DIMTEST rejected the null hypothesis of essential unidimensionality for all two—dimensional tests: both real and simulated. The H&R approach confirmed the lack of unidimensionality for two—dimensional simulated tests, provided the correlation between abilities was low (ρ=.3). For simulated test data with high correlation between abilities (ρ=.7), the H&R approach was unable to detect multidimensionality. Also, for all two—dimensional real test data, the H&R approach was unable to detect multidimensionality.

The performance of the nonlinear factor analysis methodology was similar to the H&R procedure for two—dimensional data sets. For simulated test data with ρ=.3, the two—factor model with linear and quadratic terms demonstrated adequate fit statistics (smaller means and standard deviations of squared residuals and absolute residuals). For simulated tests with ρ=.7, however, the difference in fit statistics between one—factor and two—factor quadratic models was not evident. Similarly for two—dimensional real test data HSTLIT2 and ARGs, the difference in fit statistics between one—factor and two—factor models with linear and quadratic terms was not evident. The difficulty in deciding about
the correct model arises because there is no concrete way of assessing what is meant by "sufficiently small" for goodness of fit statistics.

In this study, the results associated with the H&R approach were consistent with the findings of the Ben-Simon and Cohen's (1990) and Zwick's (1987) studies. The number of significant negative partial associations for unidimensional tests was far below the expected five percent level, making it a very conservative test. Consequently, it did not exhibit high power. The reason one observes fewer than the nominal level of negative partial associations is that the conditional score used in computing the covariances is not perfectly correlated with the latent variable (Zwick, 1987). According to the theorems proved by Holland and Rosenbaum (1986), the conditional score used to compute the covariances can be any function of the latent trait. An appropriate choice of conditional score, therefore, could maximize the power of H&R approach.

The results of nonlinear factor analyses were consistent with the findings of Hambleton and Rovinelli (1986). Factor models with linear and quadratic terms were able to fit the data better than models with just linear terms. The problem with nonlinear factor analysis is determining the appropriate number of polynomial terms to retain in the model. This problem suggests that some type of adequacy of fit statistics with associated sampling distribution would be necessary to aid in assessing the fit of nonlinear models.

In terms of assessing the degree of multidimensionality, both the DIMTEST and nonlinear factor analysis approaches can be useful. The T-values associated with DIMTEST and the fit statistics associated with nonlinear factor analysis can be helpful in assessing the degree of multidimensionality. For example, both HIST and AR are considered as essentially unidimensional data sets, but the associated T-values are -1.53 and 1.18 respectively. By contrast, for a two-dimensional data set HSTLIT2, T=2.03. The difference in the T-values mirrors the degree of multidimensionality present in the data. Similarly, the difference in fit statistics between one-factor and two-factor quadratic models for DATA1 and DATA4 reflects the degree of multidimensionality.
In the present study, the test length is more than 25 items, and the sample sizes are around 2000 examinees. It is not known if the results would hold up for small test lengths and sample sizes. De Champlain and Gessaroli (1991) have shown that DIMTEST loses power when both the test length and the sample size are small (for example, \( N=25 \) and \( J=500 \)). Their results show support for the use of incremental fit index (IFI) using the nonlinear factor analysis program, NOHARM II, to assess dimensionality in cases of smaller test lengths and sample sizes. Ben–Simon and Cohen (1990) have found that the test length and the sample size had a marked effect on the M–H \( Z \)-statistic in the detection of multidimensionality. In their study they tried test lengths of 20, 30, 40, and 50 and sample sizes of 1000, 2000, 3000, and 4000. They found that larger samples and larger tests facilitated the detection of multidimensionality. They urge a cautious interpretation of M–H test results in light of test lengths and sample sizes.

Just as linear and nonlinear methodologies share the same philosophical theory, DIMTEST and H&R approaches share the same theoretical framework. The basic rationale for the H&R approach is to reject the locally independent, monotone, unidimensional model if the conditional covariances are significantly negative. By contrast, DIMTEST rejects the essentially independent, monotone, essentially unidimensional model if the conditional covariances are significantly positive (it can be shown that the expected value of the numerator of Stout’s statistic \( T \) is mathematically equivalent to average conditional covariances among AT1 items, Stout (1987)). This apparent contradiction in the criterion for assessing unidimensionality may be resolved by noting the subtle difference in item pair covariances under consideration. In the H&R approach, one expects the conditional covariance between items measuring different traits to be negative; whereas in Stout’s approach, one expects the asymptotic conditional covariance between items measuring the same trait to approach zero. DIMTEST is specifically designed to assess unidimensionality and thus looks for the existence of at least two dominant dimensions. By contrast, the H&R approach looks at all item pairs and detects items that are not measuring the same
As for the computational time involved, DIMTEST is most efficient. The computational time involved for other procedures is significantly more. For example, for a 25 item test with 2000 examinees, DIMTEST uses 4 seconds of CPU time, H&R approach uses 24 seconds, and nonlinear factor analysis uses 42 seconds; for a 50 items test with 2000 examinees, DIMTEST uses 8 seconds, H&R approach uses 106 seconds, and nonlinear factor analysis uses 191 seconds. As the test length increases, the H&R approach requires disproportionately more time, and the same is true for the nonlinear factor analysis as test length increases and/or the model gets more complex.
Notes

1. The reader is reminded that testing for unidimensionality is not synonymous to testing for model–data fit. If a unidimensional model is to be applied to the data, testing for unidimensionality is the first step. If item responses are essentially unidimensional, then as a second step, one can test for model–data fit, such as, one-parameter logistic, two-parameter logistic, etc.
Assessing Dimensionality—Comparison

References


Assessing Dimensionality—Comparison


Table 1  
Description of Data Sets

<table>
<thead>
<tr>
<th>Name</th>
<th>$J^1$</th>
<th>Traits</th>
<th>$\rho^2$</th>
<th>$N^3$</th>
<th>Trait1</th>
<th>Trait2</th>
<th>Mixed$^4$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Simulated data sets</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DATA1</td>
<td>2000</td>
<td>1</td>
<td></td>
<td>25</td>
<td>25</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>DATA2</td>
<td>2000</td>
<td>1</td>
<td></td>
<td>40</td>
<td>40</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>DATA3</td>
<td>2000</td>
<td>1</td>
<td></td>
<td>50</td>
<td>50</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>DATA4</td>
<td>2000</td>
<td>2</td>
<td>0.3</td>
<td>25</td>
<td>8</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>DATA5</td>
<td>2000</td>
<td>2</td>
<td>0.7</td>
<td>25</td>
<td>8</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>DATA6</td>
<td>2000</td>
<td>2</td>
<td>0.3</td>
<td>50</td>
<td>16</td>
<td>16</td>
<td>17</td>
</tr>
<tr>
<td>DATA7</td>
<td>2000</td>
<td>2</td>
<td>0.7</td>
<td>50</td>
<td>16</td>
<td>16</td>
<td>17</td>
</tr>
<tr>
<td><strong>Real data sets</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIT</td>
<td>2439</td>
<td>1</td>
<td></td>
<td>30</td>
<td>30</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>HIST</td>
<td>2428</td>
<td>1</td>
<td></td>
<td>31</td>
<td>31</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AR</td>
<td>1984</td>
<td>1</td>
<td></td>
<td>30</td>
<td>30</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GS</td>
<td>1990</td>
<td>1</td>
<td></td>
<td>25</td>
<td>25</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>HSTLIT1</td>
<td>2428</td>
<td>2</td>
<td></td>
<td>36</td>
<td>31</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>HSTLIT2</td>
<td>2428</td>
<td>2</td>
<td></td>
<td>41</td>
<td>31</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>ARG5</td>
<td>1853</td>
<td>2</td>
<td></td>
<td>40</td>
<td>30</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>HSTGEO</td>
<td>2440</td>
<td>2</td>
<td></td>
<td>36</td>
<td>31</td>
<td>5</td>
<td>0</td>
</tr>
</tbody>
</table>

$^1J$ denotes the number of examinees  
$^2\rho$ denotes the correlation between traits  
$^3N$ denotes the test length  
$^4mixed$ items are a combination of both traits 1 and 2.
Table 2
Results of DIMTEST and H&R Analyses

<table>
<thead>
<tr>
<th>Name</th>
<th>T</th>
<th>p&lt;</th>
<th>Decision based on DIMTEST</th>
<th>H&amp;R Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>No. of item pairs</td>
<td>No. of pairs significant at level α</td>
</tr>
<tr>
<td>Simulated test data</td>
<td></td>
<td></td>
<td>300</td>
<td>1</td>
</tr>
<tr>
<td>DATA1</td>
<td>-1.05</td>
<td>.85</td>
<td>accept H.</td>
<td>300</td>
</tr>
<tr>
<td>DATA2</td>
<td>-0.75</td>
<td>.77</td>
<td>accept</td>
<td>780</td>
</tr>
<tr>
<td>DATA3</td>
<td>-0.94</td>
<td>.83</td>
<td>accept</td>
<td>1225</td>
</tr>
<tr>
<td>DATA4</td>
<td>7.19</td>
<td>.000</td>
<td>reject</td>
<td>300</td>
</tr>
<tr>
<td>DATA5</td>
<td>3.62</td>
<td>.000</td>
<td>reject</td>
<td>300</td>
</tr>
<tr>
<td>DATA6</td>
<td>10.13</td>
<td>.000</td>
<td>reject</td>
<td>1225</td>
</tr>
<tr>
<td>DATA7</td>
<td>2.41</td>
<td>.008</td>
<td>reject</td>
<td>1225</td>
</tr>
<tr>
<td>Real test data</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIT</td>
<td>1.70</td>
<td>.045</td>
<td>accept</td>
<td>435</td>
</tr>
<tr>
<td>HIST</td>
<td>-1.53</td>
<td>.937</td>
<td>accept</td>
<td>465</td>
</tr>
<tr>
<td>AR</td>
<td>1.18</td>
<td>.118</td>
<td>accept</td>
<td>435</td>
</tr>
<tr>
<td>GS</td>
<td>-0.14</td>
<td>.555</td>
<td>accept</td>
<td>300</td>
</tr>
<tr>
<td>HSTLIT1</td>
<td>3.01</td>
<td>.036</td>
<td>reject</td>
<td>630</td>
</tr>
<tr>
<td>HSTLIT2</td>
<td>2.03</td>
<td>.021</td>
<td>reject</td>
<td>820</td>
</tr>
<tr>
<td>ARG5</td>
<td>6.15</td>
<td>.000</td>
<td>reject</td>
<td>780</td>
</tr>
<tr>
<td>HSTGEO</td>
<td>6.19</td>
<td>.000</td>
<td>reject</td>
<td>630</td>
</tr>
</tbody>
</table>

*significant at .05 level
Table 3
Results of Linear and Nonlinear Factor Analysis
For Simulated Test data: Goodness of Fit Statistics

|                  | \( \overline{\tau}_{ij}^2 \) | SD(\( \tau_{ij} \)) | |\( \tau_{ij} \)| | SD(|\( \tau_{ij} \)|) | \( p^* \) |
|------------------|-------------------------------|---------------------|--------------|----------------|----------------|
| RANDOM           |                               |                     |              |                |                |
| Linear Factor Analysis |                               |                      |              |                |                |
| 1 Factor         | .0009                         | .0308               | .0250        | .0182          |                |
| 2 Factor         | .0008                         | .0283               | .0225        | .0169          |                |
| 3 Factor         | .0007                         | .0246               | .0207        | .0160          |                |
| 4 Factor         | .0006                         | .0245               | .0196        | .0147          |                |
| DATA1            |                               |                      |              |                |                |
| Linear Factor Analysis |                               |                      |              |                |                |
| 1 Factor         | .0017                         | .0412               | .0333        | .0242          | .006           |
| 2 Factor         | .0013                         | .0359               | .0286        | .0218          | .350           |
| 3 Factor         | .0011                         | .0332               | .0262        | .0204          | .610           |
| 4 Factor         | .0009                         | .0303               | .0236        | .0191          | .860           |
| Nonlinear Factor Analysis |                               |                      |              |                |                |
| 1 Factor Quadratic |                               |                      |              |                |                |
| (\( Y_i = b_{i10} + b_{i1} \theta + b_{i2} \theta^2 + d_{1i} \)) | .0003               | .0185             | .0147        | .0113          |                |
| 1 Factor Cubic   | .0003                         | .0185               | .0147        | .0113          |                |
| (\( Y_i = b_{i10} + b_{i1} \theta + b_{i2} \theta^2 + b_{i3} \theta^3 + d_{1i} \)) |                   |                     |              |                |                |
| DATA2            |                               |                      |              |                |                |
| Linear Factor Analysis |                               |                      |              |                |                |
| 1 Factor         | .0110                         | .1049               | .0982        | .0369          | .000           |
| 2 Factor         | .0091                         | .0954               | .0896        | .0327          | .000           |
| 3 Factor         | .0070                         | .0834               | .0774        | .0310          | .000           |
| 4 Factor         | .0061                         | .0779               | .0720        | .0278          | .000           |
| Nonlinear Factor Analysis |                               |                      |              |                |                |
| 1 Factor Quadratic |                               |                      |              |                |                |
| (\( Y_i = b_{i10} + b_{i1} \theta + b_{i2} \theta^2 + d_{1i} \)) | .0003               | .0186             | .0148        | .0113          |                |
| 1 Factor Cubic   | .0003                         | .0185               | .0148        | .0113          |                |
| (\( Y_i = b_{i10} + b_{i1} \theta + b_{i2} \theta^2 + b_{i3} \theta^3 + d_{1i} \)) |                   |                     |              |                |                |
| DATA3            |                               |                      |              |                |                |
| Nonlinear Factor Analysis |                               |                      |              |                |                |
| 1 Factor Quadratic |                               |                      |              |                |                |
| (\( Y_i = b_{i10} + b_{i1} \theta + b_{i2} \theta^2 + d_{1i} \)) | .0003               | .0186             | .0147        | .0115          |                |
| 1 Factor Cubic   | .0003                         | .0175               | .0138        | .0108          |                |
| (\( Y_i = b_{i10} + b_{i1} \theta + b_{i2} \theta^2 + b_{i3} \theta^3 + d_{1i} \)) |                   |                     |              |                |                |
Table 3 continued...

**DATA4**

<table>
<thead>
<tr>
<th>Linear Factor Analysis</th>
<th>1 Factor</th>
<th>2 Factor</th>
<th>3 Factor</th>
<th>4 Factor</th>
<th>5 Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>.0203</td>
<td>.1425</td>
<td>.1108</td>
<td>.0900</td>
</tr>
<tr>
<td>1 Factor Quadratic</td>
<td></td>
<td>.0017</td>
<td>.0412</td>
<td>.0334</td>
<td>.0240</td>
</tr>
<tr>
<td>2 Factor Quadratic</td>
<td></td>
<td>.0012</td>
<td>.0346</td>
<td>.0276</td>
<td>.0212</td>
</tr>
</tbody>
</table>

Nonlinear Factor Analysis

| 1 Factor Quadratic     | .0009    | .0003    | .0174    | .0138    | .0107    |

$Y_i = b_{i0} + b_{i1} \theta + b_{i2} \theta^2 + d_i u_i$

**DATA5**

<table>
<thead>
<tr>
<th>Linear Factor Analysis</th>
<th>1 Factor</th>
<th>2 Factor</th>
<th>3 Factor</th>
<th>4 Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>.0047</td>
<td>.0686</td>
<td>.0556</td>
</tr>
<tr>
<td>1 Factor Quadratic</td>
<td></td>
<td>.0014</td>
<td>.0374</td>
<td>.0313</td>
</tr>
<tr>
<td>2 Factor Quadratic</td>
<td></td>
<td>.0012</td>
<td>.0346</td>
<td>.0289</td>
</tr>
<tr>
<td>3 Factor Quadratic</td>
<td></td>
<td>.0010</td>
<td>.0316</td>
<td>.0254</td>
</tr>
</tbody>
</table>

Nonlinear Factor Analysis

| 1 Factor Quadratic     | .0009    | .0003    | .0174    | .0138    | .0107    |

$Y_i = b_{i0} + b_{i1} \theta + b_{i2} \theta^2 + d_i u_i$

**DATA6**

<table>
<thead>
<tr>
<th>Nonlinear Factor Analysis</th>
<th>1 Factor Quadratic</th>
<th>2 Factor Quadratic</th>
<th>3 Factor Quadratic</th>
<th>4 Factor Quadratic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.0005</td>
<td>.0003</td>
<td>.0182</td>
<td>.0145</td>
</tr>
</tbody>
</table>

$Y_i = b_{i0} + b_{i1} \theta + b_{i2} \theta^2 + d_i u_i$

**DATA7**

<table>
<thead>
<tr>
<th>Nonlinear Factor Analysis</th>
<th>1 Factor Quadratic</th>
<th>2 Factor Quadratic</th>
<th>3 Factor Quadratic</th>
<th>4 Factor Quadratic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.0005</td>
<td>.0003</td>
<td>.0175</td>
<td>.0140</td>
</tr>
</tbody>
</table>

$Y_i = b_{i0} + b_{i1} \theta + b_{i2} \theta^2 + d_i u_i$

$r_{ij}$ are the residual correlations

**p**-value associated with the chi-square test of goodness of fit.
Table 4
Results of Linear and Nonlinear Factor Analysis
For Real Test data: Goodness of Fit Statistics

|                  | $\overline{r_{ij}}^2$ | SD($r_{ij}$) | $|r_{ij}|$ | SD($|r_{ij}|$) | $p <$ |
|------------------|------------------------|--------------|-----------|----------------|-------|
| **LIT**          |                        |              |           |                |       |
| **Linear Factor Analysis** |                     |              |           |                |       |
| 1 Factor         | .0034                  | .0584        | .0465     | .0354          | .000  |
| 2 Factor         | .0028                  | .0526        | .0428     | .0307          | .000  |
| 3 Factor         | .0019                  | .0439        | .0349     | .0267          | .000  |
| 4 Factor         | .0015                  | .0391        | .0310     | .0240          | .000  |
| **Nonlinear Factor Analysis** |                     |              |           |                |       |
| 1 Factor Quadratic | .0008                  | .0278        | .0216     | .0176          |       |
| $(Y_i = b_{i0} + b_{i1}^2 + b_{i2}^2 + d_{i1} u_i)$ | | | | | |
| 2 Factor Quadratic | .0004                  | .0207        | .0162     | .0130          |       |
| $(Y_i = b_{i0} + b_{i11}^2 + b_{i12}^2 + b_{i21}^2 + b_{i22}^2 + d_{i1} u_i)$ | | | | | |
| **AR**           |                        |              |           |                |       |
| **Linear Factor Analysis** |                     |              |           |                |       |
| 1 Factor         | .0047                  | .0683        | .0569     | .0378          | .000  |
| 2 Factor         | .0032                  | .0561        | .0468     | .0310          | .000  |
| 3 Factor         | .0024                  | .0499        | .0400     | .0281          | .000  |
| 4 Factor         | .0020                  | .0447        | .0362     | .0262          | .000  |
| **Nonlinear Factor Analysis** |                     |              |           |                |       |
| 1 Factor Quadratic | .0007                  | .0265        | .0200     | .0174          |       |
| $(Y_i = b_{i0} + b_{i1}^2 + b_{i2}^2 + d_{i1} u_i)$ | | | | | |
| 2 Factor Quadratic | .0004                  | .0190        | .0146     | .0122          |       |
| $(Y_i = b_{i0} + b_{i11}^2 + b_{i12}^2 + b_{i21}^2 + b_{i22}^2 + d_{i1} u_i)$ | | | | | |
| **HSTLIT1**      |                        |              |           |                |       |
| **Nonlinear Factor Analysis** |                     |              |           |                |       |
| 1 Factor Quadratic | .0008                  | .0275        | .0213     | .0175          |       |
| $(Y_i = b_{i0} + b_{i1}^2 + b_{i2}^2 + d_{i1} u_i)$ | | | | | |
| 2 Factor Quadratic | .0003                  | .0185        | .0143     | .0118          |       |
| $(Y_i = b_{i0} + b_{i11}^2 + b_{i12}^2 + b_{i21}^2 + b_{i22}^2 + b_{i23}^2 + d_{i1} u_i)$ | | | | | |
Table 4 continued...

**HSTLIT2**

Nonlinear Factor Analysis

1 Factor Quadratic

\( Y_i = b_{i0} + b_{i1} \theta + b_{i2} \theta^2 + d_{i1} u_1 \)

<table>
<thead>
<tr>
<th></th>
<th>.0006</th>
<th>.0236</th>
<th>.0181</th>
<th>.0152</th>
</tr>
</thead>
</table>

2 Factor Quadratic

\( Y_i = b_{i0} + b_{i11} \theta_1 + b_{i12} \theta_1^2 + b_{i21} \theta_2 + b_{i22} \theta_2^2 + b_{i23} \theta_2 + d_{i1} u_1 \)

<table>
<thead>
<tr>
<th></th>
<th>.0004</th>
<th>.0191</th>
<th>.0150</th>
<th>.0119</th>
</tr>
</thead>
</table>

**ARGS**

Nonlinear Factor Analysis

1 Factor Quadratic

\( Y_i = b_{i0} + b_{i1} \theta + b_{i2} \theta^2 + b_{i3} e_i \)

<table>
<thead>
<tr>
<th></th>
<th>.0021</th>
<th>.0462</th>
<th>.0268</th>
<th>.0376</th>
</tr>
</thead>
</table>

2 Factor Quadratic

\( Y_i = b_{i0} + b_{i11} \theta_1 + b_{i12} \theta_1^2 + b_{i21} \theta_2 + b_{i22} \theta_2^2 + b_{i23} \theta_2 + d_{i1} u_1 \)

<table>
<thead>
<tr>
<th></th>
<th>.0004</th>
<th>.0192</th>
<th>.0003</th>
<th>.0123</th>
</tr>
</thead>
</table>

3 Factor Quadratic

\( Y_i = b_{i0} + b_{i11} \theta_1 + b_{i12} \theta_1^2 + b_{i21} \theta_2 + b_{i22} \theta_2^2 + b_{i23} \theta_2 + b_{i31} \theta_3 + b_{i32} \theta_3^2 + b_{i33} \theta_3 + b_{i34} \theta_3 + b_{i35} \theta_3 + d_{i1} u_1 \)

<table>
<thead>
<tr>
<th></th>
<th>.0004</th>
<th>.0175</th>
<th>.0003</th>
<th>.0111</th>
</tr>
</thead>
</table>

* \( r_{ij} \) are residual correlations

** \( p \)-value associated with the chi-square test of goodness of fit.
Distribution List

Dr. Terry Ackerman
Educational Psychology
200C Education Bldg.
University of Illinois
Champaign, IL 61801

Dr. Terry Allard
Code 1142C5
Office of Naval Research
800 N. Quinby St.
Arlington, VA 22217-5000

Dr. Nancy Allen
Educational Testing Service
Princeton, NJ 08541

Dr. Gregory Antis
Office of Naval Research
Code 1142C5
Arlington, VA 22217-5000

Dr. Gregory Boodoo
American College Testing
P.O. Box 168
Iowa City, IA 52243

Dr. Susan Chipman
Cognitive Science Program
Office of Naval Research
800 North Quincy St.
Arlington, VA 22217-5000

Dr. Raymond E. Christal
U.S. Army Research Laboratory
Brooke AFB, TX 78235

Dr. Norman Cliff
Department of Psychology
UCLA, 2050 10910

Dr. Lisa Clements
Office of Naval Research
Arlington, VA 22217-5000

Commanding Officer
Naval Research Laboratory
Code 4827
Washington, DC 20375-5000

Dr. John H. Coggeshall
Department of Psychology
U.S. Army Research Laboratory
Arlington, VA 22217-5000

Dr. Linda Current
Defense Manpower Data Center
Suite 400
1600 Wilson Blvd.
Arlington, VA 22209

Dr. Charles E. Davis
Educational Testing Service
Madison, WI 53704

Dr. Ralph J. De Ayala
Measurement, Statistics, and Evaluation
Benjamin Bldg., Rm. 1229F
University of Maryland
College Park, MD 20742

Dr. Sharon Derry
Florida State University
Department of Psychology
Tallahassee, FL 32304

Dr. Richard Dineen
Graduate School of Education
University of California
Santa Barbara, CA 93106

Dr. Susan Eisenstadt
University of Kansas
Psychology Department
426 Insull
Lawrence, KS 66045

Dr. George Engelhardt, Jr.
Division of Educational Studies
Emory University
210 North Avenue NW
Atlanta, GA 30322

ERIC Facility-Aplications
2440 Research Blvd., Suite 550
Rockville, MD 20850-3238

Dr. Marshall J. Farr
Fairleigh Dickinson University
2579 North Vernon Street
Arlington, VA 22207

Dr. Leonard Feitl
Rand Dartmouth Institute for Measurement
University of Iowa
Iowa City, IA 52242

Dr. Richard F. Fieser
American College Testing
P.O. Box 648
Iowa City, IA 52243

Dr. Gerhard Fischer
Apostolova 5
A-1011 Vienna
AUSTRIA

Dr. Byron Flord
U.S. Army Headquarters
DARPA-HR
The Pentagon
Washington, DC 20310-1000

Mr. Paul Foley
Navy Personnel R&D Center
San Diego, CA 92152-6600

Chair, Department of
Computer Science
George Mason University
Fairfax, VA 22030

Dr. Robert D. Gibbons
University of Illinois at Chicago
NIP 1064, M/C 913
922 South Wood Street
Chicago, IL 60612

Dr. Janice Gifford
University of Massachusetts
School of Education
Amherst, MA 01003

Dr. Robert Glaser
Learning Research & Development Center
University of Pittsburgh
3439 O'Hara Street
Pittsburgh, PA 15213

Dr. Susan D. Goldman
Psychology Department
University of New Mexico
Albuquerque, NM 87105

Dr. Robert Glaser
Learning Research & Development Center
University of Pittsburgh
3439 O'Hara Street
Pittsburgh, PA 15213

Dr. Susan D. Goldman
Psychology Department
University of New Mexico
Albuquerque, NM 87105

Dr. Patricia A. Gooden
Department of Psychology
University of California
Santa Barbara, CA 93106

Dr. Susan Eisenstadt
University of Kansas
Psychology Department
426 Insull
Lawrence, KS 66045

Dr. George Engelhardt, Jr.
Division of Educational Studies
Emory University
210 North Avenue NW
Atlanta, GA 30322

ERIC Facility-Aplications
2440 Research Blvd., Suite 550
Rockville, MD 20850-3238

Dr. Marshall J. Farr
Fairleigh Dickinson University
2579 North Vernon Street
Arlington, VA 22207

Dr. Leonard Feitl
Rand Dartmouth Institute for Measurement
University of Iowa
Iowa City, IA 52242

Dr. Richard F. Fieser
American College Testing
P.O. Box 648
Iowa City, IA 52243

Dr. Gerhard Fischer
Apostolova 5
A-1011 Vienna
AUSTRIA

Dr. Byron Flord
U.S. Army Headquarters
DARPA-HR
The Pentagon
Washington, DC 20310-1000

Mr. Paul Foley
Navy Personnel R&D Center
San Diego, CA 92152-6600

Chair, Department of
Computer Science
George Mason University
Fairfax, VA 22030

Dr. Robert D. Gibbons
University of Illinois at Chicago
NIP 1064, M/C 913
922 South Wood Street
Chicago, IL 60612

Dr. Janice Gifford
University of Massachusetts
School of Education
Amherst, MA 01003

Dr. Robert Glaser
Learning Research & Development Center
University of Pittsburgh
3439 O'Hara Street
Pittsburgh, PA 15213

Dr. Susan D. Goldman
Psychology Department
University of New Mexico
Albuquerque, NM 87105

Dr. Patricia A. Gooden
Department of Psychology
University of California
Santa Barbara, CA 93106

Dr. Susan Eisenstadt
University of Kansas
Psychology Department
426 Insull
Lawrence, KS 66045

Dr. George Engelhardt, Jr.
Division of Educational Studies
Emory University
210 North Avenue NW
Atlanta, GA 30322

ERIC Facility-Aplications
2440 Research Blvd., Suite 550
Rockville, MD 20850-3238

Dr. Marshall J. Farr
Fairleigh Dickinson University
2579 North Vernon Street
Arlington, VA 22207

Dr. Leonard Feitl
Rand Dartmouth Institute for Measurement
University of Iowa
Iowa City, IA 52242

Dr. Richard F. Fieser
American College Testing
P.O. Box 648
Iowa City, IA 52243