DOCUMENT RESUME

ED 270 478 TM 860 359

AUTHOR Hambleton, Ronald K.; Rovinelli, Richard J.
TITLE Assessing the Dimensionality of a Set of Test Items.
PUB DATE [86]
NOTE 37p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Analysis; Computer Simulation; Correlation; *Factor Analysis; Graduate Medical Education; Higher Education; *Item Analysis; *Latent Trait Theory; *Mathematical Models; Occupational Tests; Statistical Studies
IDENTIFIERS Bejar Model; Linear Models; Nonlinear Models; Residuals (Statistics); *Unidimensionality (Tests)

ABSTRACT

Four methods for determining the dimensionality of a set of test items were compared: (1) linear factor analysis; (2) residual analysis; (3) nonlinear factor analysis; and (4) Bejar's method. Five artificial test data sets (for 40 items and 1500 examinees) were generated, consistent with the three-parameter logistic model and the assumption of either a one- or a two-dimensional latent space. Two variables were manipulated: the correlation between the traits (either .10 or .60) and the percent of test items measuring each trait (either 50 percent measuring each trait, or 75 percent measuring the first trait and 25 percent measuring the second trait). The results indicated that linear factor analysis in all instances overestimated the number of underlying dimensions in the data. Nonlinear factor analysis, with linear and quadratic terms, led to the correct determination of the item dimensionality in the three data sets where it was used. Both residual analysis and Bejar's method provided disappointing results. The results suggested the need for extreme caution in using linear factor analysis, residual analysis, and the Bejar method, until further investigations confirm their adequacy. Nonlinear factor analysis appeared to be the most promising of the four methods. (Author/GDC)

Assessing the Dimensionality of a Set of Test Items

Ronald K. Hambleton
University of Massachusetts at Amherst

and

Richard J. Rovinelli
Educational Services for the Professions

Abstract

The main purpose of the study was to compare the determination of the dimensionality of a set of test items with four methods: linear factor analysis, non-linear factor analysis, residual analysis, and a method developed by Bejar. Five artificial test datasets (for 40 items and 1500 examinees) were generated to be consistent with the three-parameter logistic model and the assumption of either a one- or a two-dimensional latent space. Two variables were manipulated: the correlation between the traits (r = .10 or r = .60) and the percent of test items measuring each trait (50% measuring each trait, or 75% measuring the first trait and 25% measuring the second trait).

The results were that linear factor analysis in all instances overestimated the number of underlying dimensions in the data. Non-linear factor analysis, on the other hand, with linear and quadratic terms led to the correct determination of the item dimensionality in the three datasets where it was used. Both the residual analysis method and Bejar's method provided disappointing results. The results suggest the need for extreme caution in using linear factor analysis, residual analysis, and the method by Bejar until more investigations of these methods can confirm their adequacy. Non-linear factor analysis appears to be the most promising of the four methods, but more experience in applying the method seems to be needed before it can be recommended for wide-scale use.

Assessing the Dimensionality of a Set of Test Items[1]

Ronald K. Hambleton[2]

University of Massachusetts, Amherst

and

Richard J. Rovinelli

Educational Services for the Professions

The assumption that a set of test items is "unidimensional" is made for all of the presently popular item response models. Despite the importance of the assumption to these item response models there is substantial confusion in the psychometric literature concerning the proper definition for the term "unidimensionality" and the methods for assessing its presence or absence in a set of test items (Hattie, 1984; Traub & Wolfe, 1981). Definitions in the literature for what it means to say that a set of test items is unidimensional are typically abstract and non-operational. A typical example is: A set of test items is unidimensional when a single ability can explain or account for examinee test performance.

Methods for assessing the unidimensionality of a set of test items range from the commonly used but unacceptable measures of internal consistency (Green, Lissitz, & Mulaik, 1977), to the more appropriate uses of eigenvalue plots and related statistics (Lord & Novick, 1968; Reckase, 1979). Hattie (1984) reported that there are 87 indices in the psychometric literature for addressing the dimensionality of a set

of test items. Unfortunately, many methods for assessing the unidimensionality of a set of test items are only loosely connected to the various definitions in the psychometric literature.

This investigation of several methods for assessing the unidimensionality of a set of test items was prompted by a practical problem which arose in connection with the in-training exam produced by the American Board of Family Practice in Lexington, Kentucky. Candidates are required to take a core exam plus three additional subtests of their choice selected from a larger set of six available subtests. These six subtests vary somewhat in their difficulty. Score reporting and subsequent comparisons among candidates must be carried out on the combined exam (core plus subtests) and since the candidates do not, in general, take the same three subtests (there are 20 possible combinations), the subtests are equated to the common scale defined by the core items. How well the equating of candidate scores is done depends upon the choice of item response model (the three-parameter model is chosen to increase the likelihood of a satisfactory model fit to the data) and the extent to which the core items and subtest items measure a common trait, i.e., a unidimensional trait. Thus, two questions arose, "What is meant by the expression 'unidimensionality of a set of test items'?" and "How should the assumption of unidimensionality be assessed to be consistent with the definition?"

Despite the confusion in the literature on these two questions there are contributions by McDonald (1980, 1982) and Hattie (1981) which proved to be helpful, and these contributions influenced the general direction for the present research. McDonald and Hattie arrived at the conclusion that the principle of local independence should be the basis for a proper definition for the assumption of unidimensionality. McDonald defined a set of test items as unidimensional if, for examinees with the same ability, the covariation between items in the set is zero. Since the relations between items is typically non-linear, he recommended the use of non-linear factor analysis to study these relations between items. Also, after fitting a single non-linear factor model to the item set, he recommended that residual covariances be calculated and used to assess the plausibility of the unidimensionality assumption. McDonald argued that the dimensionality of a set of test items should be determined by the number of factors or abilities needed for describing examinees, so that the principle of local independence is satisfied.

In this research, interest was centered on three promising methods for addressing the unidimensionality of a set of test items: (1) non-linear factor analysis (NLFA) because of McDonald's recommendation, (2) residual analysis, and (3) the Bejar analysis. The first method appeared promising because NLFA does not require the implausible assumption of linear relationships among the variables and between the variables and the underlying traits to be made. In fact, one of the

fundamental assumptions of IRT is that these relationships are non-linear (Lord, 1980). The second method is an assessment of the overall fit of a unidimensional model to a dataset through the analysis of residuals. When the fit is adequate it would seem that the assumption of a unidimensional model is plausible too (see, for example, Rentz & Rentz, 1979). Of course when the fit is poor, specific reasons for the misfit may be unknown. For example, the assumption of unidimensionality may not be violated by the data set but some other assumption of the model is. The Bejar (1980) method appeared useful for assessing item dimensionality because it does not involve questionable linearity assumptions about the test data. Also, the method provides a straightforward check on one of the expected outcomes of a unidimensional set of test data: the subset of items from a test in which an item is calibrated is irrelevant.

The specific purpose of the investigation was to compare the assessments of the dimensionality of a set of test items with three methods, referred to in this paper, as non-linear factor analysis, residual analysis, and Bejar analysis. To provide a basis for comparing the merits of the methods, linear factor analysis was also studied on the same datasets. The four methods were applied to five datasets. The datasets were artificial and generated to reflect one and two dimensional datasets.

Method

Methods for Assessing Item Dimensionality

Linear Factor Analysis (LFA)

LFA is probably the most commonly used method for studying item dimensionality. Using (1) the matrix of phi or tetrachoric correlations to summarize the linear relationships between pairs of items in a test,[3] and (2) communality estimates (often, squared multiple correlations) in the diagonal entries of the correlation matrix, eigenvalues are extracted from the correlation matrix and plotted (from largest to smallest). The number of "significant" factors is determined by looking for the "elbow" in the plot. The number of eigenvalues to the left of the "elbow" is normally taken to be the number of "significant" factors underlying test performance. The method fails in those instances where an "elbow" cannot be found.

A second procedure for determining the number of factors is found by applying an identical LFA to a dataset consisting of the same number of items and examinees and with random normal deviates substituted for the actual data. This procedure appears to have been first suggested by Horn (1965) and was later studied by Linn (1968), Humphreys and Ilgen (1969), and Humphreys and Montanelli (1975). For the purpose of computing phi and tetrachoric correlations, the data are dichotomized at $z=0$,[4] normal deviates above 0 are coded as "correct" and below 0 as "incorrect." Since all of the data are randomly generated, the value of

the largest eigenvalue can be used as a cut-off score with the eigenvalues obtained from the actual data to determine the number of "significant factors." Thus, if the largest eigenvalue obtained from the random data is 1.5, all eigenvalues over 1.5 in the analysis of the real data are considered to be associated with significant factors.

Non-Linear Factor Analysis (NLFA)

McDonald (1967) sought to improve upon LFA by developing non-linear factor analysis (NLFA). In NLFA, non-linear relationships between the variables and the traits or factors measured by the variables are assumed. The application of NLFA to the study of item dimensionality seems especially desirable, within the context of item response theory, because one of the principal assumptions (i.e., the mathematical form of the item characteristic curves) specifies a particular non-linear relationship between item performance and ability. One version of NLFA takes the form

$$y_i = a_{i0} + \sum_{\ell=1}^{t} \sum_{p=1}^{s} a_{i\ell p}\theta_\ell^p \qquad (i=1, 2, \ldots, n)$$

where y_i represents an examinee's score on item i, t is the number of traits necessary to account for examinee test performance, s is the degree of the polynomial used to fit the model with each factor, $a_{i\ell p}$ is the factor loading of the i^{th} item on the ℓ^{th} trait for the p^{th} degree element in the polynomial. For example, with a one factor model (t=1) with polynomial terms to the third power (s=3) the model has the following form:

$$y_i = a_{i0} + a_{i11}\,\theta_1 + a_{i12}\,\theta_1^2 + a_{i13}\,\theta_1^3 \qquad (i=1, 2, \ldots, n)$$

The two factor model (t=2) with polynomial terms to the third power (s=3) has the form

$$y_i = a_{i0} + a_{i11}\,\theta_1 + a_{i12}\,\theta_1^2 + a_{i13}\,\theta_1^3$$

$$+ a_{i21}\,\theta_2 + a_{i22}\,\theta_2^2 + a_{i23}\,\theta_2^3 \qquad (i=1, 2, \ldots, n)$$

The one- and two-factor models above with linear, with linear and quadratic, and with linear, quadratic, and cubic terms were fitted to the various datasets in the study.

Residual Analysis

The method for addressing the unidimensionality of a set of test items through a residual analysis involves fitting a unidimensional item response model of interest to the test data, using the model parameter estimates to predict the item performance data, and then summarizing the discrepancies or residuals (see, for example, Hambleton & Swaminathan, 1985). Specifically, ability categories are chosen to divide the ability scale into equal intervals. Examinees are assigned to categories based upon their ability estimates. For examinees in each ability category on each item, a comparison is made between actual performance (proportion-correct) and the predicted proportion-correct level from the corresponding item characteristic curve (icc). In this study the proportion-correct estimate was obtained at the mid-point of each ability category. (A slightly better estimate is the average of the probabilities for a correct answer associated with the ability scores for examinees in the category.) The difference between the actual and predicted proportion-correct score (called a residual or a

raw residual score) in each ability category and for each item can also be divided by the corresponding standard error of the proportion-correct estimate to obtain a standardized residual. When the chosen model fits the dataset, these standardized residuals might be expected to be small and randomly distributed about the value 0. It is common within the framework of regression theory to assume the distribution of standardized residuals is normal. Of course, the distribution of residuals would only be (at best) approximately normal because of non-normal distributions of the SRs when the icc's approach values of 0 or 1.

The rationale for the appropriateness of residuals as a check on item unidimensionality is that when a unidimensional model fits a dataset, all of the model assumptions must be met to a reasonable degree.

Bejar Analysis

Bejar (1980) argued that if the set of items in a test is unidimensional, then the grouping of test items from the test for the purpose of item calibration will be irrelevant. Parameter estimates for items calibrated with different subsets of items, aside from sampling errors, should be identical. Bejar's method (with minor modifications) can be implemented in four steps:

1. Identify a subset of items in the test which appears to be measuring a trait different from the trait measured by the total test.
2. Conduct a three-parameter model analysis of only the items in the subtest.

3. Repeat the three-parameter model analysis using the total set of items.
4. Compare the two sets of b-value estimates for items in the subtest. (Because b-values are estimated with smaller sampling errors than the a-values, b-values are more useful for studying the relationships between item parameter estimates obtained in two samples.) Bejar has a simple statistical test that can be used to compare the parameter estimates. Alternately, the pairs of b-values can be plotted to determine the extent to which the two sets are linearly related.

The pairs of parameter estimates for items in the subtest and test, respectively, should be linearly related unless the subset of items is measuring a trait or traits which are not common to the trait or traits measured in the total test.

Criteria for Assessing the Usefulness of the Methods

Linear Factor Analysis

One way to evaluate LFA as a measure of item dimensionality is to compare the number of factors retained in a solution to the dimensionality of the latent space in the artificial data. Two ways for determining the number of factors were used:

1. the "elbow" in the plot of eigenvalues,
2. eigenvalues greater than the largest eigenvalues obtained with the random data.

Also, to facilitate a comparison between LFA and NLFA, four additional criteria were used. After fitting one, two, three, four, and five factors to the reduced correlation matrix, the matrix of residuals was calculated. The off-diagonal elements of the matrix were summarized by (1) the average residual, (2) the standard deviation of the residuals, (3) the average of the absolute-valued residuals, and (4) the standard deviation of the absolute-valued residuals. When the chosen model fits the dataset well, the true values of these off-diagonal elements are (near) zero. In practice, because of errors of measurement and of sampling, the residuals should be small, and evenly distributed around 0.0.

Non-Linear Factor Analysis

The four statistics described above were calculated after fitting one and two factor models with linear, quadratic, and cubic terms to the inter-item correlation matrices for four of the five datasets.

Residual Analysis

For each artificial dataset there were 40 items and 12 ability categories. In total, 480 residuals were produced. Since interest centered on the size and not the direction of the discrepancies, absolute-valued residuals and standardized residuals were substituted for residuals and standardized residuals. Criterion measures chosen were (1) average absolute-valued residual, (2) average absolute-valued standardized residual, and (3) the distribution of absolute-valued standardized residuals. With respect to (2), when a model fits the

data, and if the SRs are normally distributed, then the average absolute-valued SR should be close to .799. (The value of .799 can be obtained by calculating the average of absolute values of normal deviates.) Since a study of absolute-valued residuals is more informative when several models have been been fit to the same test data, in this phase of the work, residual analyses were carried out with the one-, two-, and three-parameter unidimensional logistic models. A computer program developed by Linda Murray and described by Hambleton (1982) was used to conduct the residual analyses.

A comparable analysis of residuals with NLFA would have been useful, but with the polynomial models the expected probabilities are not defined on the interval [0,1] across the total ability score range and therefore an analysis of residuals would have no value.

Bejar Analysis

The results from this analysis were summarized by a correlation coefficient between b-values for a subset of test items: items calibrated in a subtest and again in the total test. The original plan was to produce the plot of item b-values obtained in the sub-test and total test for each of the five datasets. But the correlations were very high in all but one analysis and so the plots did not seem necessary.

Generation of Artificial Data

The generation of examinee item response data to fit a three-parameter logistic unidimensional model is straightforward. Using item parameters which are generated according to the specifications described in the next section, for each examinee j, and given the ability level, θ_j, a vector of probabilities (P_1, P_2, ..., P_n) associated with answering the test items correctly is obtained from the expression

$$P_i(\theta_j) = c_i + (1-c_i)\,[1 + \exp(-Da_i(\theta_j - b_i))]^{-1},\ i=1, 2, \ldots, n\ .$$

Using a random number generator to produce numbers uniformly distributed on the interval [0, 1], the probabilities can be converted to 0's and 1's to reflect examinee item scores. This is done by assigning a "1" to the examinee for item i when the random number selected is below P_i, which will happen P_i of the time, and "0" otherwise.

The simulation of two-dimensional data was a substantially more difficult problem. First, there are several possible multi-dimensional models to select from. Second, there are no guidelines in the IRT literature for choosing reasonable item parameter values for multidimensional models.

Sympson (1978) offered one model which took the form

$$P_i(\theta_{j1}, \theta_{j2}, \ldots, \theta_{jk}) = c_1 + (1-c_i) \prod_{k=1}^{K} [1+\exp(-Da_{ik}(\theta_{jk}-b_{ik}))]^{-1}\ .$$

The problem encountered in applying this model was that the probabilities even for the two-dimensional model (k=2), quickly converged to the value c_i and as a consequence there was little variation in item performance for examinees, and little variation in test scores among examinees.

A similar problem was encountered in applying a model used by Christofferson (1975) and Hattie (1981). In Hattie's formulation

$$P_i(\theta_{j1},\theta_{j2},\ldots,\theta_{jk}) = c + (1-c)\,[1+\exp -D\sum_{k=1}^{K} a_{ik}\theta_{jk} - b_i]^{-1}$$

where c is a constant value over all items, b_i is the item difficulty, a_{ik}, k=1, 2, ..., K are the item discriminating powers on the K underlying traits and θ_{jk}, k=1, 2, ..., K are the trait or ability scores for examinee j on the K traits. In preliminary simulations it was observed that the probabilities for small changes in θ quickly approached values of c and 1. Perhaps the main problem encountered concerned the choice of item discriminating powers which were taken to be values often observed with one-dimensional models. In any case, a simpler model than the models proposed by Sympson and Hattie was chosen. The model was not only simpler (a special case of the Christofferson-Hattie model with independent item clusters) but guidelines for selecting item parameters were readily available. Specifically, item parameter values were assigned to all test items in the same way that the assignments were made with the one-dimensional model described in the next section. Then, examinees were assigned two trait scores (with the specified correlation). The simulation of two traits was easily carried out with the aid of a formula developed by Hoffman (1959). First, two uncorrelated normally distributed random variables X_j, Z_j(Mean=0, SD=1) are generated with the aid of a random number generator. Then the variable Y_j is obtained from the expression

$$Y_j = X_j + \frac{k}{r} Z_j$$

where $k = \sqrt{1-r^2}$

and r is the desired correlation between variables X_j and Y_j. X_j and Y_j are used in the simulations as the two trait scores for an examinee. Pairs of trait scores with the desired correlation were generated for the 1500 examinees. Finally, item probabilities and item scores were generated for examinees. For the first 20 items (or 30 items for datasets 3 and 5) the first set of trait scores were used in generating item probabilities and for the remaining items, the second set of trait scores for examinees were used.

In summary, first, artificial data generated from a one-dimensional model were used. Of interest was whether or not the four methods could identify unidimensional data. Second, two-dimensional data representing the situations where the items could be organized into two clusters measuring different traits were used. Two variables were manipulated: the correlation between the traits and the percentage of total items in each cluster.

Description of the Test Data

The five artificial test datasets were generated to be consistent with the assumption of either a one- or a two-dimensional latent space. Each test consisted of 40 test items. The item performance for 1500 examinees was simulated with the three-parameter logistic model. These

numbers were assumed to be large enough to avoid parameter estimation problems when using LOGIST. In dataset 1, the latent space was assumed to be one-dimensional. In datasets 2 to 5, the latent space was chosen to be two-dimensional. The only difference between datasets 2 and 3, and 4 and 5 was that in datasets 2 and 3 the correlation between the two latent traits was .10 whereas in datasets 4 and 5 the correlation between the two traits was .60. In addition, items were assumed to measure one trait or the other. In datasets 2 and 4, the first 20 items measured trait one and the second 20 items measured trait two. In datasets 3 and 5, the first 30 items measured trait one and the remaining 10 items measured trait two. The chart below summarizes the pertinent information:

			Number of Items	
Dataset	Trait(s)	$r(\theta_1, \theta_2)$	First Trait	Second Trait
1	1	--	40	0
2	2	.10	20	20
3	2	.10	30	10
4	2	.60	20	20
5	2	.60	30	10

Parameter values were assigned to items on each trait in the following way:

b parameters were drawn at random from a uniform distribution on the interval [-2.0 to +2.0]

a parameters were drawn at random from a uniform distribution on the interval [.40, 2.00]

c parameters were set to a value of .25.

The choice of item parameters reflected values often found in practice (Hambleton & Swaminathan, 1985).

Results

One-Dimensional Data

Linear Factor Analyses

The eigenvalues for the random data and the one-dimensional data are reported in Table 1. The plots of eigenvalues for the one-dimensional data using phi correlations to measure the relationships between pairs of items (or tetrachoric correlations) suggest that two significant factors are present. The analysis of the tetrachoric correlations was more revealing than the phi correlations in the sense that more of the variance was associated with the first factor which would be expected when the data are unidimensional. Whether the criterion for the number of significant factors is determined from the "elbows" in the plots or the largest eigenvalue from the matrix of correlations of normal random deviates, two factors would be retained.

Non-Linear Factor Analyses

The results of fitting from one to five linear factors, and one and two factors with linear, linear and quadratic, and linear, quadratic, and cubic terms to the one-dimensional dataset are reported in Table 2. The first two criteria ($\bar{r}_{ij}$, $s(r_{ij})$) show simply that the mean off-diagonal elements after fitting one or more factors are centered close to .00 (as compared to .127 in the original correlation matrix) and that the standard deviation of the distribution of the

residuals approaches zero as the number of factors is increased. From the statistics in the third and fourth columns of Table 2 it is clear that a NLFA with one factor with linear and quadratic terms fits the data better than the two factor solution provided by LFA. In fact, even three linear factors did not produce as accurate a fit to the data.

Residual Analyses

The residual analyses for the one-dimensional data with the three logistic models are reported in Table 3. Not surprisingly, since the data were generated to fit the three- parameter model, this model provided the best fit to the data. More importantly, the distribution of SRs was (approximately) normal and the mean absolute-valued SR was close to .799. With the one dimensional data and when the particular IRT model closely fits the data, the SRs appear to have the desired distribution. The slight predicted bias in this distribution is also apparent.

Bejar Analyses

Since all of the test items were generated to fit a one-dimensional model there was no reason to suspect that a second trait was necessary to account for the inter-item correlations. As a rather simple check on the method, the last 20 items were presumed to measure a second trait. The b-values for the last 20 items calibrated both in the sub-test composed of the last 20 items and the total test are reported in Table 4. The correlation between the b-values was in

excess of .99. Clearly, the assumption of unidimensionality could not be rejected on the basis of the available evidence, nor should the assumption be rejected for this dataset.

Two-Dimensional Data

Linear Factor Analyses

If the largest eigenvalue of the random data (λ_1=1.48) is used as the criterion for determining the number of factors, for all four two-dimensional datasets three significant factors emerged (see Table 2). If the "elbow" of each eigenvalue plot is used, possibly two significant factors might emerge for the two-dimensional data (r=.10; 20/20) but choosing three or four factors is not totally unreasonable. With the other two-dimensional datasets, the "elbow" revealed (at least) three significant factors also. Again, the linear factor analysis method resulted in more factors than the underlying dimensionality of the data. However, the LFA did reveal the dominance of one factor over the other. With the two- dimensional data (r=.10; 20/20) the first two factors accounted for roughly the same amount of variance. With the two-dimensional data (r=.10; 30/10) the ratio of variance accounted for by the first two factors (15.7/5.2 or roughly 3/1) was proportional to the number of items measuring each factor (30/10 or 3/1). For the final two datasets a second order factor appeared to be emerging.

Non-Linear Factor Analyses

Again, Table 2 shows that the NLFA method provided promising results. With the two-dimensional data (r=.10; 20/20) and the two-

dimensional data (r=.10; 30/10) the mean and standard deviation of absolute-valued residuals associated with a two-factor model with quadratic terms were smaller than the corresponding residuals obtained from a three-factor solution using LFA. Thus, if the three factor solution with LFA is acceptable, then the two factor solution from NLFA would be, too. The two factor model with cubic terms was not obtained because of the high costs associated with running the computer program and the acceptability of the two factor solution with quadratic terms solution.

Residual Analyses

Table 3 provides a summary of the absolute-valued residuals and standardized residuals obtained from fitting logistic models to the four two-dimensional datasets. Several findings are evident:

1. The one-parameter model did not fit any of the datasets. However, rather than suggesting multidimensionality in the data, the likely explanation in view of the results of fitting the one-parameter model to the one-dimensional data (and point two below) is that the misfit is due to the failure of the model to account for variations in item discrimination power and the guessing behavior of low-ability examinees.

2. A comparison of the SRs from the two- and three-parameter models showed substantially smaller SRs than those obtained with the one-parameter model, and the three-parameter model fitting the datasets slightly better than the two-parameter model. In fact, on the basis of a study of the SRs for the two- and three-parameter models, and assuming of course the validity of the residual analysis method, a researcher would accept the hypothesis that the test items in each dataset were unidimensional.
3. There was also evidence that the overall fits were better when the traits were correlated (r=.60), than when the traits were not (r=.10). The average absolute-valued standardized residual dropped from .84 and .76 (in datasets 2 and 3, with r=.10) to .79 and .73 (in datasets 4 and 5, with r=.60), respectively.

How could the three-parameter model fit the four two-dimensional datasets? The failure to identify multidimensionality in datasets 4 and 5 was surprising but in view of the moderately high correlation between the two traits the results were not totally unexpected although larger SRs had been predicted. Apparently LOGIST simply proceeds to estimate the second order factor which incorporates the two related factors. Why multidimensionality could not be detected in datasets 2 and 3 is not completely clear. It appears that LOGIST estimates an average ability of the two unrelated traits and also attaches low a-values to all of the test items. In doing so, a reasonable fit

between the model and each dataset can be achieved. When there is an imbalance in the test data (i.e., 30/10), LOGIST assigns high a-values to items measuring the "dominant trait" and relatively low values to the remaining items. In this way, a one-dimensional model can fit the data. With a more even split (i.e., 20/20), the values assigned to the a-values are relatively low.

In any case, because of the way LOGIST handles multidimensionality in the test data, residual analyses cannot identify it when it is present.

Bejar Results

The results of the Bejar analyses on the four two-dimensional datasets are reported in Table 4. For the purposes of these analyses the last 20 items in datasets 2 and 4 and the last 10 items in datasets 3 and 5 constituted the sub-tests. These, of course, were the test items in the datasets that measured the second traits. The following observations were noted:

1. With r=.10, and a split of 20/20, the test items had comparable b-values.
2. With r=.10, and a split of 30/10, the b-values were substantially different and appeared to be poorly estimated. This analysis would lead to a rejection of the unidimensionality assumption.

3. When r=.60, and for the two splits 20/20, and 30/10, the Bejar analyses suggested that the assumption of unidimensionality could not be rejected.

In only one of the four analyses was the Bejar method sensitive to the multidimensionality in the data. This result also was a surprise because the method appeared to have been successful in at least one other study (Bejar, 1980).

Real Data

Though the results are not reported here, the four methods for assessing item dimensionality were also applied to the 80 item section of the 1982 ABFP In-Training Exam. The four methods provided different answers to the question of unidimensionality. Had the simulation studies described earlier not been carried out the results from the residual analyses or the Bejar analyses would have been used to support the assumption of unidimensionality. The LFA of the data suggested that anywhere from 4 to as many as 8 significant factors would need to be retained for a satisfactory accounting of the data. The NLFA also appears to indicate that more than one factor may be needed. If for example, the ratio of the average absolute residual after fitting a one factor model with a cubic term, to the average correlation in the initial matrix is used as a criterion, the ratio is .018 to .042 or .438 whereas with clearly one-dimensional data the ratio was .017 to .127 or .134. It would seem that more factors are needed to obtain a satisfactory solution.

In summary, the four methods provided contradictory information about the item dimensionality. Based upon the results from the simulations, it would seem that the most likely conclusion is that more than one dimension is operating.

Discussion

On the basis of a single simulation study with limited scope, generalizability of the findings is obviously limited. But several findings of the study do appear to suggest directions for some future work. First, the linear factor analysis model in all instances overestimated the number of underlying dimensions in the data. Of course this result along with the result that the tetrachoric correlations are more useful than phi correlations in addressing item dimensionality are well-known. Second, non-linear factor analysis with linear and quadratic terms led to the correct determination of the item dimensionality in the three datasets where it was used. In subsequent work the NLFA with two traits with a correlation of .60 will be used to see if it can detect the multidimensionality. Two problems however emerged in our work with NLFA. First, the appropriate number of factors and polynomial terms to retain in a solution was determined by comparing the size of the residuals to those obtained from a satisfactory linear factor analysis. When working with real data another criterion will be needed to determine the adequacy of model fit. McDonald (personal communication) has indicated that the standard error of a binary covariance gives a good rule of thumb for the expected sizes of the residual covariances.

Both the residual analysis method and Bejar's method provided disappointing results. The residual analysis method using the one parameter model is certainly of limited value in addressing item dimensionality because large residuals may be due to the violations of several model assumptions including unidimensionality. This problem can be reduced somewhat by fitting (say) a three-parameter logistic model to the dataset. But even in the two-dimensional case where the traits were nearly orthogonal, the residual analysis method with the three-parameter model could not detect the violation in the unidimensionality assumption. It appears that the three-parameter model can accommodate multidimensionality by assigning low "a" values to these "deviant" items. Good fit is achieved, but in doing so, the "deviant" items are essentially removed from the test since those items neither contribute much to ability parameter estimation or to the test information function.

Likewise, the Bejar method was unable to detect the two underlying traits except when the correlation between the traits was low and a disproportionate number of the test items measured one of the traits. A highly plausible explanation for the failure of the Bejar method with the datasets is as follows: While it is certainly true in datasets 2 and 3 that the traits have a low correlation (r=.10), there was a very high correlation between the trait scores used in calibrating test items in the subtest and the total test. In fact if θ_1 is assumed to measure the first trait and θ_2 the second trait, then the trait measured by the combined set of test items might be closely

approximated by $w_1\theta_1+w_2\theta_2$ where w_1 and w_2 represent the proportion of items in the test measuring trait 1 and 2, respectively. When the subtest of items is calibrated, θ_2, is operating, whereas when the subset of items are calibrated in the total test $w_1\theta_1+w_2\theta_2$ is the underlying trait being measured and not θ_1. If ability scores on θ_1 and θ_2 are standardized (as they were in the simulations), and if $w_1=w_2=.50$, the correlation between the trait measured by the subtest (denoted θ_2) and the total test (approximated by $.5\theta_1 + .5\theta_2$) is at least .70. (And when the correlation between the traits is .60, the expected correlation between θ_2 and $.5\theta_1 + .5\theta_2$ is over .88!) Perhaps then it is not so surprising that the test items had highly similar difficulty estimates when calibrated in the sub-test and the total test. In the only application where the Bejar method indicated a violation of the undimensionality assumption, the correlation between θ_2 and $w_1\theta_1+w_2\theta_2$ was relatively low because $w_1=.75$ and $w_2=.25$. Why was a similar result not obtained in dataset 5? The same trend was not observed in dataset 5 because while the weighting of test items to the two traits was disproportionate (3 to 1), the traits were moderately correlated (r=.60).

The problem identified in this study with the Bejar method can be resolved by plotting ability estimates for examinees obtained in two independent samples -- one from the subtest and the other from the remaining test items. While the b-value plots, in theory, are preferable because of their higher precision, there is not an obvious way to remove the high overlap in the ability estimates apart from

choosing relatively short subtests in relation to the total test. But this solution has problems too because only relatively unstable ability estimates will be available for item calibration.

In conclusion, despite the limited scope of the present investigation, the results do suggest the need for extreme caution in using linear factor analysis, residual analysis, or Bejar's method to address questions about item unidimensionality. Clearly, more investigations of these methods showing some positive results are needed before they can be strongly recommended for use by practitioners. On the other hand, while non-linear factor analysis produced the most promising results in this study, an accepted criterion for determining the minimum number of factors to retain in a non-linear factor solution is not available, nor is an easy-to-use non-linear factor analysis program available. More research along these lines must be carried out first before NLFA can be recommended. Also, in our subsequent work, attention will be focused on the use of non-linear factor analysis in studying item dimensionality. Our plan is to determine the usefulness of NLFA with multiple traits and various correlational structures.

References

Bejar, I. I. (1980). A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. Journal of Educational Measurement, 17, 283-296.

Christofferson, A. (1975). Factor analysis of dichotomized variables. Psychometrika, 40, 5-22.

Green, S. B., Lissitz, R. W., & Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. Educational and Psychological Measurement, 37, 827-838.

Hambleton, R. K. (1982). Applications of item response models to NAEP mathematics exercise results. Final Report - ECS Contract No. 02-81-20319. Denver, CO: Educational Commission of the States.

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

Hattie, J. A. (1981). Decision criteria for determining unidimensionality. Unpublished doctoral dissertation, University of Toronto.

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

Hoffman, P. J. (1959). Generating variables with arbitrary properties. Psychometrika, 24, 265-267.

Horn, J. L. (1965). A rationale and technique for estimating the number of factors in factor analysis. Psychometrika, 30, 179-185.

Humphreys, L. G., & Ilgen, D. R. (1969). Note on a criterion for the number of common factors. Educational and Psychological Measurement, 29, 571-578.

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

Linn, R. L. (1968). A Monte Carlo approach to the number of factors problem. Psychometrika, 33, 37-72.

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

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

McDonald, R. P. (1967). Non-linear factor analysis. Psychometric Monograph, No. 15.

McDonald, R. P. (1980). The dimensionality of tests and items. British Journal of Mathematical and Statistical Psychology, 34, 100-117.

McDonald, R. P. (1982). Linear versus nonlinear models in item response theory. Applied Psychological Measurement, 6, 379-396.

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

Rentz, R. R., & Rentz, C.C. (1979). Does the Rasch model really work? NCME Measurement in Education, 10, 1-11.

Sympson, J. B. (1978). Estimation of latent trait status in adaptive testing procedures. In D.J. Weiss (Ed.), Proceedings of the 1977 Computerized Adaptive Testing Conference. Minneapolis, MN: University of Minnesota.

Traub, R. E., & Wolfe, R. G. (1981). Latent trait theories and the assessment of educational achievement. In D.C. Berliner (Ed.), Review of Research in Education - Volume 9. Washington: American Educational Research Association.

Footnotes

[1] Funding for this research study was provided by the American Board of Family Practice, Lexington, Kentucky. We are grateful to Jamshid Etezadi for providing us with a copy of NOFA, a computer program for conducting non-linear factor analyses.

[2] The authors are grateful to I. Bejar, R. McDonald, F. Lord, Jane Rogers, R. Traub, and two anonymous reviewers for comments on an earlier draft of the paper.

[3] Usually the choice of phi correlations or tetrachoric correlations is not a major consideration in conducting factor analysis studies. But, McDonald (personal communication) notes that as Lord originally showed, if the normal ogive model fits the data, and if ability scores are normally distributed, the matrix of tetrachoric correlations should fit a Spearman model. This is therefore an approximate method for fitting the normal ogive model.

[4] In several pilot runs, the z scores were dichotomized to simulate the actual difficulty levels of the test items. However, the effect on the results was minimal.

Table 1
Eigenvalues (λ) and Percent of Variance Accounted for in Random and One-Dimensional Datasets Using Phi and Tetrachoric Correlations
(40 items; 1500 examinees)

	Random Data[1]		One-Dimensional Data			
	Tetrachoric		Phi		Tetrachoric	
Factor	λ	%	λ	%	λ	%
1	1.48	3.7	8.86	22.2	15.00	37.5
2	1.44	3.6	2.09	5.2	2.21	5.5
3	1.37	3.4	1.11	2.8	1.15	2.9
4	1.34	3.3	1.05	2.6	1.10	2.7
5	1.32	3.3	1.03	2.6	1.08	2.7
6	1.30		1.02		1.02	
7	1.28		.98		.96	
8	1.25		.96		.95	
9	1.22		.95		.92	
10	1.21		.93		.90	
11	1.19		.93		.88	
12	1.18		.93		.84	
13	1.15		.91		.80	
14	1.13		.89		.77	
15	1.10		.86		.74	

[1] Squared multiple correlations used as communality estimates.

Table 2
Residual Matrices After Fitting Linear
and Non-Linear Factor Analysis Models

Dataset	Model	λ_1	% Var.	Goodness of Fit $\bar{r}_{ij}$	$s(r_{ij})$	$\overline{\lvert r_{ij} \rvert}$	$s(\lvert r_{ij} \rvert)$
1-DIM	Correlation Matrix			.127	.079	.127	.079
	Factor Analysis						
	1 Factor	6.64	16.6	.006	.078	.060	.050
	2 Factors	1.84	4.6	-.002	.030	.022	.021
	3 Factors	1.13	2.8	-.003	.025	.019	.016
	4 Factors	1.11	2.8	.000	.021	.016	.013
	5 Factors	1.10	2.7	.000	.019	.015	.012
	Non-Linear Factor Analysis						
	1 Factor, Linear Term			.002	.033	.026	.021
	1 Factor, Quad Term			.001	.022	.017	.014
	1 Factor, Cubic Term			.000	.022	.017	.014
	2 Factors, Linear Terms			-.006	.030	.022	.020
	2 Factors, Quad Terms			.000	.020	.015	.012
2-DIM	Correlation Matrix			.075	.090	.081	.084
(r=10;	Factor Analysis						
20/20)	1 Factor	4.41	11.0	.016	.074	.054	.054
	2 Factors	3.59	9.0	.000	.033	.024	.022
	3 Factors	1.64	4.1	.000	.025	.019	.016
	4 Factors	1.41	3.5	.000	.020	.016	.012
	5 Factors	1.15	2.9	.000	.018	.014	.011
	Non-Linear Factor Analysis						
	1 Factor, Linear Term			.025	.072	.050	.057
	1 Factor, Quad Term			.011	.037	.028	.027
	1 Factor, Cubic Term			.007	.029	.022	.020
	2 Factors, Linear Terms			-.005	.039	.027	.028
	2 Factors, Quad Terms			.000	.020	.016	.012

-continued on next page-

Table 2, continued

Dataset	Model	λ_1	% Var.	Goodness of Fit $\bar{r}_{ij}$	$s(r_{ij})$	$\overline{\lvert r_{ij} \rvert}$	$s(\lvert r_{ij} \rvert)$
2-DIM (r=.10; 30/10)	Correlation Matrix			.104	.100	.109	.095
	Factor Analysis						
	1 Factor	6.27	15.7	.004	.047	.033	.034
	2 Factors	2.10	5.3	.002	.036	.029	.021
	3 Factors	1.88	4.7	.000	.023	.017	.015
	4 Factors	1.28	3.2	.000	.020	.016	.012
	5 Factors	1.09	2.7	.000	.018	.015	.011
	Non-Linear Factor Analysis						
	1 Factor, Linear Term			.008	.046	.033	.034
	1 Factor, Quad Term			.007	.036	.032	.029
	1 Factor, Cubic Term			.005	.039	.027	.028
	2 Factors, Linear Terms			-.004	.042	.031	.028
	2 Factors, Quad Terms			.000	.020	.016	.012
2-DIM (r=.60; 20/20)	Correlation Matrix			.111	.069	.111	.068
	Factor Analysis						
	1 Factor	5.7	14.3	-.001	.046	.038	.028
	2 Factors	2.2	5.6	.000	.030	.022	.020
	3 Factors	1.6	3.9	.000	.023	.018	.014
	4 Factors	1.2	3.1	.000	.020	.016	.013
	5 Factors	1.1	2.8	.000	.019	.015	.012
2-DIM (r=.60; 30/10)	Correlation Matrix			.132	.080	.132	.080
	Factor Analysis						
	1 Factor	6.8	16.9	.000	.042	.032	.027
	2 Factors	2.0	5.1	.000	.028	.021	.019
	3 Factors	1.6	3.9	.000	.021	.017	.013
	4 Factors	1.2	3.1	.000	.020	.015	.012
	5 Factors	1.1	.7	.000	.028	.014	.011

Table 3
Summary of Standardized Residuals (SRs)

Data Set and Model	% of Absolute-Valued SRs				Average Absolute-Valued Residual	Average Absolute-Valued SR
	|0 to 1|	|1 to 2|	|2 to 3|	|3 and over|		
1-DIM						
1	32.3	28.2	18.6	21.4	.067	1.86
2	66.6	26.8	5.5	1.1	.033	.89
3	76.8	21.1	1.8	.2	.031	.71
2-DIM (r=.10; 20/20)						
1	49.8	32.7	13.0	4.6	.048	1.20
2	63.6	32.1	3.0	1.4	.036	.86
3	68.2	26.8	3.9	1.1	.035	.84
2-DIM (r=.10; 30/10)						
1	33.2	26.6	16.4	23.9	.075	1.99
2	61.8	27.3	7.7	3.2	.038	.99
3	69.8	26.6	3.6	0.0	.027	.76
2-DIM (r=.60; 20/20)						
1	44.3	26.8	16.6	12.3	.060	1.51
2	67.1	24.8	7.1	1.1	.035	.88
3	72.7	22.7	3.6	0.9	.030	.79
2-DIM (r=.60; 30/10)						
1	39.1	25.7	15.0	20.2	.065	1.79
2	61.6	29.1	5.0	4.3	.038	1.00
3	73.2	24.1	2.7	0.0	.026	.73

Table 4
b-Values with the Simulated Test Data

Last Items	Dataset: 1-DIM Total	1-DIM Subtest	2-DIM r=.10; 20/20 Total	2-DIM r=.10; 20/20 Subtest	2-DIM r=.10; 30/10 Total	2-DIM r=.10; 30/10 Subtest	2-DIM r=.60; 20/20 Total	2-DIM r=.60; 20/20 Subtest	2-DIM r=.60; 30/10 Total	2-DIM r=.60; 30/10 Subtest
21	1.64	1.71	-0.91	-0.88	--	--	-0.43	0.23	--	--
22	-1.05	-1.01	-2.17	-1.88	--	--	0.13	0.26	--	--
23	-1.96	-1.80	0.55	0.74	--	--	-0.63	-0.25	--	--
24	-2.40	-2.24	0.71	0.70	--	--	-1.55	-1.45	--	--
25	1.12	1.13	0.61	0.73	--	--	-1.95	-1.81	--	--
26	1.15	1.11	-1.87	-1.87	--	--	-0.03	0.05	--	--
27	-0.63	-0.59	0.34	0.50	--	--	1.64	1.34	--	--
28	1.81	1.85	1.55	1.40	--	--	-0.61	-0.46	--	--
29	1.52	1.46	-0.39	-0.22	--	--	-1.64	-1.47	--	--
30	-1.52	-1.43	1.55	1.17	--	--	2.07	1.60	--	--
31	-1.37	-1.28	-0.33	-0.31	-10.59	-1.08	-1.97	-1.73	-2.51	-1.25
32	-1.56	-1.48	-1.08	-1.03	-5.38	-0.67	-1.21	1.09	-1.28	-0.82
33	1.60	1.57	-1.54	-1.55	2.70	7.03	1.93	1.52	2.28	1.88
34	-1.88	-1.76	-1.52	-1.53	3.27	10.83	2.12	2.03	2.37	2.07
35	-1.15	-1.21	-1.79	-1.73	2.96	3.48	1.01	0.82	1.31	0.95
36	0.25	0.27	-1.20	-1.11	1.95	-0.22	-0.41	-0.28	-0.52	-0.39
37	0.36	0.37	0.82	0.99	3.80	16.37	2.22	2.25	2.58	2.23
38	0.39	0.31	1.51	1.35	-5.60	0.37	-1.43	-1.35	-1.70	-0.38
39	-0.69	-0-.59	1.90	1.48	5.33	-0.53	-1.69	-1.38	-2.42	-0.69
40	-0.82	-0-.79	-0.70	-0.53	-5.64	-1.09	-1.56	-1.40	-2.04	-1.24
Mean	-0.26	-0.22	-0.20	-0.18	-0.72	3.45	-0.20	-0.13	-0.22	0.24
SD	1.38	1.33	1.30	1.21	5.50	6.07	1.51	1.34	2.12	1.40
r	.99		.99		.56		.99		.98	