DOCUMENT RESUME

ED 395 966 TM 025 104

AUTHOR McKinley, Robert
TITLE Confirmatory Analysis of Test Structure Using Multidimensional Item Response Theory.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-89-31
PUB DATE Jun 89
NOTE 44p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Ability; *Chi Square; *Goodness of Fit; *Item Response Theory; Matrices; Maximum Likelihood Statistics; Simulation; *Test Construction; Test Items
IDENTIFIERS Akaike Information Criterion; *Confirmatory Factor Analysis; Likelihood Ratio Criterion; *Multidimensional Approach

ABSTRACT

A confirmatory approach to assessing test structure using multidimensional item response theory (MIRT) was developed and evaluated. The approach involved adding to the exponent of the MIRT model an item structure matrix that allows the user to specify the ability dimensions measured by an item. Various combinations of item structures were fit to two sets of simulation data with known true structures, and the results were evaluated using a likelihood ratio chi-square statistic and two information-based model selection criteria. The results of these analyses support the use of the confirmatory MIRT approach, since it was found that the procedures could recover true item structures. It was also found that adding an additional ability dimension that forces together items that ought not to be together noticeably deteriorates the quality of the solution. On the other hand, imposing structures different from, but not inconsistent with, the true structures does not necessarily yield worse fit. Finally, in terms of model fit statistics, the consistent Akaike information criterion performed better than the simple Akaike information criterion, while the likelihood ratio chi-square was clearly inadequate. (Contains 12 tables and 30 references.) (Author/SLD)

RESEARCH REPORT

CONFIRMATORY ANALYSIS OF TEST STRUCTURE USING MULTIDIMENSIONAL ITEM RESPONSE THEORY

Robert McKinley

Educational Testing Service
Princeton, New Jersey
June 1989

Confirmatory Analysis of Test Structure using Multidimensional Item Response Theory

Abstract

The purpose of this research was to develop and evaluate a confirmatory approach to assessing test structure using multidimensional item response theory (MIRT). The approach investigated involves adding to the exponent of the MIRT model an item structure matrix that allows the user to specify the ability dimensions measured by an item. Various combinations of item structures were fit to two sets of simulation data with known true structures, and the results were evaluated using a likelihood ratio chi-square statistic and two information-based model selection criteria. The results of these analyses support the use of the confirmatory MIRT approach, since it was found that the procedures could recover the true item structures. It was also found that adding an additional ability dimension that forces together items that ought not to be together noticeably deteriorates the quality of the solution. On the other hand, imposing structures different from, but not inconsistent with, the true structures does not necessarily yield worse fit. Finally, in terms of model fit statistics, the consistent Akaike information criterion performed better than the simple Akaike information criterion, while the likelihood ratio chi-square was clearly inadequate.

Confirmatory Analysis of Test Structure

using Multidimensional Item Response Theory

Introduction

Although item response theory (IRT) has proven to be a very powerful and useful measurement tool, use of IRT models has been somewhat limited because the available models require the assumption that the test being analyzed measures only a single ability dimension. This unidimensionality assumption often limits the application of IRT-based methods to tests consisting of relatively homogeneous sets of items, such as might be found on a vocabulary test. Tests including items sampled from several content areas, such as a science test containing both physics and chemistry items, are probably not sufficiently homogeneous as to permit analysis using IRT. Such may also be the case with tests containing multi-faceted items, such as a mathematics test containing problem-solving items requiring a high level of reading comprehension or vocabulary skill.

Because it is clear that many tests measure more than a single ability dimension (Traub, 1983), attempts have been made to extend IRT to multidimensional tests. In multidimensional IRT, or MIRT, examinee responses are modeled as a function of a set of examinee traits, and the assumption of unidimensionality is replaced by the less restrictive requirement that the dimensionality of the item responses matches the dimensionality of the set of examinee traits used in the MIRT model.

This assumption is less restrictive in that it permits the application of IRT methods to a much broader range of instruments, but it has its own disadvantages. The most serious difficulty encountered in the application of MIRT methodology is matching the model dimensionality to the dimensionality of the test. This problem is particularly serious in view of the fact that there are no generally accepted procedures for determining the dimensionality of a test, especially if the test items are dichotomously scored. In addition, in the case of correlated dimensions, a number of multidimensional solutions might fit the data almost equally well. If this is so, it may be desirable to choose a solution with a theoretical foundation in cognitive psychology or test content rather than a solution with a slightly greater likelihood, perhaps obtained by fitting error.

The purpose of this research was to develop and evaluate multidimensional IRT procedures designed to permit the extraction of theory-based solutions. The procedures developed include a MIRT model, a mechanism for imposing a priori structures on the data, procedures for estimating the model parameters, and model selection criteria for choosing among alternative structures. The evaluation performed on these procedures included assessing the reasonableness of simulation data generated to fit the model, evaluating the estimation procedures, and evaluation of the model selection criteria.

Background

Multidimensional IRT

Most of the recent progress in MIRT research has occurred in two areas -- the development of nonlinear factor analysis models (Bock, Gibbons, & Muraki, 1985; Christoffersson, 1975; and Muthen, 1978), and the development of multidimensional two- and three-parameter logistic MIRT models (McKinley, 1983, 1987, in preparation; Reckase, 1985; Reckase, Ackerman, & Carlson, 1988; Reckase & McKinley, 1985). Work on these procedures is still at an early stage, but has progressed to the point that estimation procedures are available.

Nonlinear factor analysis model. The basic nonlinear factor analysis model is based on the two-parameter logistic normal ogive (2PNO) model, although the method does allow for the input of previously estimated c-parameter values. The 2PNO model assumes there is an unobservable response variable which is on a continuous scale, and which is dichotomized into an observed score of 0 or 1 depending on whether the examinee is above or below some threshold point. The 2PNO model is given by

$$P_i(\theta_j) = \Phi((Y_i - \underline{a}_i'\underline{\theta}_j)/s_i) \quad , \qquad (1)$$

where $P_i(\theta_j)$ is the probability of a correct response to item i by examinee j, $\Phi(x)$ represents the cumulative normal distribution, a_i is a column vector of k item factor loadings j, θ_j is a column vector of k factor scores for examinee j, s is the standard deviation of the normal distribution, Y represents the

threshold beyond which an examinee will respond correctly to the item, and there are k dimensions.

Applications of nonlinear factor analysis procedures include analysis of the item pool for the computerized version of the Armed Services Vocational Aptitude Battery, or ASVAB, by Zimowski and Bock (1987), analysis of the Graduate Management Admission Test, or G MAT, by Kingston (1986), and analysis of the Graduate Records Examination (GRE) Subject Test in Mathematics by McKinley and Kingston (1987).

Logistic models. In the logistic model approach, the normal ogive model is replaced with the logistic distribution. Although the parallel to factor analysis is lost in this approach, all of the desirable properties of IRT are maintained, and the computation is greatly simplified. The multidimensional three-parameter logistic, or M3PL, model is given by

$$P_i(\underline{\theta}_j) = c_i + (1-c_i)/(1+\exp(-1.702(b_i + \underline{a}_i'\underline{\theta}_j))) \quad , \quad (2)$$

where $P_i(\underline{\theta}_j)$ is the probability of a correct response to item i by examinee j, $\underline{\theta}_j$ is a column vector of k ability parameters for examinee j, $\underline{a}_i$ is a column vector of k discrimination parameters for item i, b_i is the threshold parameter for item i, and c_i is the lower asymptote parameter for item i. There are k dimensions, and the ability and discrimination parameter vectors contain one element for each dimension.

Uses of logistic MIRT models thus far have been somewhat limited, primarily due to the recency of the development of estimation procedures for

applying the models. The multidimensional two-parameter logistic (M2PL) model, which is formed by holding the M3PL c-parameter fixed at zero for all items, was applied to a French proficiency exam by Kaya-Carton (1988). In this application, parameters of the M2PL model were obtained using the MULTIDIM (McKinley, 1987) program with item lower asymptote parameters fixed at zero. The method was compared to maximum likelihood factor analysis and boolean factor analysis. Despite the shortness of the test (18 items) and the small sample size (between 700 and 800 examinees), the results obtained were positive. The MIRT solution was found to be interpretable and consistent with the factor analysis solutions.

Estimation. For the nonlinear factor analysis procedure, the TESTFACT program (Wilson, Wood, & Gibbons, 1984) is available. The TESTFACT program uses marginal maximum likelihood estimation (MMLE) to estimate the item parameters of the 2PNO factor analysis model, and provides a mechanism for using item guessing parameters obtained from a previous analysis. In this approach to estimation, examinee factor scores are treated as nuisance parameters, and are removed from the estimation process by specifying a distribution for them, and integrating over that distribution.

For the M2PL model, the MAXLOG (McKinley & Reckase, 1983) and MIRTE (Carlson, 1987) programs are available, and for the M2PL and M3PL models the MULTIDIM program (McKinley, 1987) is available. The MAXLOG and MIRTE programs are based on a simultaneous, or joint, maximum likelihood estimation (MLE) algorithm. In MLE item parameters are estimated while ability parameters are

held fixed, and then ability parameters are estimated while item parameters are held fixed. This two-step process is repeated until the procedure converges. The MULTIDIM program uses MMLE.

Model selection. In MIRT, model selection is relatively straightforward. Solutions for relatively simple models (such as the unidimensional 3PL model) are obtained first. Models of increasing complexity are then created by adding parameters. These more complex models subsume simpler models, making it possible to test the significance of the contribution of the additional parameters using procedures such as are implemented in TESTFACT.

For example, assume one- and two-dimensional MIRT solutions have been obtained on the same data using the M2PL model. Comparing the solutions can be accomplished by computing, for each solution, a measure of fit such as the likelihood ratio chi-square statistic (Bock, Gibbons & Muraki, 1985). This statistic is given by

$$G^2 = 2 \sum_{j=1}^{J} r_j \ln(r_j/NP_j) \quad , \qquad (3)$$

where J is the number of possible unique response strings for the item set to be calibrated, r_j is the number of examinees with response string j, N is the total number of examinees in the calibration sample, and P_j is computed as

$$P_j = \sum_{k=1}^{q} L_j(X_k) W_k \quad , \qquad (4)$$

where P_j represents the total likelihood of observing response string j, $L_j(\underline{x}_k)$ is the likelihood of observing response string j given an ability vector equal to $\underline{x}_k$, and $\underline{x}_k$ and W_k are the quadrature nodes and weights used for numerically integrating over the ability distribution.

The degrees of freedom for the statistic given by Equation 3 are given by

$$df = 2^n - 2(m+2) \quad , \qquad (5)$$

where n is the number of items and m is the number of dimensions. If the c-parameter is not estimated the term in parentheses is m+1.

While it is doubtful the statistic given by Equation 3 is actually distributed as a chi-square, the difference between the values of G^2 for subsuming models ought to be distributed as a chi-square (Haberman, 1977). The degrees of freedom for the difference between two values of G^2 is equal to the difference between Equation 5 for the two solutions. For MIRT models, this equals the difference in the number of item parameters estimated.

MIRT Limitations

Logistic and factor analytic MIRT procedures share a very serious shortcoming -- they are prone to overfitting the data. This occurs in part because these procedures depend on large sample chi-square tests to assess the significance of the contribution of additional ability dimensions. This results in a very powerful test that often results in retention of statistically significance, yet uninterpretable dimensions, perhaps based on nothing more than chance relations in the data.

Confirmatory MIRT

Confirmatory MIRT was developed largely as a response to this shortcoming in MIRT. The goals of confirmatory MIRT, then, are to avoid overfitting the data and to enhance the interpretability of obtained solutions by forcing a correspondence between estimated ability dimensions and the content and cognitive processes the instrument was intended to measure. This is accomplished by inserting into the MIRT model an item structure matrix that determines for a given item which ability dimensions are measured. Items that in theory ought to measure the same ability dimensions are thereby clustered together by assigning them identical structure matrices. Similarly, items that in theory ought to differ as to which dimensions are measured are forced apart by assigning them dissimilar structure matrices. Thus, ability dimensions are defined prior to estimation based on a priori considerations.

Confirmatory MIRT model. The confirmatory MIRT, or CMIRT, procedure used in this research is based on a modification of the M3PL model. As indicated above, the modification consists of adding an item structure matrix. The confirmatory M3PL, or CM3PL, model is given by

$$P_i(\underline{\theta}_j) = c_i + (1-c_i)/(1+\exp(-1.702(b_i + \underline{a}_i'\underline{S}_i\underline{\theta}_j))) \quad , \qquad (6)$$

where $\underline{S}_i$ is the item structure matrix for item i, and the remaining terms are as previously defined.

Item structure matrices. The item structure matrix identifies for a given item the ability dimensions measured. This is accomplished by

specifying either a 0 or a 1 for each element of $\underline{S}$ in such a way that, when it is premultiplied by the transpose of the item discrimination vector, item discrimination parameters are zeroed out for those ability dimensions the item does not measure. For example, if $\underline{S}$ is the identity matrix, the item measures all of the ability dimensions. If an item is to measure only the first and third dimensions in a three-dimensional solution, $\underline{S}$ would be given by

$$\underline{S} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} . \quad (7)$$

Estimation. The estimation procedure used in this study was the CONFIRM program, which is based on an EM algorithm similar to those described by Bock and Aitkin (1981), Bock, Gibbons, and Muraki (1985), Mislevy and Bock (1985), and Reckase and McKinley (1985). The algorithm has been modified for this application to allow collapsing across extraneous ability dimensions in accordance with the hypothesized item structures.

In this algorithm, item parameter estimation is performed using a two-step marginal maximum likelihood procedure. In this procedure, examinee ability is treated as a random variable, and is eliminated from the estimation process by specifying a form for the ability distribution and integrating over that distribution. The integration over the ability distribution, which is

accomplished through numerical quadrature, is performed during the first step, the E-step, and produces an expected sample size and number-correct score at each quadrature node for each item.

In MIRT applications, the values produced by the E-step are immediately input into the M-step. In CMIRT applications, however, there is an intermediate step performed prior to execution of the M-step. This additional step involves a process that collapses expected sample size and number-correct scores over ability dimensions not measured by an item. The output of this step, then, is the expected sample size and number-correct scores for each item collapsed in accordance with the specified item structures.

These values are then input into the M-step, which uses these values to perform marginal maximum likelihood item parameter estimation. This process is repeated until the item parameter estimates converge. As a final, optional step, the item parameter estimates can be used as input into an expected a posteriori, or EAP, ability estimation routine. For a complete description of the CONFIRM program, see McKinley (in preparation).

Model selection. Clearly many different sets of S matrices can be applied to a given set of data. Consequently, a procedure for selecting from among them is necessary. Unfortunately, the chi-square procedure described above often cannot be applied, since in CMIRT alternative models are not necessarily subsuming. For example, consider a four item test in which the first two items are intended to measure vocabulary, and the last two items are intended to measure reading comprehension. One test structure that might be

evaluated using CMIRT involves fitting one common ability dimension that all four items are assumed to measure, and one additional dimension that only the last two items are assumed to measure.

An alternative test structure that might be considered would be to assume all four items measure a common dimension, and the middle two items measure a second dimension. This structure would not correspond to any content-based hypothesis about the structure of the test, but would serve as a useful baseline for evaluating the first model. That is, it would provide an indication of the extent of improvement (or deterioration, as the case may be) in the quality of the solution to be expected simply from adding a second dimension for two items.

Unfortunately, the model selection procedure described above cannot be applied in this case. Not only are the competing models not subsuming, but they result in equal degrees of freedom. Although the resulting chi-squares could still be visually compared, the significance of the difference in the chi-squares could not be tested.

One way these two competing models of test structure might be compared is based on the work of Akaike (1973, 1987). This approach is based on a criterion called the entropic information criterion (Bozdogan, 1987), also known as the AIC, and involves evaluating model fit in terms of the natural logarithm of the likelihood of the solution, which is presumed to be an approximation of the expected natural logarithm of the likelihood of the true

model. The greater the likelihood of the solution (in practice, the lower the negative log likelihood), the closer the fitted model is presumed to approximate the true model.

The AIC statistic is given by

$$AIC = -2 \log(L) + 2k \quad , \qquad (8)$$

where log(L) denotes the natural logarithm of the likelihood, and k is the number of parameters estimated. The 2k term constitutes a sort of a penalty function that penalizes over-parameterization.

A variation on the AIC, called the consistent AIC (CAIC), was proposed by Bozdogan (1987). This statistic was derived in response to the criticism that the AIC statistic does not provide an asymptotically consistent estimate of model order (Bozdogan, 1987).

The CAIC statistic is given by

$$CAIC = -2 \log(L) + k(\log(n)+1) \quad , \qquad (9)$$

where n is the sample size. This modification of the AIC has the effect of increasing the penalty for over-parameterization and, consequently, tends to lead to the selection of simpler models.

One reason these statistics are desirable is that they are designed to identify which of a class of models is the closest approximation to the true model. Unlike classical chi-square tests of model fit, in which a constrained

model is typically evaluated by comparing it to a more saturated, subsuming model, with the AIC and CAIC statistics each model under consideration is evaluated in terms of its closeness to the true model. Different models are not directly compared, so there is no requirement that competing models be subsuming.

The AIC and CAIC statistics have a very interesting property that makes them very desirable even in situations where use of the chi-square statistic is possible -- the level of significance for testing whether a particular model is the best-fitting model is implicit to the model-selection criterion (Bozdogan, 1987). In effect, the critical value is embodied in the penalty for over-parameterization, and the probability of a Type I error is determined by the sample size. Thus, selecting the CAIC over the AIC is tantamount to selecting a larger critical value, which results in a reduction in the Type I error rate. Moreover, once a statistic has been selected, increasing the sample size has the effect of decreasing the probability of a Type I error. Indeed, the Type I error rate decreases exponentially with increased sample size. In fact, for the CAIC statistic, the error rate asymptotically goes to zero.

Method

Overview

The evaluation of the CMIRT procedure described above was performed using simulation data. Two sets of simulation data were generated using different true structure matrices. Several different solutions, based on different

structure matrices, were then obtained and compared in terms of model fit using the indices described above.

Although the model proposed above was a multidimensional extension of the unidimensional 3PL model, for the purpose of evaluating the proposed CMIRT procedures the lower asymptote parameter, c, was held constant. This was done to avoid complications arising from errors in estimation of the c-parameter. At this point it is unclear how such errors in estimation affect the solution obtained, but experience with unidimensional IRT estimation procedures suggest the effect is potentially serious.

Data Generation

As was indicated above, two sets of data were used for this evaluation. As was pointed out previously, for both sets of data true c-parameters were not allowed to vary. Rather, a constant value of 0.15 was used. For each simulation dataset responses were generated for 1000 examinees and 80 items.

The first set of simulation data was generated to have only one ability dimension. For these data examinee true abilities were selected randomly from a standard normal distribution. The same item structure matrix was used for all items. The matrix used was, in effect, a scalar with a value of 1.

The second set was generated using three uncorrelated ability dimensions. Examinee true abilities were selected from a trivariate normal distribution with a mean vector equal to zero and a covariance matrix equal to the identity matrix. Two different item structure matrices were used to generate these data. The first matrix, used for the first 40 items, is given by

$$\underline{S} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad . \qquad (10)$$

Thus, these items each measured the first two ability dimensions. The second item structure matrix, used for items 41 through 80, is given by

$$\underline{S} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad . \qquad (11)$$

These items each measured the first and third ability dimensions.

Table 1 provides summary statistics for each model parameter used to generate each dataset. Note that for the three-dimensional data, there were only 40 item discrimination values on the second and third dimensions. The remaining 40 values were set equal to zero.

For the unidimensional data, the correlation between the true a-values and b-values was 0.01. For the three-dimensional data, the correlation between the a-values was -0.14 for the first and second dimensions, 0.21 for the first and third dimensions, and 0.0 for the second and third dimensions (note that no item had a-values for both the second and the third dimensions). The b-values for the three-dimensional data had correlations of -0.09, 0.23, and 0.17 with the a-values on the first, second and third dimensions, respectively. For the three-dimensional data, the correlations of the true

ability parameters were 0.0 for the first and second dimensions and for the first and third dimensions, and -0.02 for the second and third dimensions.

Table 1

True Parameter Distribution Summary Statistics

Dataset/Parameter	N	Mean	Std. Dev.	Min.	Max.
1-Dimensional					
a	80	1.00	0.29	0.52	1.48
b	80	-0.14	0.93	-2.28	1.83
c	80	0.15	0.00	0.15	0.15
θ	1000	-0.05	0.93	-2.85	2.77
3-Dimensional					
a_1	80	1.01	0.28	0.53	1.48
a_2	40	1.00	0.26	0.53	1.43
a_3	40	0.92	0.29	0.53	1.48
b	80	0.01	0.93	-3.34	2.28
c	80	0.15	0.00	0.15	0.15
θ_1	1000	-0.05	0.99	-4.03	3.76
θ_2	1000	-0.01	0.99	-3.13	2.83
θ_3	1000	0.00	0.99	-3.12	3.70

Solutions

Several different solutions were obtained for each set of data. For both sets of data, the first solution obtained was unidimensional, and used for each item an item structure matrix that was, in effect, a scalar with a value of 1. For the unidimensional data, this solution, signified by the code 1DU (one-dimensional unconstrained), represents the true structure of the data.

The second solution obtained for each set of data was two-dimensional, with each item measuring both dimensions. This was accomplished by using for

all items an item structure matrix equal to a two-by-two identity matrix. This solution is signified by the code 2DU (two-dimensional unconstrained).

The third matrix used for both sets of data was also two-dimensional. However, for this solution, designated 2DC (two-dimensional constrained), the two-by-two identity matrix was used only for items 41 through 80. For items 1 through 40, the matrix used is given by

$$\underline{S} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \quad . \qquad (12)$$

For the three-dimensional simulation data, two additional solutions were obtained. The first of these, designated as 3DCa (three-dimensional constrained, solution a), used for each item the matrix used to generate the data (given by Equations 10 and 11). The other solution, designated 3DCb (three-dimensional constrained, solution b), used for items 1 through 20 and items 61 through 80 the matrix given by Equation 10, and for items 21 through 60 the matrix given by Equation 11. Thus, the first and last 20 items measured the first two ability dimensions, while the middle 40 items measured the first and third ability dimensions.

For all solutions derived during this evaluation, estimates of the c-parameter were held fixed at their true value, 0.15. All other item parameters were estimated, with the restriction that a maximum value of 1.8 and a minimum value of 0.1 was imposed on the a-values. The estimation procedure was allowed to cycle through the EM process until the likelihood of

the response matrix, given the CMIRT model and current item parameter estimates, ceased to increase, or until the item parameter estimates ceased to change, whichever came first.

Analyses

For each set of data the following analyses were performed. First, a principal components analysis of phi coefficients and a traditional item analysis were performed to evaluate the reasonableness of the generated data, and to provide information to aid in the interpretation of the results of the CMIRT analyses. (It should be noted that tetrachoric correlations were also computed for both sets of data, but in both cases the correlation matrix was found to be non-Gramian.)

Second, several different CMIRT solutions were obtained for each set of data. Then, these solutions, which varied not only in dimensionality, but also in the item structure matrices used, were evaluated using the AIC and CAIC model selection criteria, as well as the likelihood ratio chi-square goodness-of-fit statistic.

In addition, as an aid to interpretation correlations between the true parameters and the parameter estimates obtained for each solution were computed. Of course, these correlations cannot be used to evaluate the quality of the solutions, since in some cases estimates are obtained for parameters not even used in the data generation. Moreover, it is possible that some solutions with estimates that have low correlations with the true parameters may, in fact, represent rotations and/or translations of the true structures used to generate the data.

Finally, residuals were computed and analyzed, using the procedure described by Divgi (1980), to determine whether there appeared to be any interpretable common variance remaining after the model was fit to the data. Residuals were computed as

$$R_{ij} = P_{ij} - u_{ij} \quad , \qquad (13)$$

where P_{ij} is the probability of the observed response to item i by examinee j predicted from the MIRT model, and u_{ij} is the observed response. Residual correlation matrices were analyzed using principal components analysis, and resulting component loadings were examined. [Note that, since residuals are on a continuous scale, the problem of using principal components analysis with binary data is avoided using this procedure.]

Results

Item Analyses

Table 2 presents the means, standard deviations, minimums, and maximums of the examinee number-correct scores, item-total biserials, and item proportion-correct (p+) scores for each dataset. The correlation between item biserials and p+ values was 0.59 for the unidimensional data and 0.65 for the three-dimensional data. The KR-20 coefficient of reliability was 0.94 for the unidimensional data, and 0.95 for the three-dimensional data.

Table 2

Traditional Item Analysis Results

Dataset/Statistic	Mean	S.D.	Min.	Max.
1-Dimensional				
Number-Correct	42.95	15.38	7.00	77.00
Item Biserial	0.51	0.15	0.09	0.77
Item p+	0.54	0.19	0.17	0.92
3-Dimensional				
Number-Correct	45.34	16.52	8.00	80.00
Item Biserial	0.56	0.11	0.22	0.73
Item p+	0.57	0.18	0.19	0.92

These results indicate that the response data simulated according to the CMIRT model were fairly realistic. The means and standard deviations of the examinee number-correct scores suggest a test of appropriate difficulty for the distribution of ability used in this research. Moreover, item p+ and biserial values varied to a reasonable degree. The only departures from what might typically be obtained with real data are: 1) the mean biserials and the KR-20 are somewhat high, indicating the relative purity of the simulation data; and, 2) the correlation between the item p+ and biserial values was high (for comparison, values of this correlation computed on items pretested for the Test of English as a Foreign Language over the eight year period from 1981 through 1988 were 0.33, 0.29, and 0.44 for Sections 1, 2, and 3, respectively). This, too, is probably a reflection of the purity of the data.

Principal Components Analyses

Table 3 summarizes the results of the principal components analyses performed on the two datsets. Provided for each dataset are the first 10 eigenvalues, along with the percent and cumulative percent of variance.

Table 3

Principal Components Analysis Results

Component	1-Dimensional Data			3-Dimensional Data		
	Eigen-value	Percent of Variance	Cumulative Percent	Eigen-value	Percent of Variance	Cumulative Percent
1	14.91	18.6	18.6	16.44	20.6	20.6
2	1.95	2.4	21.1	6.51	8.1	28.7
3	1.33	1.7	22.7	1.97	2.5	31.2
4	1.31	1.6	24.4	1.54	1.9	33.1
5	1.27	1.6	26.0	1.34	1.7	34.8
6	1.24	1.6	27.5	1.18	1.5	36.2
7	1.22	1.5	29.1	1.13	1.4	37.6
8	1.20	1.5	30.6	1.11	1.4	39.0
9	1.18	1.5	32.1	1.09	1.4	40.4
10	1.17	1.5	33.6	1.08	1.3	41.7

As has been pointed out by many researchers (see, for example, Bock, Gibbons & Muraki, 1985; Carroll, 1945; Lord & Novick, 1968; Reckase, 1981; and, Tucker, Humphreys, & Roznowski, 1986), a principal components analysis of phi coefficients is fraught with dangers. Among these are the likelihood of obtaining spurious components due to item difficulty and nonlinearity. The results reported above illustrate these problems. Although the first set of data were generated to have only one dimension, the principal components analysis suggests the presence of a second component. However, further

examination reveals that the second component is essentially a difficulty component. The correlation between the item loadings on the second component and item proportion-correct score was -0.86. In contrast, the correlation between the first component and the item proportion-correct scores was 0.48, which is about what would be expected in light of the previously reported finding that the item proportion-correct scores and item biserials had a correlation of 0.59 for these data. The correlation between the first component and the item biserials was 0.98.

A similar, though more complex, pattern emerged for the three-dimensional data. The data were generated to have three ability dimensions -- one dimension which all items measured, one dimension only the first 40 items measured, and one dimension only items 41 through 80 measured. This resulted in a principal components solution in which there were two components -- one common component, on which all items had positive loadings, and one bipolar component, on which the first 40 items all had positive loadings, and the last 40 items all had negative loadings. However, the principal components analysis results shown in Table 3 indicate the presence of a small third component. As was the case with the unidimensional data, this additional component is the result of using phi coefficients. The correlation between the loadings on the third component and the item proportion-correct scores was -0.89. The correlation of item proportion-correct scores and item loadings was 0.49 and -0.11 for the first and second components, respectively. As

reported above, the correlation between the item biserials and item proportion-correct scores for these data was 0.65. The item biserials had correlations of 0.95, -0.12 and -0.53 with the first, second, and third components, respectively.

CMIRT Analyses

Unidimensional data. Table 4 presents summary statistics for the parameter estimates obtained for the unidimensional data. Shown are the means, standard deviations, minimum, and maximum values for each parameter estimated in each solution, along with the number of values estimated for each parameter. It should be noted that no attempt was made to place the estimates on the same scale as the true parameters. Nor have different sets of estimates been placed on the same scale. Consequently, the summary statistics shown in Table 4 should not be used to assess similarity of estimates to true parameters or other estimates.

It can be seen from Table 4 that the summary statistics for the a-values on the first dimension (the only dimension for the 1DU solution) were similar across solutions, although the mean was a little higher for the 2DU solution. Likewise, the ability estimate distributions for the first dimension and the b-values did not vary much across solutions. The second dimension summary statistics were also similar across solutions for both the a-values and ability estimates, and in both cases the statistics were different from those obtained for the first dimension. The ability estimates on the second

dimension has relatively small standard deviations, and the a-values on the second dimension had low means compared to the first dimension, suggesting the second estimated dimension may be nothing but noise.

Table 4

Parameter Estimate Distribution Summary Statistics for the Unidimensional Data

Solution/Parameter	N	Mean	Std. Dev.	Min.	Max.
1DU					
a	80	0.77	0.23	0.37	1.26
b	80	-0.12	0.95	-3.41	1.91
θ	1000	-0.08	1.13	-2.78	2.99
2DC					
a_1	80	0.71	0.22	0.33	1.20
a_2	40	0.30	0.15	0.10	0.74
b	80	-0.15	0.92	-2.65	1.92
θ_1	1000	-0.04	1.18	-2.77	3.15
θ_2	1000	-0.01	0.54	-1.54	1.72
2DU					
a_1	80	0.85	0.27	0.34	1.42
a_2	80	0.44	0.20	0.10	0.86
b	80	-0.08	0.97	-3.60	1.98
θ_1	1000	-0.11	0.90	-2.37	2.58
θ_2	1000	-0.06	0.69	-1.82	1.97

Table 5 shows the correlations of the true item parameters with the estimated item parameters for each solution for the unidimensional data. [Note that, since the c-parameter was held fixed, it is not included in Table 5.] It can be seen from the data shown in Table 5 that, for each solution the a-parameter estimates for the first dimension were highly correlated with the true parameters. The meaning of the moderate correlations obtained between

the a-parameter estimates on the second dimension and the true a-values is unclear, and may be a result of overfitting. It is interesting to note that the correlation between the a-parameter estimates on the second dimension and the true a-values was higher for the 2DU solution than for the 2DC solution, and that the correlation between the a-parameter estimates on the first dimension and the true values was lower for the 2DU solution. It appears as though increasing the number of parameters estimated on the second dimension produced a deterioration of the fitting of the first dimension.

Table 5

True and Estimated Item Parameter Correlations for the Unidimensional Data

Solution/Parameter	N	True Parameter	
		a	b
1DU			
a	80	0.89	0.07
b	80	0.05	0.99
2DC			
a_1	80	0.91	0.15
a_2	40	0.53	-0.18
b	80	0.03	0.99
2DU			
a_1	80	0.81	-0.04
a_2	80	0.67	0.22
b	80	0.07	0.98

The true and estimated ability parameter correlation was 0.96 for the 1DU solution. For the 2DC solution, the correlation between the true ability

parameter and the ability estimates on the first dimension was 0.96, and for the second dimension it was 0.32. For the 2DU solution, the correlation between the true abilities and the ability estimates was 0.92 for the first dimension and 0.65 for the second.

Table 6 shows the chi-square, AIC, and CAIC values obtained for each solution for the unidimensional data. For these data, all three pair-wise comparisons could be tested for the significance of the differences in the associated chi-square values. All three chi-square differences were significant.

Table 6

Model Selection Criteria Values
for The Unidimensional Data

Solution	Chi Square	AIC	CAIC
1DU(true)	70947.6	85083.2	86028.4
2DC	71102.7	85318.2	86499.8
2DU	70618.0	84913.5	86331.4

As shown in Table 6, if the chi-square criterion is used, the unconstrained two-dimensional solution would be selected as optimal, even though the data are actually unidimensional. The unidimensional solution would be chosen over the constrained two-dimensional solution. Use of the AIC would result in the same ordering of solutions. Using the CAIC as a

criterion, however, would result in selection of the unidimensional solution. The unconstrained two-dimensional solution would be selected over the constrained two-dimensional.

Three-dimensional data. Table 7 provides summary statistics for the parameter estimates obtained for the three-dimensional data. As was the case previously, no attempt at scaling these estimates has been made.

The summary statistics for the a-parameter estimates on the first dimension were similar across solutions, although for the 2DU and 3DCb solutions the mean a-value tended to be a little lower than for the other solutions, and the mean a-value was a little higher for the 3DCa solution than for the others. There was very little variation across solutions in the ability estimate distributions for the first dimension, or for the b-values. For the two-dimensional solutions, the a-values on the second dimension differed noticeably in mean value, with the mean being 0.3 higher for the 2DC solution, and the 2DC a-values on the second dimension were less variable than for the 2DU solution. There was not a difference in the second dimension ability estimate distributions for these two solutions.

For the three-dimensional solutions, the a-values on the second and third dimensions had similar means and standard deviations, and on both dimensions the means were higher for the 3DCa solution than for the 3DCb solution. The ability estimate distributions were similar for the second and third dimensions for both three-dimensional solutions.

Table 7

Parameter Estimate Distribution Summary Statistics for the Three-dimensional Data

Solution/Parameter	N	Mean	Std. Dev.	Min.	Max.
1DU					
a	80	0.78	0.19	0.41	1.26
b	80	0.02	0.85	-2.54	2.26
θ	1000	-0.05	1.17	-3.04	3.31
2DC					
a_1	80	0.77	0.23	0.31	1.25
a_2	40	0.92	0.22	0.53	1.31
b	80	-0.06	1.02	-3.37	2.57
θ_1	1000	0.08	1.20	-2.61	3.06
θ_2	1000	-0.07	1.12	-3.11	3.48
2DU					
a_1	80	0.65	0.40	0.10	1.40
a_2	80	0.62	0.36	0.10	1.33
b	80	0.10	1.01	-3.19	2.68
θ_1	1000	-0.09	1.18	-2.74	2.75
θ_2	1000	-0.09	1.16	-2.61	3.06
3DCa					
a_1	80	0.81	0.24	0.41	1.58
a_2	40	0.69	0.19	0.24	1.04
a_3	40	0.68	0.17	0.31	1.10
b	80	0.00	1.03	-3.18	2.62
θ_1	1000	-0.03	1.03	-2.51	2.76
θ_2	1000	-0.04	0.96	-2.72	2.64
θ_3	1000	0.03	1.00	-2.89	2.61
3DCb					
a_1	80	0.62	0.32	0.10	1.32
a_2	40	0.55	0.38	0.10	1.17
a_3	40	0.58	0.37	0.10	1.30
b	80	-0.05	0.99	-3.36	2.46
θ_1	1000	0.01	1.27	-2.78	2.83
θ_2	1000	0.08	1.04	-2.10	2.65
θ_3	1000	-0.01	1.03	-2.65	2.45

Table 8 shows the intercorrelations for the true and estimated item parameters for the three-dimensional data. Because the number of values in common to the true structure and the imposed structures varies across dimensions and solutions, the number of items on which each correlation is based is shown in parentheses after each correlation.

The values shown in Table 8 indicate that the dimension estimated for the 1DU solution was most strongly related to the first true dimension. For the 2DC solution the two estimated dimensions appeared to be equally strongly related to the first true dimension. The first estimated dimension appeared to be slightly more strongly related to the first true dimension, while the second estimated dimension appeared to be more strongly related to the second true dimension.

For the 2DU solution, the first estimated dimension appeared to be related to the second true dimension, while the second estimated dimension was related to the third true dimension. Neither dimension appeared to be strongly related to the first true dimension.

For the 3DCA solution, the first estimated dimension appeared to correspond to the first true dimension, while the second and third estimated dimensions corresponded to the second and third true dimensions, respectively. For the 3DCb solution, the first estimated dimension was most strongly related to the second true dimension, while the second and third estimated dimensions both appeared most strongly related to the third true dimension.

Table 8

True and Estimated Item Parameter Correlations for the Three-Dimensional Data

Solution/Parameter	True Parameter			
	a_1(N)	a_2(N)	a_3(N)	b(N)
1DU				
a	0.76(80)	0.44(40)	-0.10(40)	.17(80)
b	-0.06(80)	0.17(40)	0.25(40)	.99(80)
2DC				
a_1	0.57(80)	0.42(40)	0.40(40)	0.00(80)
a_2	0.58(40)	0.80(40)	0.00(0)	0.16(40)
b	-0.10(80)	0.16(40)	0.23(40)	0.99(80)
2DU				
a_1	0.37(80)	0.70(40)	-0.40(40)	0.11(80)
a_2	0.24(80)	0.17(40)	0.67(40)	-0.07(80)
b	-0.08(80)	0.17(40)	0.24(40)	0.99(80)
3DCa				
a_1	0.86(80)	0.31(40)	-0.17(40)	0.02(80)
a_2	0.19(40)	0.86(40)	0.00(0)	0.15(40)
a_3	-0.09(40)	0.00(0)	0.85(40)	0.11(40)
b	-0.09(80)	0.17(40)	0.24(40)	0.99(80)
3DCb				
a_1	0.40(80)	0.68(40)	-0.29(40)	0.09(80)
a_2	0.26(40)	0.10(20)	0.52(20)	-0.13(40)
a_3	0.07(40)	-0.02(20)	0.69(20)	-0.01(40)
b	-0.10(80)	0.15(40)	0.25(40)	0.99(80)

The true and estimated ability parameter intercorrelations are shown in Table 9. These data indicate that, for the 1DU solution, the estimated abilities were most similar to the first dimension true abilities. For the 2DC solution, the first dimension ability estimates were most highly correlated with the first dimension true abilities, while the second dimension

ability estimates were most strongly related to the second dimension true abilities. The first dimension ability estimates were also fairly strongly related to the third dimension true abilities.

For the 2DU solution, the ability estimates on the first dimension were most strongly related to the second dimension true abilities, and the second dimension ability estimates were most strongly related to the first dimension true abilities. The first and second dimension ability estimates were equally strongly related to the first dimension true abilities.

For the 3DCa solution the first dimension estimates were most strongly related to the first dimension true estimates, the second dimension ability estimates were most strongly related to the second dimension true abilities, and the third dimension estimates were most strongly related to the third dimension true abilities. For the 3DCb solution, the first dimension ability estimates were strongly related to both the first and second dimension true abilities, though the correlation was slightly higher for the second dimension. The second and third dimension estimates were both most strongly related to the third dimension true abilities.

Table 10 shows the chi-square, AIC, and CAIC values obtained for each solution for the three-dimensional data. For these data not all chi-square differences could be tested for significance. Table 11 summarizes which pairs of chi-squares could be tested. All testable pairs were significantly different.

Table 9

True and Estimated Ability Parameter Correlations for the Three-Dimensional Data

Solution/Dimension	True Ability Dimension		
	1	2	3
1DU			
1	0.80	0.42	0.33
2DC			
1	0.73	0.03	0.62
2	0.32	0.78	-0.40
2DU			
1	0.57	0.74	-0.19
2	0.58	-0.20	0.72
3DCa			
1	0.87	0.29	0.21
2	0.19	0.85	-0.23
3	0.24	-0.27	0.84
3DCb			
1	0.64	0.70	-0.09
2	0.54	-0.25	0.70
3	0.49	-0.25	0.72

Like the unidimensional case, the chi-square and AIC would result in the same ordering of the models for the three-dimensional data. Using either the absolute magnitude of the chi-square criterion or the AIC, the 3DCa solution would have been selected as best. The unconstrained two-dimensional solution was next, while the constrained two-dimensional solution was third. The 3DCb solution was fourth, and the unidimensional solution was last. Of course, since not all of the pair-wise comparisons are testable, this rank-ordering isn't entirely objective.

Using the CAIC resulted in a slightly different ordering of models. The 3DCa solution was first, as it was using the chi-square and AIC. However, using the CAIC, the constrained two-dimensional solution was second. This ordering is more reasonable than that obtained using the chi-square and AIC, since the constrained two-dimensional solution used an item discrimination parameter that was consistent with the true pattern. The unconstrained two-dimensional solution was third, the 3DCb solution was fourth, and the unidimensional solution was last.

Table 10

Model Selection Criteria Values
for The Three-Dimensional Data

Solution	Chi-Square	AIC	CAIC
1DU	70706.4	84823.8	85769.0
2DC	65828.9	80022.6	81204.1
2DU	65592.6	79871.3	81289.2
3DCa(true)	65450.4	79723.6	81141.5
3DCb	67548.8	81815.3	83233.2

Table 11

Testable Pair-Wise Chi-Square Comparisons
for The Three-Dimensional Data

Solution	1DU	2DC	2DU	3DCa	3DCb
1DU	-	*	*	*	*
2DC		-	*	*	-
2DU			-	-	-
3DCa(true)				-	-

Note. Dash (-) indicates not testable, asterisk (*) indicates testable.

Analysis of Residuals

Table 12 summarizes the results of the residual analyses for both the unidimensional and three-dimensional data. Shown are the first three eigenvalues obtained from a principal components analysis of Pearson product moment correlations computed on the matrix of residuals for each CMIRT solution. Also shown is the percent of variance and cumulative percent of variance corresponding to each eigenvalue.

For the unidimensional data, the results reported in Table 12 indicate no meaningful variation remaining in the residuals. This is consistent with the fact that the data were truly unidimensional. It is interesting to note that increasing the number of parameters estimated did not reduce the size of the first eigenvalue of the residuals to any meaningful degree.

For the three-dimensional data, the pattern is quite different. For these data, increasing the number of estimated parameters noticeably reduced the size of the first eigenvalue, and correctly clustering the items in the 3DCa solution reduced the first eigenvalue to a smaller value than was obtained for the 2DU solution, even though the number of item parameters estimated did not increase. Incorrectly clustering the items in the 3DCb solution, on the other hand, did not produce a smaller first eigenvalue than was obtained for the 2DU solution.

Table 12

Principal Components Analysis of Residuals

Dataset/ Component/ Statistic	Solution				
	1DU	2DC	2DU	3DCa	3DCb
Unidimensional					
1					
Eigenvalue	1.62	1.62	1.61		
% Variance	2.02	2.02	2.02		
Cumulative %	2.02	2.02	2.02		
2					
Eigenvalue	1.56	1.56	1.57		
% Variance	1.97	1.95	1.96		
Cumulative %	3.99	3.97	3.98		
3					
Eigenvalue	1.54	1.55	1.55		
% Variance	1.92	1.94	1.94		
Cumulative %	5.91	5.91	5.92		
Three-dimensional					
1					
Eigenvalue	8.37	2.07	1.95	1.63	1.93
% Variance	10.46	2.58	2.44	2.04	2.42
Cumulative %	10.46	2.58	2.44	2.04	2.42
2					
Eigenvalue	1.96	1.61	1.61	1.61	1.63
% Variance	2.45	2.01	2.02	2.01	2.03
Cumulative %	12.91	4.60	4.45	4.05	4.45
3					
Eigenvalue	1.72	1.57	1.57	1.58	1.58
% Variance	2.15	1.96	1.97	1.98	1.98
Cumulative %	15.07	6.56	6.42	6.02	6.43

Summary and Conclusions

The purpose of this research was to develop and evaluate a confirmatory approach to assessing test structure using multidimensional item response theory. The approach investigated involves adding to the exponent of the MIRT model an item structure matrix that allows the user to specify what ability dimensions are measured by an item. Various combinations of item structures were fit to two sets of simulation data with known true structures, and the results were evaluated using three different model selection criteria and an analysis of residuals procedure. In addition, item and principal components analyses were performed to assess the reasonableness of the data.

The results of the item and principal components analyses tend to support the reasonableness of the CMIRT model. The data generated according to both the unidimensional and three-dimensional models appeared to be realistic with respect to item difficulty and discrimination, and the structure of each test, as revealed by the principal components analysis, was neither unrealistic nor uncommon. The reliabilities of the tests did appear to be a little higher than normally obtained with real data, as did the correlations between the item biserials and difficulties, but these results were most likely a reflection of the purity of the simulated data.

The comparisons among the various solutions derived for each set of data using the three model selection criteria were encouraging. The likelihood ratio chi-square statistic was clearly inadequate, since its significance could not always be tested, and both the chi-square and AIC statistics tended to result in over-parameterization. However, the CAIC criterion appeared to

function quite well. For both the unidimensional and three-dimensional data, the CAIC criterion resulted in selection of the true structure.

In addition to finding that the procedures could recover the true item structures, it was also found that adding an additional ability dimension that forces together items that ought not to be together (the 3DCb solution) noticeably deteriorates the quality of the solution. On the other hand, imposing structures different from, but not inconsistent with, the true structure (the 2D solutions) does not necessarily yield worse fit.

The residual analyses indicated that, for the unidimensional data, adding additional dimensions did not reduce the proportion of common variance remaining in the residuals below what was obtained for the unidimensional solution. For the three-dimensional data, however, adding dimensions did reduce the remaining common variance below what was obtained for the unidimensional solution, and correctly clustering items reduced the remaining common variance below what was obtained when items were incorrectly clustered, even when the number of dimensions (or parameters) did not increase.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & B. F. Csaki (Eds.), Second International Symposium on Information Theory. Academiai Kiado: Budapest.

Akaike, H. (1987). Factor analysis and AIC. Psychometrika, 52, 317-332.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika, 46, 443-459.

Bock, R. D., Gibbons, R., & Muraki, E. (1985). Full-information item factor analysis (MRC Report No. 85-1). Chicago: University of Chicago.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. Psychometrika, 52, 345-370.

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

Carroll, J. B. (1945). The effect of difficulty and chance success on correlations between items or between tests. Psychometrika, 10, 1-19.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. Psychometrika, 40, 5-32.

Divgi, D. (1980, April). Dimensionality of binary items: Use of mixed models. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston.

Haberman, J. S. (1977). Log-linear models and frequency tables with small expected cell counts. Annals of Statistics, 5, 1148-1169.

Kaya-Carton, E. (1988, March). Empirical comparisons of three methods in calibrating items for French reading proficiency levels. Paper presented at the Language Testing Research Colloquium, New York.

Kingston, N. M. (1986). Assessing the dimensionality of the GMAT verbal and quantitative measures using full information factor analysis (ETS Research Report 86-19). Princeton, NJ: Educational Testing Service.

Lewis, C., & Sheehan, K. (in preparation). Using Bayesian decision theory to design a computerized mastery test. Draft Research Report.

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.

McKinley, R. L. (1983, April). A multidimensional extension of the two-parameter logistic latent trait model. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.

McKinley, R. L. (1987). User's guide to MULTIDIM. Princeton: Educational Testing Service.

McKinley, R. L. (in preparation). Exploratory and confirmatory analysis of test structure using multidimensional item response theory models. Draft Research Memorandum.

McKinley, R. L., & Kingston, N. M. (1987). Exploring the use of IRT equating for the GRE Subject Test in Mathematics (ETS Research Report 87-21). Princeton, NJ: Educational Testing Service.

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

Mislevy, R. J., & Bock, R. D. (1985). Implementation of the EM algorithm in the estimation of item parameters: The BILOG computer program. In D. J. Weiss (Ed.), Proceedings of the 1982 item response theory and computerized adaptive testing conference. Minneapolis: University of Minnesota.

Muthen, B. (1978). Contributions to factor analysis of dichotomized variables. Psychometrika, 43, 551-560.

Reckase, M. D. (1981). The formation of homogeneous item sets when guessing is a factor in item responses (Research Report 81-5). Columbia, MO: University of Missouri.

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

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

Reckase, M. D., & McKinley, R. L. (1985). Some latent trait theory in a multidimensional latent space. In D. J. Weiss (Ed.), Proceedings of the 1982 item response theory and computerized adaptive testing conference. Minneapolis: University of Minnesota.

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

Tucker, L. R., Humphreys, L. G., & Roznowski, M. A. (1986). Comparative accuracy of five indices of dimensionality of binary items (Technical Report 1). Urbana: University of Illinois.

Wilson, D., Wood, R., & Gibbons, R. (1984). TESTFACT user's guide. Mooresville, IN: Scientific Software.

Zimowski, M. F., & Bock, R. D. (1987). Full-information item factor analysis of test forms from the ASVAB CAT pool (MRC Report No. 87-1). Chicago: University of Chicago.