DOCUMENT RESUME

ED 385 583 TM 024 021

AUTHOR	McKinley, Robert L.; Way, Walter D.
TITLE	The Feasibility of Modeling Secondary TOEFL Ability Dimensions Using Multidimensional TRT Models.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-92-16; TOEFL-TR-5
PUB DATE	Feb 92
NOTE	31p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Ability; Goodness of Fit; *Identification; *Item Response Theory; Models; *Performance; Second Languages; Test Format
IDENTIFIERS	*Multidimensionality (Tests); Secondary Analysis; *Test of English as a Foreign Language

ABSTRACT

An analysis of the skills necessary for performance on the Test of English as a Foreign Language (TOEFL) tends to support the view that there are important, although subtle, secondary dimensions present in the test. This research explored the feasibility of an item response theory (IRT) based method of modeling examinee performance on these secondary ability dimensions. Both exploratory multidimensional IRT (MIRT) and confirmatory multidimensional IRT (CMIRT) models were investigated in the study. The work performed included the application of unidimensional IRT, MIRT, and CMIRT models in two TOEFL forms to evaluate the extent to which model fit is enhanced by using a multidimensional model and to determine to what extent the additional fitted ability dimensions correspond to meaningful cognitive processes or content areas. Results indicate that the MIRT and CMIRT procedures were successful in modeling secondary ability dimensions on TOEFL and that they provide corroborative evidence in interpreting the structure of the test that is consistent with previous structure interpretations. The data also illustrate how the consistent Akaike information criterion can identify the best competing models of test structure. Four figures (plots) and seven tables illustrate the discussion. (Contains 34 references.) (SLD)

TOEFL®

February 1992

Technical Report

TR- 5

The Feasibility of Modeling Secondary TOEFL Ability Dimensions Using Multidimensional IRT Models

Robert L. McKinley
Walter D. Way

The Feasibility of Modeling
Secondary TOEFL Ability Dimensions
Using Multidimensional IRT Models

Robert L. McKinley
Walter D. Way

Educational Testing Service
Princeton, New Jersey

RR-92-16

The Test of English as a Foreign Language (TOEFL®) was developed in 1963 by a National Council on the Testing of English as a Foreign Language, which was formed through the cooperative effort of more than thirty organizations, public and private, that were concerned with testing the English proficiency of nonnative speakers of the language applying for admission to institutions in the United States. In 1965, Educational Testing Service (ETS) and the College Board assumed joint responsibility for the program, and in 1973, a cooperative arrangement for the operation of the program was entered into by ETS, the College Board, and the Graduate Record Examinations (GRE) Board. The membership of the College Board is composed of schools, colleges, school systems, and educational associations; GRE Board members are associated with graduate education.

ETS administers the TOEFL program under the general direction of a Policy Council that was established by, and is affiliated with, the sponsoring organizations. Members of the Policy Council represent the College Board and the GRE Board and such institutions and agencies as graduate schools of business, junior and community colleges, nonprofit educational exchange agencies, and agencies of the United States government.

A continuing program of research related to the TOEFL test is carried out under the direction of the TOEFL Research Committee. Its six members include representatives of the Policy Council, the TOEFL Committee of Examiners, and distinguished English as a second language specialists from the academic community. Currently the Committee meets twice yearly to review and approve proposals for test-related research and to set guidelines for the entire scope of the TOEFL research program. Members of the Research Committee serve three-year terms at the invitation of the Policy Council; the chair of the committee serves on the Policy Council.

Because the studies are specific to the test and the testing program, most of the actual research is conducted by ETS staff rather than by outside researchers. However, many projects require the cooperation of other institutions, particularly those with programs in the teaching of English as a foreign or second language. Representatives of such programs who are interested in participating in or conducting TOEFL-related research are invited to contact the TOEFL program office. All TOEFL research projects must undergo appropriate ETS review to ascertain that the confidentiality of data will be protected.

Current (1991-92) members of the TOEFL Research Committee are:

James Dean Brown	University of Hawaii
Patricia Dunkel (Chair)	Pennsylvania State University
William Grabe	Northern Arizona University
Kyle Perkins	Southern Illinois University at Carbondale
Elizabeth C. Traugott	Stanford University
John Upshur	Concordia University

ABSTRACT

An analysis of the skills necessary for performance on the TOEFL® test tends to support the view that there are important, although perhaps subtle, secondary dimensions present in the test. Given that these subtle secondary ability dimensions may be present in examinee response data, and that they do represent meaningful psychological variables, the purpose of this research was to explore the feasibility of an IRT-based method of modeling examinee performance on these secondary ability dimensions. The procedure investigated is based on a multidimensional extension of the IRT model currently used for equating the TOEFL test. Both exploratory multidimensional IRT (MIRT) and confirmatory multidimensional IRT (CMIRT) models were investigated in the study. The work performed included the application of unidimensional IRT, MIRT, and CMIRT models to two TOEFL forms to evaluate the extent to which model fit is enhanced by using a multidimensional model, and to determine to what extent the additional fitted ability dimensions correspond to meaningful cognitive processes or content areas.

The results of this study indicate that the MIRT and CMIRT procedures were successful in modeling secondary ability dimensions on TOEFL. The two procedures provided corroborative evidence in interpreting the structure of the test that was consistent with previous interpretations of the test's structure. The data presented in this study also provide an illustration of how a particular criterion for assessing model fit--the consistent Akaike information criterion--can be utilized to identify the best of several competing models of test structure.

TABLE OF CONTENTS

	Page
Introduction	1
Item Response Theory	2
Unidimensional Item Response Theory	2
Multidimensional IRT	2
Exploratory Models	3
Confirmatory Models	3
Method	4
Data	4
Estimation	4
Model Evaluation	4
Analyses	6
Results	7
Sampling	7
Model Fit - Exploratory Analyses	8
Model Selection Criteria	8
Analysis of Residuals	9
Model Fit - Confirmatory Analyses	10
Model Selection Criteria	10
Analysis of Residuals	12
Summary of Model Fit	13
Interpretability of Results	13
Exploratory Results	14
Confirmatory Results	19
Discussion	19
Conclusions	20
References	21

LIST OF TABLES

		Page
Table 1	Summary of CMIRT Target Test Structures	7
Table 2	Number-Correct Score Summary Statistics by Form and Sample	8
Table 3	Exploratory Analysis Model Selection Criteria	9
Table 4	Summary of the Principal Components Analysis of Residuals for the Exploratory Analyses	10
Table 5	Confirmatory Analysis Model Selection Criteria	11
Table 6	Summary of the Principal Components Analysis of Residuals for the Confirmatory Analyses	13
Table 7	Means and Standard Deviations of Item Parameter Estimates by Content Area for Confirmatory Solution 3DC1	19

LIST OF FIGURES

		Page
Figure 1	Plots of 3-Dimensional A-Values — Form JM	15
Figure 2	Plots of 3-Dimensional A-Values — Form JP	16
Figure 3	Plots of 3-Dimensional A-Values — Form K9	17
Figure 4	Plots of 3-Dimensional A-Values — Form KA	18

INTRODUCTION

Factor analytic research on the Test of English as a Foreign Language (TOEFL) seems to lead to the conclusion that the test measures primarily one factor. For example, in a factor analytic study of seven different language groups, Swinton and Powers (1980) obtained first-to-second eigenvalue ratios ranging from 7.1 to 9.3, depending on the language group. When the language groups were combined, a ratio of 8.9 was obtained. Similar results were obtained by Hale, Stansfield, Rock, Hicks, Butler, and Oller (1988) using both factor analytic and item response theory (IRT) based methods for assessing dimensionality, by Dunbar (1982) and Hale, Rock, and Jirele (1989), who used confirmatory factor analysis approaches, and by Boldt (1988) using latent structure analysis.

However, an analysis of the skills necessary for performance on the TOEFL test tends to support the view that there are important, although perhaps subtle, secondary dimensions present in the test (Duran, Canale, Penfield, Stansfield, & Liskin-Gasparro, 1985). Note, for instance, that the test is constructed to have five content components: listening comprehension, structure, written expression, vocabulary, and reading comprehension. Each section includes items measuring a variety of content subareas. For example, the listening comprehension section includes items requiring knowledge of syntax, lexical items, and items focusing on phonology, stress, and intonation. Moreover, there are a variety of required skills in common to items, not only within sections but also across sections (Duran et al., 1985). Clearly, then, the TOEFL test includes items measuring a variety of content areas and cognitive processes. Consequently, it would seem reasonable to expect to find at least some empirical evidence of these dimensions in examinee response data. To some extent, in fact, this expectation is borne out by the research cited above.

For instance, Swinton and Powers (1980) concluded that there was evidence of three factors underlying examinee performance on the TOEFL test, although the interpretation of the factors tended to vary with native language and ability level. Hale et al. (1988) concluded that there was evidence of two factors, one related to listening comprehension, and one related to the remainder of the test. A latter study by Hale, Rock, and Jirele (1989) suggested a consistent two-factor structure of the TOEFL test across several language groups. Dunbar (1982) found evidence of four factors: one general factor and one secondary factor associated with each of the three TOEFL sections. Oltman, Stricker, and Barrows (1988) used a three-way multidimensional scaling approach to examine the effects of native language and English proficiency on the structure of the TOEFL test and found evidence of three dimensions that corresponded to the sections of the examination, and a fourth that was identified as an "end-of-test phenomenon."

Given that these subtle secondary ability dimensions may be present in examinee response data, and that they do represent meaningful psychological variables, it would be useful to have an IRT-based procedure that could (1) confirm that the such dimensions are or are not present in a particular form of the test and (2) extract information about these abilities, when present, for individual examinees. Such a procedure could yield valuable dividends. For instance, it might be possible to use such a procedure to provide meaningful feedback to examinees regarding their performance on specific content areas and item types. In fact, the ability to provide useful diagnostics might be enhanced by providing feedback from the procedure to test developers, who could use the information as a guide to test construction.

The purpose of this research was to explore the feasibility of an IRT- based method of modeling examinee performance on these secondary dimensions. The procedure investigated is based on a multidimensional extension of the IRT model currently used for equating the TOEFL test. The work performed included the application of both unidimensional and multidimensional IRT models to two forms of the TOEFL test to evaluate the extent to which model fit is enhanced by using various multidimensional models, and to determine to what extent the additional fitted ability dimensions correspond to test content.

Item Response Theory

Unidimensional Item Response Theory

In recent years, item response theory has become a very popular tool both in research and in practical measurement applications. The attractiveness of IRT derives primarily from its parameter invariance properties and the availability of well-defined standard errors of estimate (Bock & Aitkin, 1981; Lord, 1980; Lord & Novick, 1968). Moreover, the ability estimate standard errors are expressed as functions of ability. Thus, the standard error of estimate is reported for each level of ability rather than for a test as a whole, as is the case with more traditional measurement procedures. As a consequence, not only does the use of IRT improve the quality of measurement, but it makes possible applications that would be prohibitively difficult or impossible with more traditional measurement procedures.

Among the item response theory models that have been developed are the two-parameter normal ogive model (Lord, 1952); the two-parameter logistic model (Birnbaum, 1958); the one-parameter logistic, or Rasch, model (Rasch, 1960); and the three-parameter logistic model (Birnbaum, 1968). By far the most widely used of these models are the one-parameter logistic (1PL) model and the three-parameter logistic (3PL) model. The model used for equating the TOEFL is the 3PL model, which was used in this research.

The 3PL model is given by

$$P_i(\theta_j) = c_i+(1-c_i)/(1.0+\exp(-Da_i(\theta_j-b_i))) \quad , \quad (1)$$

where c_i is the pseudo-guessing parameter for item i; a_i is the item discrimination parameter for item i; b_i is the item difficulty parameter for item i; and Θ_j is the ability parameter for examinee j. The D in the exponent is equal to 1.702, and is included to make the logistic curve more closely approximate what would be obtained using a normal ogive model.

Although item response theory 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 that include items sampled from several content areas, such as a science test containing both physics and chemistry items, are probably not sufficiently homogeneous to permit analysis using IRT. Such may also be the case with tests containing multifaceted items, such as a language test containing English structure items requiring a high level of reading comprehension or vocabulary skill.

In recent years, attempts have been made to extend IRT to the case of 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.

Multidimensional IRT

As with factor analysis, there are two basic approaches to MIRT analysis--exploratory and confirmatory. In exploratory procedures, the emphasis is on discovering the best fitting model, while in confirmatory approaches the focus is on evaluating the extent to which the data follow a hypothetical model developed a priori on the basis of content and process analysis of the instrument to be analyzed. Both approaches were examined in this research.

Exploratory Models

Most of the recent progress in MIRT research has occurred in two areas--the development of a multidimensional two-parameter normal ogive nonlinear factor analysis model (Bock, Gibbons, & Muraki, 1985), and the development of multidimensional two- and three-parameter logistic IRT models (McKinley, 1983, 1987). Work on these procedures is still at an early stage, but it has progressed to the point that estimation procedures are available. For the nonlinear factor analysis procedure, the TESTFACT program (Wilson, Wood, & Gibbons, 1984) is available. For the two-parameter logistic IRT model, the MAXLOG program (McKinley & Reckase, 1983) is available, and for the three-parameter model, the MULTIDIM program (McKinley, 1987) is available.

For this research, the MIRT model employed was the multidimensional three-parameter logistic (M3PL) model. This model was selected for two reasons: microcomputer-based estimation procedures are available for use with the M3PL model, and it is closely related to the model currently used with the TOEFL test.

The M3PL model is given by

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

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

Uses of the M3PL model thus far have been somewhat limited, primarily due to the recency of the development of estimation procedures for applying the model. The model was, however, applied to a French proficiency exam by Kaya-Carton (1988). In this application, parameters of the M2PL model were obtained using the MULTIDIM program (item lower asymptote parameters were 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.

Confirmatory Models

As mentioned above, the goal of confirmatory MIRT, or CMIRT, analysis is to confirm or disconfirm the presence of some hypothesized test structure. In CMIRT, competing hypothesized models are applied to data and compared using some measure of model fit. The CMIRT procedure used in this research was based on a modification of the M3PL model.

Adaptation of the M3PL model for use in confirmatory analysis consists of imposing a set of constraints on the item discrimination parameters in accordance with an a priori target test structure. As an example, consider a six-item English test containing three vocabulary items and three reading comprehension items. One possible structure for such a test might be given by

$$\underline{S} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \quad , \qquad (3)$$

where rows are dimensions and columns are items. A 1 indicates that an item discrimination parameter will be estimated for the item on that dimension, and a 0 indicates that the item discrimination parameter on that dimension will be constrained to 0.0. Thus, in this example, the first dimension corresponds to a general latent trait, the second dimension represents a trait specific to the first three items, and the third dimension represents a trait specific to the last three items. (For computational details about CMIRT procedures, see McKinley, 1988, 1989.)

Like the M3PL MIRT model, CMIRT models have not been widely used due to the recency of their development. In one application reported by Kingston and McKinley (1988), CMIRT procedures were applied to the Graduate Record Examination's General Test. Results, which were consistent with previous research on that test, indicated the presence of a general dimension, a verbal dimension, a quantitative dimension, and a weak analytical dimension defined by logical reasoning items (but not analytical reasoning items).

METHOD

Data

Data for this study comprised the responses to the 146 operational TOEFL items for the November 1987 and May 1988 administrations. Responses were sampled for 2,500 randomly selected examinees from domestic test centers and for 2,500 randomly selected examinees from foreign test centers. Sampling was performed by selecting all candidates with low and high number-correct scores and equal numbers of candidates at each score level in between.[1] Foreign and domestic examinees were analyzed separately, as were the two forms, as a means of cross-validating the results.

Estimation

For all analyses in this study, solutions were based on the M3PL model, except that the c-parameter was not estimated. Rather, it was held fixed at a value of 0.2. The parameters of the model were estimated using an EM algorithm based procedure similar to those described by Bock and Aitkin (1981), Mislevy and Bock (1985), and Bock, Gibbons, and Muraki (1985). The algorithm has been implemented in the MULTIDIM program (McKinley, 1987), which is designed for exploratory MIRT analysis, and in the CONFIRM program (McKinley, 1989), which is designed for both exploratory and confirmatory analyses.

In this algorithm, item parameter estimation is performed using a two-step marginal maximum likelihood procedure. Multiple latent abilities are hypothesized, each is treated as a random variable, and integration over the joint distributions of these random variables is performed. The integration over the ability distribution, accomplished through numerical quadrature, is performed during the first step, the E (expectation) step, and produces an expected sample size and number-correct score at each quadrature node for each item. These values are used in the second step, the M (maximization) step, to perform marginal maximum likelihood item parameter estimation.

Model Evaluation

In MIRT, evaluation of model-data fit is relatively straightforward. Solutions for 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 a chi-square procedure such as is implemented in TESTFACT (Wilson, Wood, & Gibbons, 1984).

---

[1]This sampling procedure was used to increase the sensitivity of the analyses, particularly at lower levels of ability. Although this procedure may have been inconsistent with the assumptions of the marginal maximum likelihood procedure, the effects of the sampling procedure on the results of the model estimations are difficult to predict, as the relationship that exists between observed score distributions and latent ability distributions is complex. It is recognized that random examinee samples may have yielded different results.

For example, assume that one- and two-dimensional MIRT solutions have been obtained on the same data using a multidimensional 2PL 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

$$G^2 = 2 \sum_{j=1}^{J} r_j \ln(r_j/NP_j) \qquad , \qquad (4)$$

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

$$P_j = \sum_{k=1}^{q} L_j(\underline{x}_k)W_k \qquad , \qquad (5)$$

where P_j represents the marginal 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 4 are given by

$$df = 2^n - n(m+2) \qquad , \qquad (6)$$

where n is the number of items and m is the number of dimensions. If the c-parameter is not estimated, then the second term on the right is n(m+1). For CMIRT models it is necessary to reduce Equation 6 by the number of parameters constrained to 0.0.

While it is doubtful that the statistic given by Equation 4 is actually distributed as a chi-square, the difference between the values of G^2 for subsuming models has been shown to be asymptotically distributed as chi-square (Haberman, 1977). The degrees of freedom for the difference between two values of G^2 is equal to the difference between Equation 6 for the two solutions. For IRT models, this equals the difference in the number of item parameters estimated.

Another way in which two competing models of test structure can 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 log 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. This approach is particularly useful in the context of CMIRT analysis, since competing models often are not subsuming and therefore cannot be compared using the chi-square procedure described above.

The AIC statistic is given by

$$AIC = -2 \log(L) + 2k \qquad , \qquad (7)$$

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

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

The CAIC statistic is given by

$$CAIC = -2 \log(L) + k(\log(N)+1) \quad , \qquad (8)$$

where N is the sample size. This modification of the AIC has the effect of increasing the penalty for overparameterization and, consequently, tends to lead to the selection of simpler models. All three statistics were used in this research.

Analyses

The first analysis performed in this study was to fit a unidimensional IRT model to both the foreign test center data and the data from the domestic test centers, and to evaluate the goodness-of-fit of the model for both sets of data using the chi-square, AIC, and CAIC statistics. Following this, residuals were computed and analyzed, using a 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

$$R_{ij} = P_{ij} - u_{ij} \quad , \qquad (9)$$

where P_{ij} is the probability of the observed response to item i by examinee j predicted from the IRT model, and u_{ij} is the observed response. Residual correlation matrices were then analyzed using principal components analysis. The advantage of this procedure is that residuals are on a continuous scale, which reduces the potential problems that may occur when principal components are applied to interitem phi or tetrachoric correlations.

After this, two-, three-, and four-dimensional MIRT models were fit to both datasets, and the goodness-of-fit evaluated. Item discrimination vectors were examined to determine whether the fitted ability dimensions corresponded to content areas or cognitive processes. Residual analyses were also performed using Divgi's (1980) procedure.

Finally, the CMIRT procedure was used to impose several hypothesized test structures on the two sets of data. CMIRT solutions were obtained for one 2-dimensional structure, three 3-dimensional structures, and two 4-dimensional structures. These target structures are summarized in Table 1, which indicates the dimensions for which discrimination parameters were estimated in a given CMIRT solution. The abbreviations for the test structures used in Table 1 will be used throughout the report. For example, 3DC3 will refer to the three-dimensional solution (3D), configuration 3 (C3).

TABLE 1 Summary of CMIRT Target Test Structures

Section /Part	No. Items	2 Dim. 2DC1	3 Dim. 3DC1	3DC2	3DC3	4 Dim. 4DC1	4DC2
1. Listening Comp.	50						
Statements	20	1,2	1,2	1,2	1,2	1,2	1,2
Dialogues	15	1,2	1,2	1,2	1,2	1,2	1,2
Minitalks	15	1,2	1,2	1,2	1,2	1,2	1,2
2. Structure & Written Exp.	38						
Structure	14	1	1,3	1,3	1,3	1,3	1,3
Written Exp.	24	1	1,3	1,3	1,3	1,3	1,3
3. Vocabulary & Reading Comp.	58						
Vocabulary	29	1	1,3	1,3	1,3	1,4	1,3
Reading Comp.	29	1	1,3	1,2	1	1,4	1,4

Of the above structures, the 2DC1, 3DC1, and 4DC1 solutions are most consistent with TOEFL structures that have been suggested by previous research. In each of these solutions, a discrimination parameter is estimated for each item on the first dimension, which can be thought of as a general factor. In Solution 2DC1, the second dimension is constrained so that discrimination parameters are estimated only for listening comprehension items. In Solution 3DC1, the second dimension is constrained as in Solution 2DC1, while discrimination parameters on the third dimension are estimated for structure and written expression and vocabulary and reading comprehension items. In Solution 4DC1, discrimination parameter estimates are obtained for the items in each of the three TOEFL test sections on three distinct secondary dimensions. Solutions 3DC2, 3DC3, and 4DC2 were hypothesized to allow reading comprehension items to measure different abilities from vocabulary items. For example, in Solution 3DC2, reading comprehension items were hypothesized to measure the same ability dimension as listening comprehension items, whereas the vocabulary items were hypothesized to measure the same ability dimension as the structure and written expression items. In Solution 3DC3, reading comprehension was hypothesized to be measured only by the general factor. In Solution 4DC2, reading comprehension items were hypothesized to measure a distinct secondary ability dimension. To evaluate the various CMIRT structures, the goodness-of-fit of the different CMIRT solutions were compared to each other and to the MIRT solutions. Residual analyses were also performed.

RESULTS

Sampling

The number-correct score means and standard deviations, along with sample sizes and KR-20 reliability estimates, for the four sets of data sampled for this study are shown in Table 2. Although the four sets of data were similar with regard to these statistics, there are some important differences. Note, for example, that the number-correct score means are somewhat higher for the samples of domestic examinees and the standard deviations are somewhat lower. Despite the nonrandom sampling procedures, these results are consistent with results typically seen on the operational TOEFL test.

TABLE 2 Number-Correct Score Summary Statistics by Form and Sample

	Statistic			
Form/Sample	N	Mean	S. D.	KR-20
November 1987				
Foreign (JM)	2,538	90.2	32.1	0.98
Domestic (JP)	2,551	92.2	30.9	0.97
May 1988				
Foreign (K9)	2,537	87.7	33.6	0.98
Domestic (KA)	2,509	90.3	31.7	0.98

It can also be seen from Table 2 that the scores were somewhat higher and less variable for the November 1987 samples (hereinafter referred to as JM for the foreign examinee sample and JP for the domestic sample) than for the May 1988 samples (K9 for the foreign sample and KA for the domestic sample). These differences are likely a reflection of differences in average item difficulty between the two test forms, rather than an indication of differences in group ability.

Model Fit--Exploratory Analyses

Model Selection Criteria

Table 3 summarizes the model selection criteria for the exploratory MIRT analyses performed on the four sets of data. For each test form and examinee sample, Table 3 shows the likelihood ratio chi-square, AIC, and CAIC statistics for the unconstrained one-dimensional (1D), two-dimensional (2D), three-dimensional (3D), and four dimensional (4D) solutions.

As the data in Table 3 indicate, the relative ordering of solutions was consistent across for the chi-square and the AIC selection criteria: for all data sets, the 4D solution would be considered optimal, followed, in order, by the 3D, 2D, and 1D solutions. In each sample, the differences between chi-square statistics were testable and were found to be statistically significant. Across solutions, the decreases in the chi-square and AIC statistics from the 1D to the 2D solutions were largest, and the decreases in these statistics from the 3D to the 4D solutions were relatively small.

Using the CAIC statistics, however, results in different orderings of the exploratory solutions. For each data sample, the CAIC statistics indicate that the 3D solutions are optimal. For two of the samples (JM and KA), the CAIC statistic is lower for the 2D solution than it is for the 4D solution. The differences between the CAIC results and those obtained for the chi-square and AIC statistics are clearly due to the penalty for overparameterization that is incorporated into the CAIC statistic (see Equation 8). McKinley (1989) points out that both the CAIC and AIC statistics essentially embody a "critical value" for testing whether a particular model is the best fitting one. Selecting the CAIC statistic over the AIC statistic is equivalent to selecting a larger critical value, which reduces the Type I error rate. In fact, the main advantage of the CAIC statistic is that the Type I error rate decreases exponentially with increased sample size. Asymptotically, Type I error for the CAIC statistic goes to zero (Bozdogan, 1987).

TABLE 3 Exploratory Analysis Model Selection Criteria

Sample	Number of Dimensions	Criterion		
		Chi-Square	AIC	CAIC
JM				
	1D	334251.6	374618.8	376615.8
	2D	327455.7	368114.8	371110.4
	3D	325720.6	366671.7	370665.8
	4D	325112.5	366355.6	371348.2
JP				
	1D	332826.3	373415.0	375413.5
	2D	328494.5	369375.2	372373.0
	3D	326372.6	367545.3	371542.3
	4D	325445.9	366910.5	371906.8
K9				
	1D	331088.4	371407.0	373403.9
	2D	324155.5	364766.1	367761.5
	3D	321895.4	362798.0	366791.8
	4D	320813.6	362008.2	367000.5
KA				
	1D	329103.0	368949.4	370943.1
	2D	325507.5	365645.9	368636.4
	3D	323915.3	364345.8	368333.1
	4D	323093.3	363815.8	368800.0

Analysis of Residuals

Table 4 provides a summary of the analysis of residuals performed on each MIRT solution. For each examinee sample for each test form, Table 4 provides the first three eigenvalues and the percentage of variance accounted for by each from a principal components analysis of Pearson correlations computed on residuals. Also shown are the ratios of the first to second and second to third eigenvalues.

TABLE 4 Summary of the Principal Components Analysis of Residuals for the Exploratory Analyses

Sample/ Number of Dimensions	Eigenvalue (% of Variance)			Ratios	
	E_1	E_2	E_3	E_1/E_2	E_2/E_3
JM					
1D	5.03(3.5)	2.31(1.6)	2.00(1.4)	2.18	1.16
2D	2.35(1.6)	2.06(1.4)	1.83(1.3)	1.14	1.13
3D	2.13(1.5)	1.86(1.3)	1.78(1.2)	1.15	1.04
4D	1.97(1.3)	1.85(1.3)	1.73(1.2)	1.07	1.07
JP					
1D	3.54(2.4)	2.94(2.0)	2.06(1.4)	1.21	1.43
2D	2.96(2.0)	2.08(1.4)	1.77(1.2)	1.42	1.18
3D	2.11(1.4)	1.79(1.2)	1.71(1.2)	1.18	1.05
4D	1.80(1.2)	1.74(1.2)	1.66(1.1)	1.03	1.05
K9					
1D	4.96(3.4)	2.88(2.0)	2.39(1.6)	1.72	1.21
2D	2.91(2.0)	2.42(1.7)	2.03(1.4)	1.20	1.19
3D	2.40(1.6)	2.06(1.4)	1.87(1.3)	1.17	1.10
4D	2.03(1.4)	1.89(1.3)	1.79(1.2)	1.07	1.05
KA					
1D	3.18(2.2)	2.51(1.7)	2.18(1.5)	1.26	1.15
2D	2.56(1.8)	2.22(1.5)	1.79(1.2)	1.15	1.24
3D	2.24(1.5)	1.79(1.2)	1.68(1.2)	1.25	1.07
4D	1.81(1.2)	1.72(1.2)	1.65(1.1)	1.05	1.04

For the 4D solution, the results reported in Table 4 indicate essentially no meaningful variation remaining in the residuals. For all data sets, increasing the number of parameters estimated reduced the magnitudes of all three eigenvalues, with the most pronounced reduction occurring when two discrimination parameters were estimated instead of one. However, other than the changes apparent in going from the 1D to 2D solutions, there are no consistent trends in these data across samples. For the 1D data, the magnitudes of the first eigenvalues were appreciably greater in the foreign data samples (forms JM and K9) than in the domestic data samples. This suggests that departures from unidimensionality were perhaps more severe for the foreign samples than for the domestic samples.

Model Fit-Confirmatory Analyses

Model Selection Criteria

Table 5 summarizes the model selection criteria for the confirmatory MIRT analyses. For each test form and examinee sample, Table 5 shows the likelihood ratio chi-square, AIC, and CAIC statistics for each of the hypothesized test structures.

TABLE 5 Confirmatory Analysis Model Selection Criteria

		Criterion		
Sample	Solution	Chi-Square	AIC	CAIC
JM				
	2DC1	329812.3	370279.5	372618.4
	3DC1	327340.4	367999.5	370995.1
	3DC2	328002.3	368661.4	371656.9
	3DC3	328383.9	368985.1	371782.3
	4DC1	328963.5	369622.6	372618.1
	4DC2	328463.5	369122.6	372118.1
JP				
	2DC1	331262.0	371950.7	374291.4
	3DC1	329167.7	370048.4	373046.1
	3DC2	329787.2	370667.9	373665.7
	3DC3	330721.3	371543.9	374343.2
	4DC1	330754.0	371634.6	374632.4
	4DC2	330739.1	371619.8	374617.6
K9				
	2DC1	327098.7	367517.3	369856.1
	3DC1	323674.9	364285.5	367280.8
	3DC2	324199.2	364809.8	367805.2
	3DC3	324640.1	365192.7	367987.7
	4DC1	325520.4	366131.0	369126.4
	4DC2	324000.2	364610.8	367606.2
KA				
	2DC1	327987.3	367933.8	370268.8
	3DC1	325888.2	366026.7	369017.2
	3DC2	326501.9	366640.4	369630.9
	3DC3	326989.9	367070.3	369867.8
	4DC1	327196.4	367334.9	370325.4
	4DC2	327001.7	367140.2	370130.7

Unlike the exploratory analyses, where the optimal solution differed according to the different model selection criteria, in the confirmatory analyses the chi-square, AIC, and CAIC criteria all indicated that Solution 3DC1 was optimal for each sample. In fact, the chi-square and AIC statistics resulted in the same ordering of solutions for each sample, and only for Sample JP was the ordering obtained using the CAIC statistic different from those obtained using the chi-square and AIC statistics.

For the most part, the model selection criteria indicated that the 3D confirmatory solutions were preferable to the 4D confirmatory solutions. This was somewhat surprising, as Solution 4DC1 was most consistent with the current configuration of the TOEFL test sections. Solution 3DC1 suggests that the TOEFL test is characterized by a general dimension, a dimension associated with listening comprehension, and a dimension associated with the other sections of the test. This solution is consistent with interpretations of the TOEFL test structure suggested by Hale et al. (1988) and Hale, Rock, and Jirele (1989).

Comparing the results in Table 5 to those in Table 3, it can be seen that for the foreign samples (JM and K9), the CAIC statistic for Solution 3DC1 was lower than all the CAIC statistics except those for the 3D exploratory solutions. However, for the domestic samples (JP and KA), the CAIC statistics for the 2D, 3D, and 4D exploratory solutions were all lower than the CAIC statistics for any of the confirmatory solutions. A possible explanation for this result is that there is probably a more salient distinction between the listening comprehension section of the TOEFL test and the other sections for foreign examinees than for domestic examinees. Foreign examinees typically have more trouble with the listening comprehension section of the test than they do with the other sections. Domestic examinees tend to perform better on listening comprehension relative to the other sections on the test because they have had more opportunities to listen to spoken English in natural settings.

Analysis of Residuals

Table 6 provides a summary of the analysis of residuals performed on each CMIRT solution. For each examinee sample for each test form, Table 6 provides the first three eigenvalues and the percentage of variance accounted for by each from a principal components analysis of Pearson correlations computed on residuals. Also shown are the ratios of the first to second and second to third eigenvalues.

The data in Table 6 indicate little variation in the magnitudes of the eigenvalues across the 3D and 4D confirmatory solutions. In general, at least one of the three eigenvalues for each of the 2D solutions tends to be higher than the corresponding eigenvalues for the 3D and 4D solutions.

TABLE 6 Summary of the Principal Components Analysis of Residuals for the Confirmatory Analyses

Sample/ Number of Dimensions	Eigenvalue (% of Variance)			Ratios	
	E_1	E_2	E_3	E_1/E_2	E_2/E_3
JM					
2DC1	2.41(1.7)	2.04(1.4)	1.85(1.3)	1.18	1.10
3DC1	2.20(1.5)	2.07(1.4)	1.82(1.3)	1.06	1.14
3DC2	2.16(1.5)	1.97(1.3)	1.84(1.3)	1.10	1.07
3DC3	2.16(1.5)	1.93(1.3)	1.82(1.2)	1.12	1.06
4DC1	2.18(1.5)	1.95(1.3)	1.85(1.3)	1.11	1.06
4DC2	2.10(1.4)	1.92(1.3)	1.81(1.2)	1.10	1.06
JP					
2DC1	3.23(2.2)	2.09(1.4)	1.78(1.2)	1.54	1.17
3DC1	2.33(1.6)	1.85(1.3)	1.82(1.3)	1.26	1.02
3DC2	2.67(1.8)	1.88(1.3)	1.80(1.2)	1.42	1.04
3DC3	2.87(2.0)	1.97(1.3)	1.81(1.2)	1.46	1.09
4DC1	2.60(1.8)	1.93(1.3)	1.79(1.2)	1.35	1.08
4DC2	2.70(1.8)	1.88(1.3)	1.80(1.2)	1.43	1.05
K9					
2DC1	3.11(2.1)	2.41(1.7)	2.07(1.4)	1.29	1.16
3DC1	2.42(1.7)	2.17(1.5)	2.08(1.4)	1.12	1.04
3DC2	2.62(1.8)	2.19(1.5)	2.08(1.4)	1.20	1.05
3DC3	2.58(1.8)	2.13(1.5)	2.08(1.4)	1.21	1.03
4DC1	2.34(1.6)	2.19(1.5)	1.85(1.3)	1.07	1.18
4DC2	2.41(1.7)	2.09(1.4)	1.84(1.3)	1.15	1.13
KA					
2DC1	2.55(1.8)	2.23(1.5)	1.79(1.2)	1.15	1.25
3DC1	2.39(1.6)	2.03(1.4)	1.79(1.2)	1.18	1.13
3DC2	2.57(1.8)	1.80(1.2)	1.77(1.2)	1.43	1.02
3DC3	2.53(1.7)	1.83(1.3)	1.79(1.2)	1.39	1.02
4DC1	2.49(1.7)	1.99(1.4)	1.75(1.2)	1.25	1.14
4DC2	2.40(1.6)	1.81(1.2)	1.78(1.2)	1.32	1.02

Summary of Model Fit

For both the exploratory and confirmatory analyses carried out in this study, the model selection criteria suggested that the 4D interpretations of the TOEFL test were not optimal. The CAIC statistics for the exploratory 4D solutions were higher than the CAIC statistics for 3D solutions and, in some samples, were higher than the 2D CAIC statistics. The principal components analyses of residuals were less conclusive, but they did not indicate that the 4D solutions were clearly preferable to the 2D and 3D solutions. In the confirmatory analyses, all of the model selection criteria indicated that Solution 3DC1 was preferable to both the 4DC1 and 4DC2 solutions.

Interpretability of Results

As in the case of factor analysis, exploratory MIRT solutions are subject to rotational indeterminancy. Therefore, extreme caution should be used in examining the unrotated multidimensional item and ability parameter estimates for the exploratory solutions. Further, comparisons between solutions for the foreign and domestic samples for the same test forms (i.e., comparisons between solutions for Forms JM and JP, and between solutions for Forms K9 and KA) are problematic without equating, in the case of both the exploratory and confirmatory solutions. Because of time and budget constraints, no rotations or equatings of solutions were attempted in the context of this study. To provide some interpretation, however, graphical representations of the exploratory 3D solutions were inspected to determine if patterns in the multidimensional a-parameter estimates were related to content areas of the test. In addition, means and standard deviations of item parameter estimates for the confirmatory solution 3DC1 were calculated by TOEFL test section.

Exploratory Results

Bivariate plots of the 3D item discrimination estimates for each form and sample are provided in Figures 1 through 4. Each figure has three plots: the a_2-estimates graphed against the a_1-estimates, the a_3-estimates graphed against the a_1-estimates, and the a_3-estimates graphed against the a_2-estimates. In each plot, the items in Section 1 (S1) are represented by squares, the items in Section 2 (S2) by plusses, the vocabulary items in Section 3 (S3-Voc) by diamonds, and the reading comprehension items in Section 3 (S3-RC) by triangles.

In Figure 1, the plot of the a_2-estimates against the a_1-estimates (top graph) indicates a separate clustering of the Section 1 items from the items in Sections 2 and 3. A similar clustering can be seen in the plot of the a_3-estimates against the a_2-estimates (bottom graph), suggesting that the second ability dimension in this solution primarily measures listening. The plots involving the a_3-estimates (particularly the middle graph) indicate that the reading comprehension items of Section 3 tend to have the highest discrimination values on this dimension.

In Figure 2, the plot of the a_2-estimates against the a_1-estimates (top graph) for Form JP is very similar to the plot seen in Figure 1 for Form JM. Again, the Section 1 items cluster distinctly from the Section 2 and 3 items. However, contrary to the plots in Figure 1, there appears to be no content-related pattern apparent in the a_3-estimates (middle and bottom graphs).

In Figure 3, the observed patterns are similar to those seen in Figure 1, except that the roles of the a_2- and a_3-estimates are reversed. That is, the a_3-estimates (rather than the a_2-estimates as in Figure 1) are clearly higher for the Section 1 items compared to the Section 2 and 3 items (middle and bottom graphs). Similarly, the highest a_2-estimates (rather than a_3-estimates as in Figure 1) are seen almost exclusively for reading comprehension items (top graph). These differences between Figures 1 and 3 are insignificant, as the dimensions defined by the a_2- and a_3-estimates are arbitrary because of the rotational indeterminancy of the MIRT solutions. In Figure 4, the clustering in the plot of the a_2-estimates against the a_1-estimates again appears to separate the Section 1 items from the Section 2 and 3 items. In this case, the a_2-estimates are clearly higher for the Section 2 and 3 items than for the Section 1 items. As was the case in Figure 2, the plots of the a_3-estimates against the a_1- and a_2-estimates did not suggest any meaningful content-related pattern.

In summary, Figures 1 through 4 did provide support for considering the listening comprehension section as measuring a latent ability distinct from the ability measured by the remaining sections of the TOEFL test. In addition, Figures 1 through 4 suggested that the pattern of the multidimensional item parameter estimates for the reading comprehension items in Section 3 differed in the foreign and domestic samples. The reasons for this are not clear, and may warrant further investigation, particularly as the TOEFL program is currently researching the possibility of revising Section 3 to eliminate discrete vocabulary items.

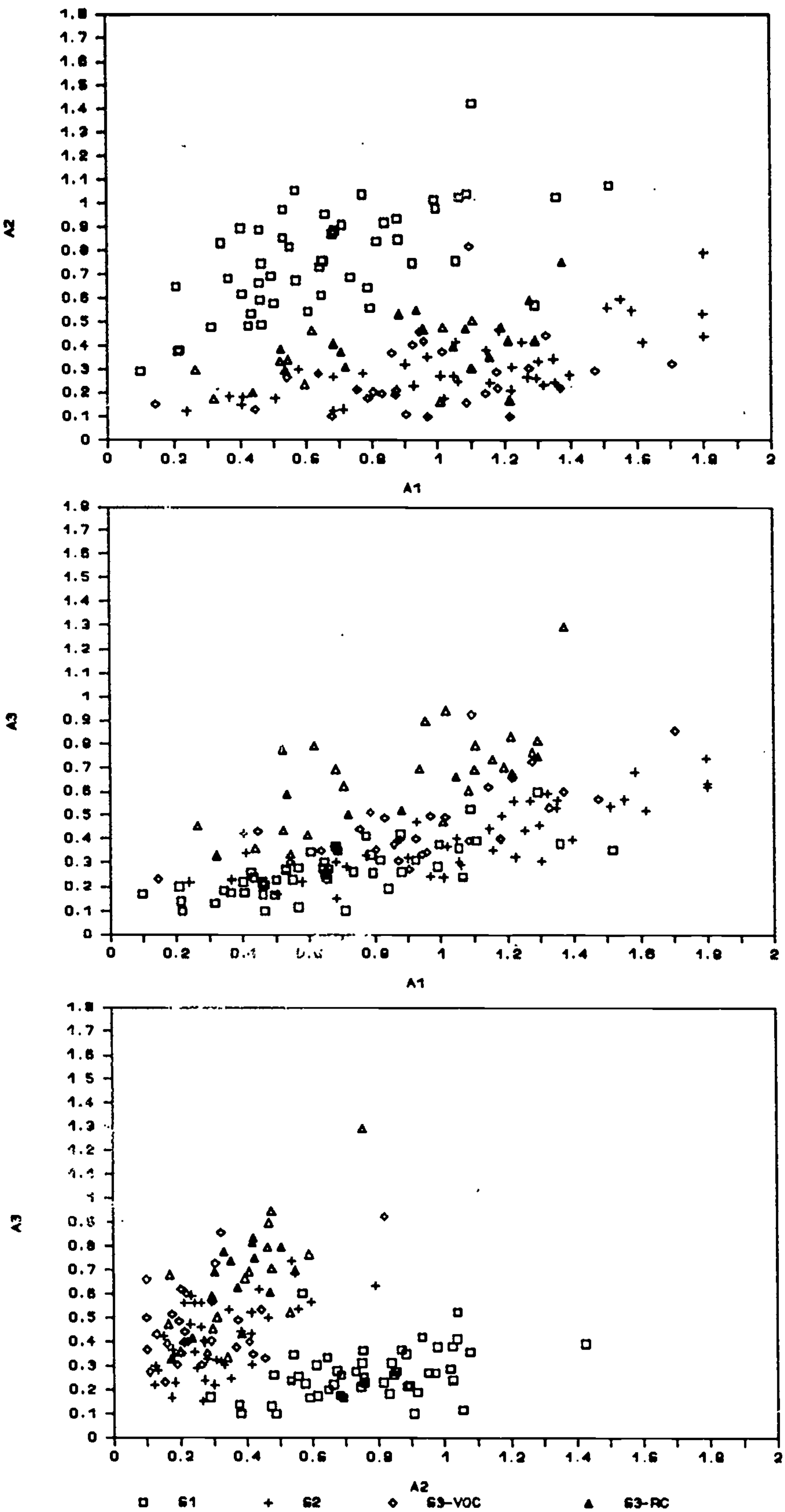

Figure 1: Plots of 3-Dimensional A-Values - Form JM

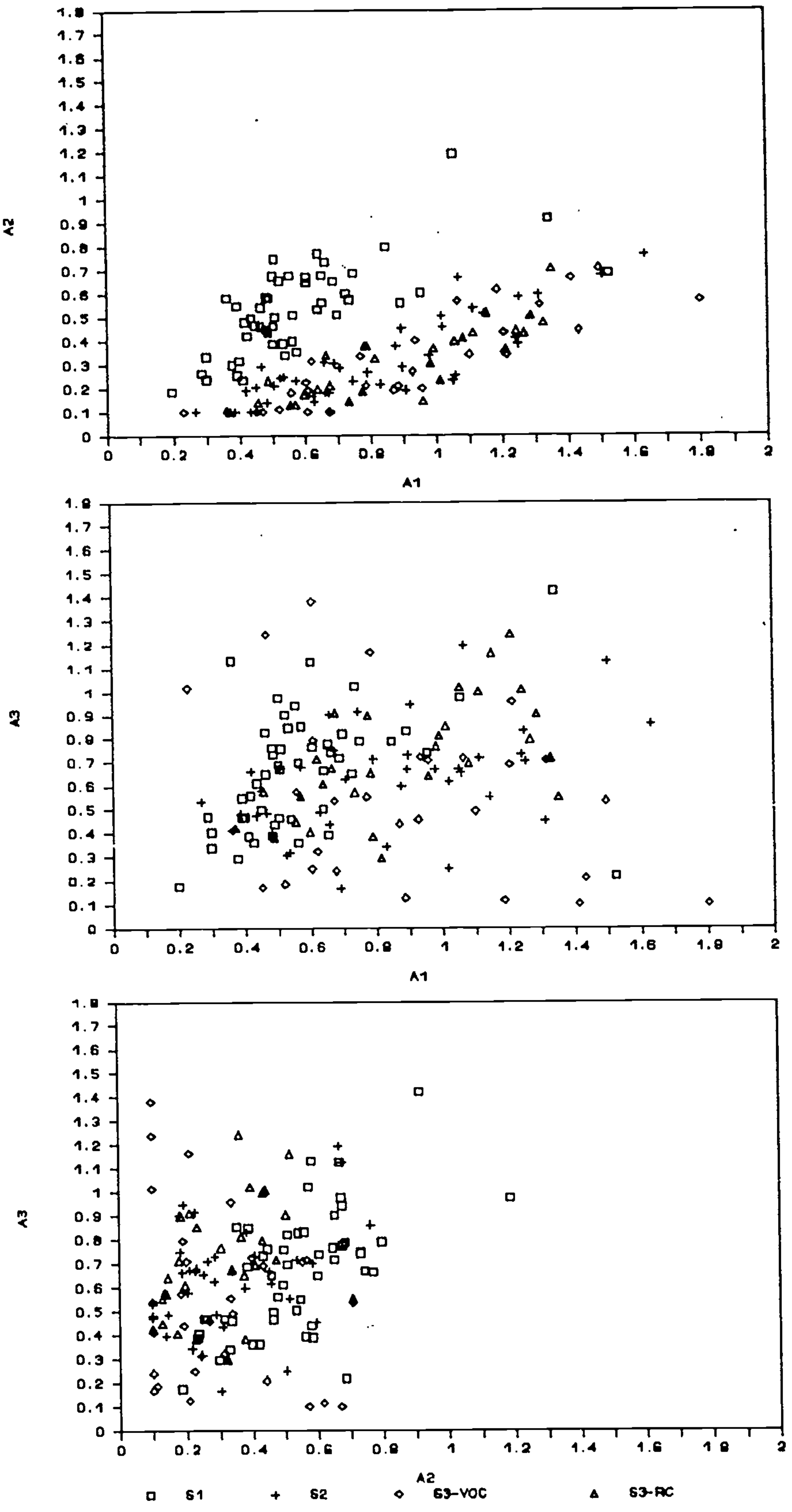

Figure 2: Plots of 3-Dimensional A-Values - Form JP

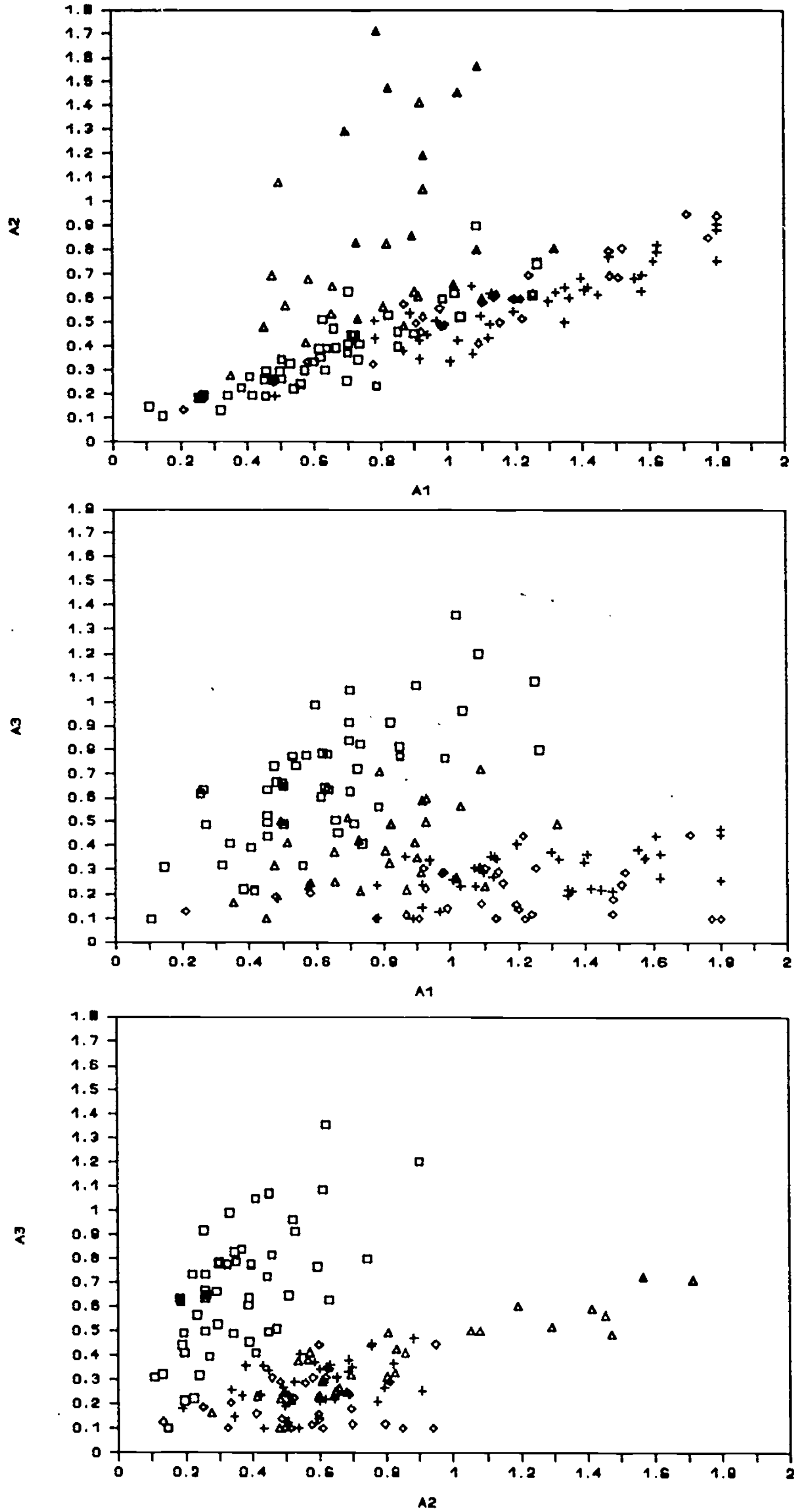

Figure 3: Plots of 3-Dimensional A-Values - Form K9

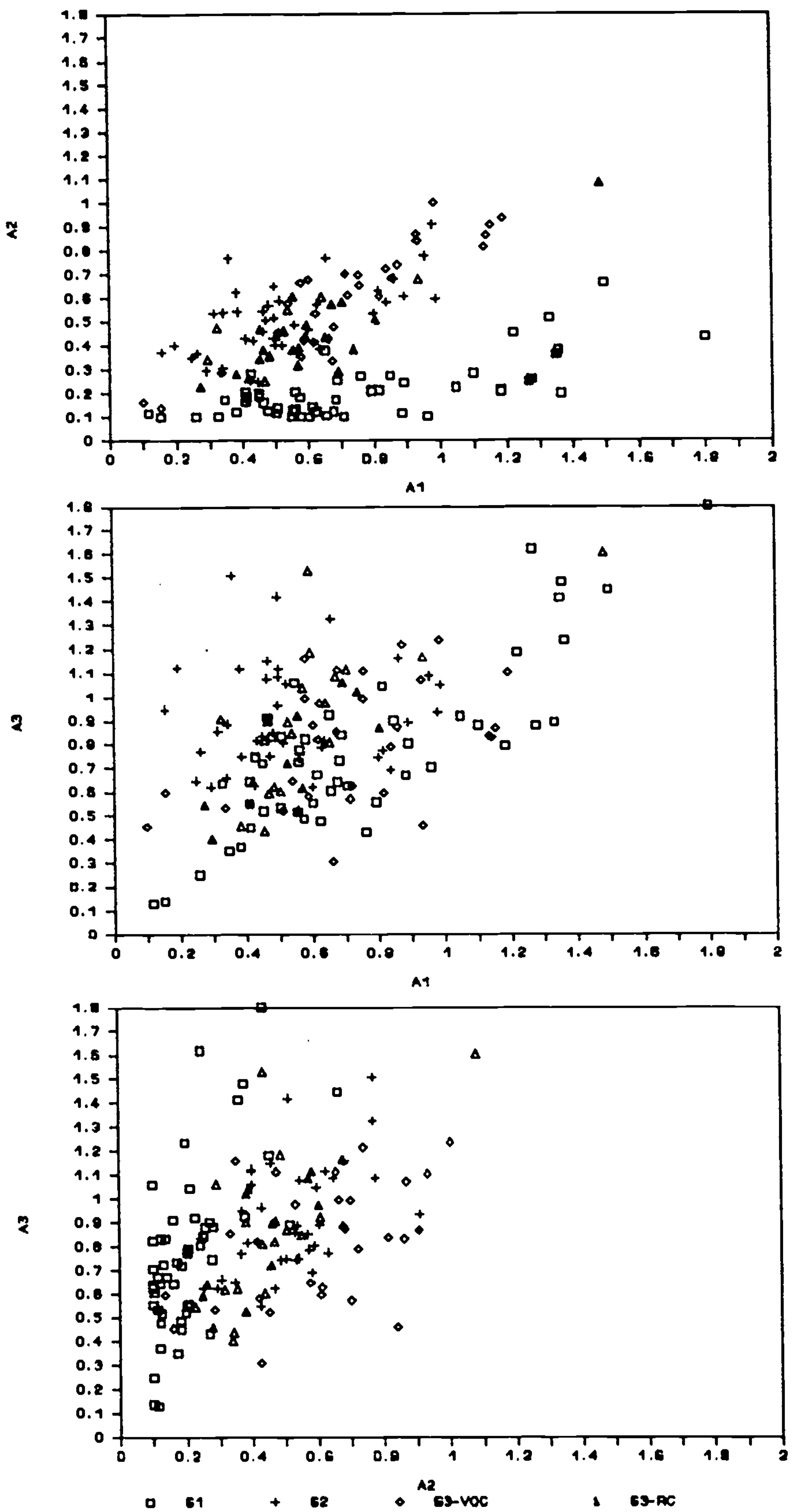

Figure 4: Plots of 3-Dimensional A-Values - Form KA

Confirmatory Results - 3DC1

Table 7 presents the means and standard deviations of the item parameter estimates for each CMIRT solution obtained for the 3DC1 data. These summary statistics are displayed for the relevant a_1, a_2-, a_3-, and b-estimates from Section 1, Section 2, Section 3 vocabulary, and Section 3 reading comprehension. Although the confirmatory solutions are not subject to rotational indeterminancy, it should be noted that the latent scales for these solutions have not been equated and should be compared with caution.

TABLE 7 Means and Standard Deviations of Item Parameter Estimates by Content Area for Confirmatory Solution 3DC1

Content/ Parameter	N	JM Mean	JM S.D.	JP Mean	JP S.D.	K9 Mean	K9 S.D.	KA Mean	KA S.D.
S1 a_1	50	0.86	0.31	0.79	0.27	0.77	0.29	0.83	0.38
S2 a_1	38	0.97	0.37	0.85	0.30	1.03	0.26	0.83	0.21
S3-Voc a_1	29	0.87	0.27	0.76	0.23	0.94	0.29	0.94	0.30
S3-RC a_1	29	0.98	0.32	0.90	0.28	1.00	0.34	0.85	0.27
S1 a_2	50	0.54	0.18	0.51	0.19	0.52	0.22	0.47	0.23
S2 a_3	38	0.53	0.21	0.38	0.22	0.67	0.19	0.51	0.16
S3-Voc a_3	29	0.54	0.20	0.48	0.34	0.64	0.24	0.43	0.16
S3-RC a_3	29	0.34	0.17	0.42	0.20	0.27	0.16	0.38	0.16
S1 b	50	-0.02	0.75	0.27	0.80	-0.26	0.75	-0.13	0.86
S2 b	38	0.38	0.69	0.21	0.75	0.68	0.76	0.30	0.82
S3-Voc b	29	-0.02	1.14	-0.02	1.05	0.37	1.01	-0.08	1.14
S3-RC b	29	-0.02	0.87	0.24	0.89	-0.14	0.75	0.10	0.77

Examinee Sample: JM, JP, K9, KA

It is worth noting that the average reading comprehension a_3-estimates for Forms JM and K9 in Table 7 are much lower than the average a_3-estimates for Section 2 or vocabulary items. In addition, the average difficulties of the listening items compared to the other items tend to follow different patterns in the foreign and domestic samples. (Contrary to more common IRT estimation programs, in the CONFIRM program, higher b-estimates correspond to easier items and lower b-estimates correspond to more difficult items.)

DISCUSSION

The results of the exploratory analyses supported the interpretation that the TOEFL test is characterized by essentially three latent ability dimensions. For each sample analyzed, the CAIC statistics were lowest for the 3D solutions, and for two of the samples (JM and KA), the CAIC statistic was lower for the 2D solution than for the 4D solution. Although the CAIC statistics suggested that a three-dimensional structure of the TOEFL test was optimal, the authors were unable to find a consistently meaningful content-related interpretation for all three dimensions. Inspection of plots of the item discrimination estimates for the exploratory 3D solutions suggested that the listening comprehension section of the TOEFL examination measures a different ability dimension than the nonlistening sections of the test. In addition, for the foreign samples, there was some evidence from the item discrimination estimates that reading comprehension items could be differentiated from Section 2 and the vocabulary items of Section 3.

The best fitting model examined in the confirmatory analyses was Solution 3DC1, which suggests the existence of a general ability dimension, a secondary ability dimension measuring listening comprehension, and a secondary ability dimension measuring a combination of structure and written expression and vocabulary and reading comprehension. An alternate model (4DC1), which allowed for a general ability dimension and secondary dimensions associated with each section of the test, proved clearly to be a less satisfactory solution, as did the 4DC2 model, which differs from the 4DC1 model in that the vocabulary portion of the reading comprehension section is associated with the secondary ability dimension corresponding to the structure and written expression section.

The results with respect to solution 3DC1 tend to confirm the interpretation of the TOEFL offered by Hale et al. (1988) and Hale, Rock, and Jirele (1989), although these authors did not extract a general ability dimension in their analyses. In the present study, the CMIRT algorithm employed required the extraction of a general ability dimension. Thus, it should be noted that other plausible confirmatory structures exist that may provide better fit to the data investigated in this study. For example, it is possible that the structure of the TOEFL test could be better supported by a solution with three correlated dimensions (one for each section). In future investigations, it might be worthwhile to fit CMIRT structures to TOEFL data that do not require every item to load on a general factor.

Although the models identified as best fitting in this study were consistent across test forms and examinee samples, there did appear to be tentative evidence of differences in the salience of different dimensions, depending upon whether the fitted data were based on foreign or domestic samples. A reasonable explanation for this finding is that the makeup of foreign and domestic examinees taking the TOEFL tends to differ in terms of native language and proficiency in various aspects of the English language. Several studies with TOEFL test data have suggested that item performance, test equating, and the structural interpretation of the test may differ according to examinees' native language and/or English proficiency (Alderman & Holland, 1981; Golub-Smith, 1986; Oltman, Stricker, & Barrows, 1988). Furthermore, in operational administrations of the TOEFL test there is consistent evidence of differences in the way foreign and domestic examinees perform on the listening comprehension section, compared to the other sections of the test. These differences appear to be related to differential experience in informal and formal exposure to English. Informal exposure would tend to emphasize communicative aspects, such as speaking and listening, while formal exposure would tend to emphasize grammar, vocabulary, and reading.

CONCLUSIONS

The results of this study indicated that the MIRT and CMIRT procedures were quite successful in modeling secondary ability dimensions on the TOEFL. The two procedures provided corroborative evidence in interpreting the structure of the test. This evidence was consistent with previous interpretations of the test's structure, and was verified by examining the characteristics of item parameter estimates in the various solutions obtained in the study. The data presented in this study also illustrated how the consistent Akaike information criterion can be utilized to identify the best of several competing models of test structure.

Several areas of research related to application of the MIRT and CMIRT models could not be investigated in this study but may be of interest for future investigations. For example, in the present study, no attempt was made to rotate different solutions (for example, the foreign and domestic samples) to a common orientation, or to equate the multidimensional estimates for different solutions to a common scale. As in the case of unidimensional IRT, parameter estimates of the same items obtained in separate calibrations are not directly comparable until they have been transformed to a common scale. Although some progress has been made in this area (cf. Ackerman, 1990; Hirsch, 1989; Reckase, Davey, & Ackerman, 1989), research must continue to address this problem if practical use of MIRT parameter estimates is to be made in applications such as equating and test development. Another possibility for investigation in future studies could be based on comparisons of ability estimates obtained in unidimensional and MIRT solutions that have been transformed to an estimated true score scale. Such transformations might provide information about how much practical impact fitting additional ability dimension would have on examinees' scores. Finally, future applications could attempt to use MIRT and CMIRT procedures to provide diagnostic feedback to test assemblers for the purpose of modifying the TOEFL test design. Such feedback could be used to enhance the measurement of important secondary abilities, or to eliminate them if they were undesired.

REFERENCES

Ackerman, T. (1990, April). An evaluation of the multidimensional parallelism of the EAAP mathematics test. Paper presented at the annual meeting of the American Educational Research Association, Boston.

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. Budapest, Academiai Kiado.

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

Alderman, D. L., & Holland, P. W. (1981). Item performance across native language groups on the Test of English as a Foreign Language (TOEFL Research Report No. 9). Princeton, NJ: Educational Testing Service.

Birnbaum, A. (1958). Statistical theory of some quantal response models [Abstract]. Annals of Mathematical Statistics, 29, 1284.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord and M.R. Novick, Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

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., and Muraki, E. (1985). Full-information item factor analysis (MRC Report No. 85-1). Chicago: University of Chicago.

Boldt, R. F. (1988). Latent structure analysis of TOEFL (TOEFL Research Report No. 20). Princeton, NJ: Educational Testing Service.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. Psychometrika, 2, 345-370.

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.

Dunbar, S. B. (1982, March). Construct validity and the internal structure of a foreign language test for several native language groups. Paper presented at the annual meeting of the American Educational Research Association, New York.

Duran, R. P., Canale, M., Penfield, J., Stansfield, C. W., & Liskin-Gasparro, J. E. (1985). TOEFL from a communicative viewpoint on language proficiency: A working paper (TOEFL Research Report No. 17). Princeton, NJ: Educational Testing Service.

Golub-Smith, M. L. (1986, April). A study of the effects of examinee native language on TOEFL parameter estimation and equating. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

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

Hale, G. A., Stansfield, C. W., Rock, D. A., Hicks, M. M., Butler, F. A., & Oller, J. W. (1988). Multiple-choice cloze items and the Test of English as a Foreign Language (TOEFL Research Report No. 26). Princeton, NJ: Educational Testing Service.

Hale, G. A., Rock, D. A., & Jirele, T. (1989). Confirmatory factor analysis of the Test of English as a Foreign Language (TOEFL Research Report No. 32). Princeton, NJ: Educational Testing Service.

Hirsch, T. M. (1989). Multidimensional equating. Journal of Educational Measurement, 26, 337-349.

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., & McKinley, R. L. (1988, April). Assessing the structure of the GRE General Test using confirmatory multidimensional item response theory. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Lord, F. M. (1952). A theory of test scores. Psychometric Monograph No. 7.

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, NJ: Educational Testing Service.

McKinley, R. L. (1988, April). Assessing dimensionality using confirmatory multidimensional IRT. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

McKinley, R. L. (1989). Confirmatory analysis of test structure using multidimensional item response theory (Research Report No. 89-21). Princeton, NJ: Educational Testing Service.

McKinley, R. L., & Reckase, M. D. (1983). MAXLOG: A computer program for estimating the parameters of a multi-dimensional 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.

Oltman, P. K., Stricker, L. J., & Barrows, T. (1988). Native language, English proficiency, and the structure of the Test of English as a Foreign Language (TOEFL Research Report No. 27). Princeton, NJ: Educational Testing Service.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research.

Reckase, M. D., Davey, T., & Ackerman, T. (1989, April). Similarity of the Multidimensional space defined by parallel forms of a mathematics test. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Swinton, S. S., & Powers, D. E. (1980). Factor analysis of the Test of English as a Foreign Language (TOEFL Research Report No. 6). Princeton, NJ: Educational Testing Service.

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

TOEFL is a program of
Educational Testing Service
Princeton, New Jersey, USA

57906-06442 • Y121M.5 • 275586 • Printed in U.S.A.