DOCUMENT RESUME

ED 386 482 TM 024 039

AUTHOR Chyn, Susan; And Others
TITLE An Investigation of IRT-Based Assembly of the TOEFL Test. TOEFL Technical Report.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-94-38; TR-9
PUB DATE Mar 95
NOTE 46p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *English (Second Language); Item Banks; *Item Response Theory; *Language Tests; Second Language Learning; Selection; Test Construction; Test Items
IDENTIFIERS *Automated Item Selection; Parallel Test Forms; *Test of English as a Foreign Language; Test Specifications

ABSTRACT

The current study, carried out jointly by Test Development and Statistical Analysis staff at Educational Testing Service investigated the feasibility of the Automated Item Selection (AIS) procedure for the Test of English as a Foreign Language (TOEFL). Item-response theory (IRT)-based statistical specifications were developed. Two TOEFL test forms were assembled using AIS, and statistical and content-related properties were evaluated. The results show that the statistical consistency (parallelism) of the tests assembled using AIS would appear to be superior to the consistency of tests assembled using traditional test-assembly procedures. The results also provided strong evidence that AIS-assembled TOEFL tests can successfully meet the IRT-based specifications. Test Development staff observed visible gains in efficiency in item selection for Sections 1 and 2 and the potential for time gains in Section 3. The implications of AIS for pool management were explored. Twelve figures and eight tables present analysis results. One appendix provides a sample AIS specification for Section 1. (Contains 13 references.) (Author/SLD)

TOEFL

March 1995

Technical Report

TR-9

An Investigation of IRT-Based Assembly of the TOEFL Test

Susan Chyn
K. Linda Tang
Walter D. Way

An Investigation of IRT-Based Assembly of the TOEFL Test

Susan Chyn
K. Linda Tang
Walter D. Way

Educational Testing Service
Princeton, New Jersey

RR-94-38

Educational Testing Service is an Equal Opportunity/Affirmative Action Employer.

Abstract

The current study, carried out jointly by Test Development and Statistical Analysis staff at Educational Testing Service (ETS), investigated the feasibility of the Automated Item Selection (AIS) procedure for the Test of English as a Foreign Language (TOEFL). IRT-based statistical specifications were developed. Two TOEFL® test forms were assembled using AIS and statistical and content-related properties were evaluated. The results of this study show that the statistical consistency (parallelism) of the tests assembled using AIS would appear to be superior to the consistency of tests assembled using traditional test-assembly procedures. The results of the study also provided strong evidence that AIS-assembled TOEFL tests can successfully meet the IRT-based specifications. Test Development staff observed visible gains in efficiency in item selection for Section 1 and Section 2 and the potential for time gains in Section 3. The implications of AIS for pool management were explored.

The Test of English as a Foreign Language was developed in 1963 by the National Council on the Testing of English as a Foreign Language. The Council was formed through the cooperative effort of more than 30 public and private organizations 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. 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, 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. 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. Many projects require the cooperation of other institutions, however, 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 data confidentiality will be protected.

Current (1994-95) members of the TOEFL Research Committee are:

Paul Angelis	Southern Illinois University at Carbondale
James Dean Brown	University of Hawaii
Carol Chapelle	Iowa State University
Joan Jamieson	Northern Arizona University
Linda Schinke-Llano	Millikin University
John Upshur (Chair)	Concordia University

Acknowledgments

The authors wish to thank Dan Eignor for reviewing early drafts of this manuscript and for providing guidance and valuable comments. Thanks are also due to Pat Carey, who reviewed early drafts and consulted with the authors on statistical issues, and to Susan Nissan, who in addition to reviewing early drafts, oversaw the development of content-related AIS specifications in Test Development. Finally, the authors are indebted to Martha Stocking and Len Swanson for their thorough and insightful critical reviews.

Table of Contents

List of Figures

List of Tables

Introduction

The Test of English as a Foreign Language has been using item-response theory (IRT) equating procedures since 1978 (Cowell, 1982; Hicks, 1983; 1984). The continued use of IRT equating methods has led to the development of an extensive bank of pretested and IRT-calibrated items. When a new final form of the TOEFL test is constructed, IRT item parameter estimates are taken from this bank and are used to equate the new form to a base form for which a transformation to the existing TOEFL scale exists. In doing this, a transformation is derived to link the new form to the TOEFL scale. Although IRT item parameter estimates are available and are used to derive transformations to the TOEFL scale, until recently TOEFL test assembly procedures have relied on classical statistics [i.e., Delta (Δ) and R-biserial correlation (R_{bis})] in the assembly of tests, rather than using the IRT parameter estimates. Initially, this was because test developers had no practical mechanism for retrieving pretest IRT statistics. Since 1990, however, the evolution of the TD/DC™ (Test Development Document Creation) system and the automated item selection (AIS) algorithm (see Stocking, Swanson, and Pearlman, 1991; Swanson and Stocking, 1993a; Stocking, Swanson, and Pearlman, 1993b) has made it feasible for TOEFL test developers to assemble a test form using the computer and statistical specifications based on IRT.

Automated item selection (AIS) procedures, i.e., the use of a combination of IRT-based statistical specifications and content-related specifications, modern computers, and a mathematical programming algorithm in test assembly, has drawn strong interest and has been recently investigated by many researchers (Baker, Cohen, and Barmish, 1988; de Gruijter, 1990; Theunissen, 1985, 1986; van der Linden and Boekkooi-Timminga, 1989). These researchers have applied linear programming techniques to select items from an item bank. The items selected minimize or maximize a specified target function subject to statistical and content-related constraints. All these studies used the IRT-based test information function as the statistical constraint because it has a very important feature: It consists entirely of independent and additive contributions from each item in a test. The models developed by these researchers have been used to assemble test forms with test lengths ranging from 20 to 65 items from simulated item banks consisting of 300, 500, or 1,000 items (Theunissen, 1985, 1986; van der Linden and Boekkooi-Timminga, 1989). The results of these studies suggested two substantial advantages of automated item selection over manual test assembly: an increase in test construction efficiency and a greater degree of statistical and content consistency (i.e., parallelism) of the test forms. However, these models were only evaluated by simulation studies.

The AIS procedure used at ETS was developed by Swanson and Stocking (1993a), and evaluated by Stocking, Swanson, and Pearlman (1991, 1993b). This procedure is based on a model that is particularly well suited to large testing programs that make use of multiple content-related constraints and statistical constraints in assembling test forms. The model used in this procedure can be described as follows:

Let i=1,...,N be the discrete items or the subsets in the item pool. Let j=1,...,J be the item properties associated with the non-psychometric constraints. Let L_j and U_j be the lower and upper bounds (which may be equal), respectively, on the number of items in the test having each property. The model optimizes an objective function described in equation (8) subject to the constraints specified in equations (1) to (7):

$$\sum_{i=1}^{N} g_i x_i = n, \quad x_i \in (0,1), \; i = 1,...,N. \tag{1}$$

In equation (1), g_i = number of items in a subset if i represents a subset, or $g_i = 1$ if i represents a discrete item; x_i denotes the decision variable, i.e., $x_i = 1$ if item or subset i is included in the test and $x_i = 0$ otherwise; and n denotes the number of items in a test.

The following constraint limits the selection to at most one subset of any item set. Let s = 1,...,S be the item sets in the item pool, and let $b_{is} = 1$ if item/subset i is a subset of set s and 0 otherwise. The constraint is specified as

$$\sum_{i=1}^{N} b_{is} x_i \leq 1, \; s = 1,...,S \; . \tag{2}$$

Let

$$d_{Lj}, \; d_{Uj}, \; e_{Lj}, \; e_{Uj} \geq 0, \; j = 1,...,J, \tag{3}$$

then,

$$\sum_{i=1}^{N} a_{ij} x_i + d_{Lj} - e_{Lj} = L_j, \quad j=1,...,J. \tag{4}$$

In equation (4), $a_{ij} = 1$ if item or subset i has property j and $a_{ij} = 0$ if it does not. The d_{Lj} is the positive (or zero) deviation of the quantity $\Sigma_i a_{ij} x_i$ from the lower bound, that is, the difference between the lower bound and the sum wherever the lower bound is not met. The e_{Lj} represents the nonnegative difference between the sum and the lower bound wherever the lower bound is exceeded. Note that for a given j, one or both of these variables must take on the value zero (that is, the sum cannot both exceed and fail to meet the lower bound). Similarly,

$$\sum_{i=1}^{N} a_{ij} x_i - d_{Uj} + e_{Uj} = U_j, \; j=1,...,J, \tag{5}$$

where d_{Uj} denotes the nonnegative difference between $\Sigma_i a_{ij} x_i$ and the upper bound wherever the upper bound is exceeded, and e_{Uj} denotes the difference between the upper bound and the sum wherever the upper bound is not exceeded.

For IRT-based tests, Swanson and Stocking (1993a) specified a target test-information function as a set of constraints. Let k = 1,... K be points on the ability metric Θ and let $I_i(\Theta_k)$ be the item information for item i at Θ_k. Let $I_L(\Theta_k)$ and $I_U(\Theta_k)$ be the lower and upper bounds, respectively, on test information at Θ_k. Then the IRT-based statistical constraints are expressed by the two equations

$$\sum_{i=1}^{N} I_i(\Theta_k)x_i + d_{Lk} - e_{Lk} = I_L(\Theta_k),\ k=1,...K, \tag{6}$$

and

$$\sum_{i=1}^{N} I_i(\Theta_k)x_i - d_{Uk} + e_{Uk} = I_U(\Theta_k),\ k=1,...,K. \tag{7}$$

Note that these equations are simply special forms of equations (4) and (5), respectively. For test information, the a_{ij} is substituted by item information at a point on the ability metric, instead of 0 or 1, and the L_j and U_j become lower and upper bounds on information, rather than bounds on the number of items having a specified property.

Subject to the above model constraints, the objective function to be minimized is

$$\sum_{j=1}^{J} w_j d_{Lj} + \sum_{j=1}^{J} w_j d_{Uj}\ , \tag{8}$$

where w_j is the weight assigned to constraint j and d_{Lj} and d_{Uj} are defined in (4) and (5). Swanson and Stocking (1993a) refer to the above model as the weighted deviations model.

Because simultaneously satisfying all the constraints is generally not possible in practice, typically because of the size of the item pool and the fairly large number of constraints, Swanson and Stocking (1993a) took a heuristic approach in solving the problem. They treated the constraints as desired properties rather than constraints in the mathematical sense of binary programming, and took into account that certain constraints are more important than others by weighting them accordingly. The heuristic algorithm selects items to decrease the expected weighted sum of positive deviations as described in (8). After the desired test length has been reached, the algorithm successively replaces previously selected items until no further decrease in the weighted sum of positive deviations can be made. Swanson and Stocking (1993a) demonstrated that the heuristic algorithm can solve much larger test-assembly problems than the standard linear programming algorithms, and is much more efficient in terms of CPU time. They also found that the algorithm produced satisfactory results in the actual assembly of tests. In another experiment, Stocking, Swanson, and Pearlman (1991) compared the performance

of AIS with manual test assembly. They found that the test information functions for the tests produced by the AIS procedure were sufficiently close to those of the manually assembled tests.

Given the recency of development of the AIS procedure, there have been few empirical studies of the application of the procedure in operational test assembly for large testing programs. The purpose of the present study was to apply the AIS procedure to TOEFL to explore the possibility of improving the consistency and efficiency of TOEFL test assembly. In this study, IRT-based statistical specifications were developed. Two TOEFL final forms were assembled using AIS and statistical and content-related properties were evaluated. The implications of AIS for pool management were explored.

Methodology

Because the application of the automated item selection procedure in TOEFL test assembly involves both statistical and test development-related issues, the methods used in this study can likewise be separated into two categories: statistical methods and test development-related methods. The statistical methods include:

1) *selecting an IRT-based information function for TOEFL statistical specifications*
2) *developing IRT-based statistical specifications*
3) *investigating the relationship between the classical and IRT-based item statistics*
4) *investigating the statistical properties of TOEFL item pools to ensure that the statistical specifications can be met by the item pools*
5) *evaluating the statistical properties of AIS assembled tests*

The test development-related methods include:

1) *developing content-related constraints (content rules)*
2) *evaluating the content of AIS assembled tests*
3) *evaluating the efficiency of the AIS assembly procedures*
4) *investigating the implications for TOEFL item-pool management*

Statistical Methods

Developing IRT-based target information function curves. As indicated in equations (6) and (7), the AIS algorithm requires that the statistical constraints (rules) be specified by certain IRT-based information functions. The IRT model used for the TOEFL test is the three-parameter logistic (or 3PL) model which is defined as

$$P\{\theta;\ a_i,\ b_i,\ c_i\} = c_i + \frac{1 - c_i}{1 + \exp[-Da_i(\theta - b_i)]}, \qquad (9)$$

where i = 1,..., number of items. In equation (9), 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, θ is the ability parameter, and D is a constant assuming the value of 1.7 (which is employed to make the logistic curve closely approximate the normal ogive model).

By definition, the information function $I\{\theta, y\}$ for any score y is inversely proportional to the square of the length of the asymptotic confidence interval for estimating ability θ from score y (Lord, 1980). Therefore, the higher the information, the narrower the confidence interval for ability θ, i.e., the higher the information, the greater the precision in ability estimation.

Three information functions were considered for use in TOEFL IRT-based statistical specifications: information for a scaled (converted) score provided by the observed score; information for ability provided by observed score (score information function); and information for ability provided by the IRT maximum likelihood ability estimator ($\hat{\theta}$) (test information function). The advantages and disadvantages of using each of these three information functions for TOEFL IRT-based statistical specifications were evaluated.

The converted score information provided by the observed score is given in Lord (1980, equations (6-9) and (6-10)), and is specified as

$$I(SS_\eta, x) = \frac{1}{\sum_i P_i(\theta)Q_i(\theta)(d\ SS_\eta/d\eta)^2}, \qquad (10)$$

where x is the observed score for a test form, η is the true score for a test form, θ is the ability parameter corresponding to η, SS_η is the converted score which is a monotone transformation of θ and is also a monotone transformation of η, $P_i(\theta)$ is defined in equation (9), and $Q_i(\theta) = 1 - P_i(\theta)$.

The score information is defined in Lord (1980, equation (5-13))

$$I(\theta, x) = \frac{(\sum_i P_i'(\theta))^2}{\sum_i P_i(\theta)Q_i(\theta)}, \qquad (11)$$

where $P'_i(\theta)$ is the derivative of $P_i(\theta)$ with respect to θ, and the other terms have been defined previously.

The test information function is defined in Lord (1980, equation (5-6))

$$I(\theta,\hat{\theta}) = \sum_i I_i(\theta, u_i) = \sum_i \frac{P_i'(\theta)^2}{P_i(\theta)Q_i(\theta)}, \qquad (12)$$

where $I_i\{\theta, u_i\}$ is the item information function of item i, $u_i \in \{0,1\}$ is the response of item i, and all other terms have been defined previously.

The information function given in Equation (10) has an interpretive advantage of being on the reported converted score metric, and is a function of the reciprocal of the IRT-based estimate of the conditional standard error of measurement. The relationship between the converted score $SS\eta$ and true score η does not exist in functional form, however, if the base form raw-to-converted score conversion is not linear; hence the required derivative has to be estimated using numerical methods. While algorithms exist

in the TD/DC™ system to handle these conditions, they have not been extensively tested in operational situations.

The score information function, defined in (11), also has an interpretive advantage: It describes the information about ability parameter θ provided by observed score x. The observed score does not provide maximum information about θ, however. Based on Theorem 5.3.2 in Lord (1980), the test information function $I\{\theta,\hat{\theta}\}$, or the information about ability parameter θ provided by the maximum likelihood estimate $\hat{\theta}$ as defined in equation (12), is the upper bound on the information about θ that can be obtained by any method of scoring the test. The test information function also has a very important feature that is lacking in the other two information functions: It consists entirely of independent and additive contributions from the items. The TD/DC™ system provides updated graphs of information functions obtained after the addition of each new item. Based on the evaluation of the three information functions, it was decided to use a test information function to define the TOEFL statistical specifications.

Developing the target information curves. IRT statistics for eight TOEFL operational forms administered from July 1989 to August 1990 were examined to develop the target information curves. For each TOEFL section, the mean and standard deviation of information across the eight forms at each of the 21 values of θ were calculated. These 21 values of θ were equally spaced, ranging from -3.0 to 3.0 in increments of 0.3. Next, a constant at each θ value was obtained by multiplying the standard deviation by 1.28. (The value 1.28 was selected because it results in an interval that encompasses 80 percent of a standard normal distribution). This constant was added to and subtracted from the mean information at each θ value to obtain the upper and lower boundaries (target curves) to serve as statistical specifications. These specifications provide values of $I_L(\Theta_k)$ and $I_U(\Theta_k)$ in equations (6) and (7).

The target information function curves are presented in Figures 1 to 3 for TOEFL Sections 1 to 3, respectively. The information function for an AIS-assembled test is required to fall within the two boundaries. This requirement ensures that a test form will provide sufficient information, and also ensures that the information functions for all the AIS-assembled tests will be sufficiently close to each other so that the tests can be considered parallel. The θ range between the two vertical lines corresponds to the range of TOEFL converted scores between 45 and 60, where measurement is of the most interest. Because the test information function is the reciprocal of the asymptotic variance of the maximum likelihood estimator of ability, and the test information functions peak in this range, examinee abilities in this range are more accurately estimated.

Figures 1 to 3 show that the shape of the information function for Section 2 is flatter than those for Sections 1 and 3. This is because the test information function is a summation of item information functions and Section 2 has fewer items than the other two sections.

Figure 1
Target Test Information Function for Section 1

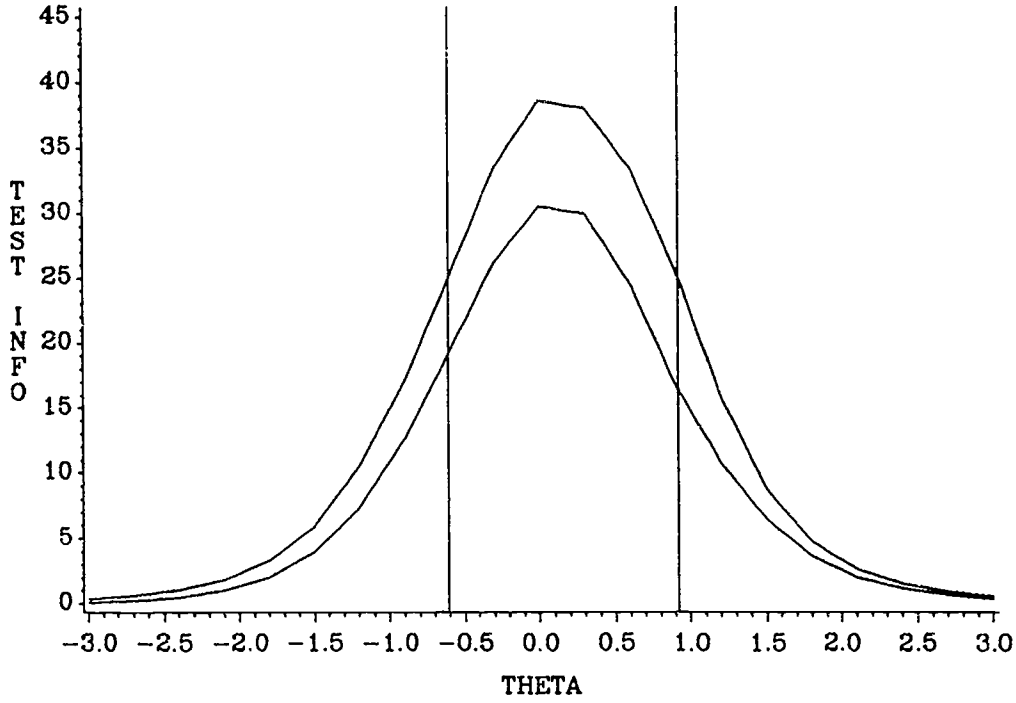


Figure 2
Target Test Information Function for Section 2

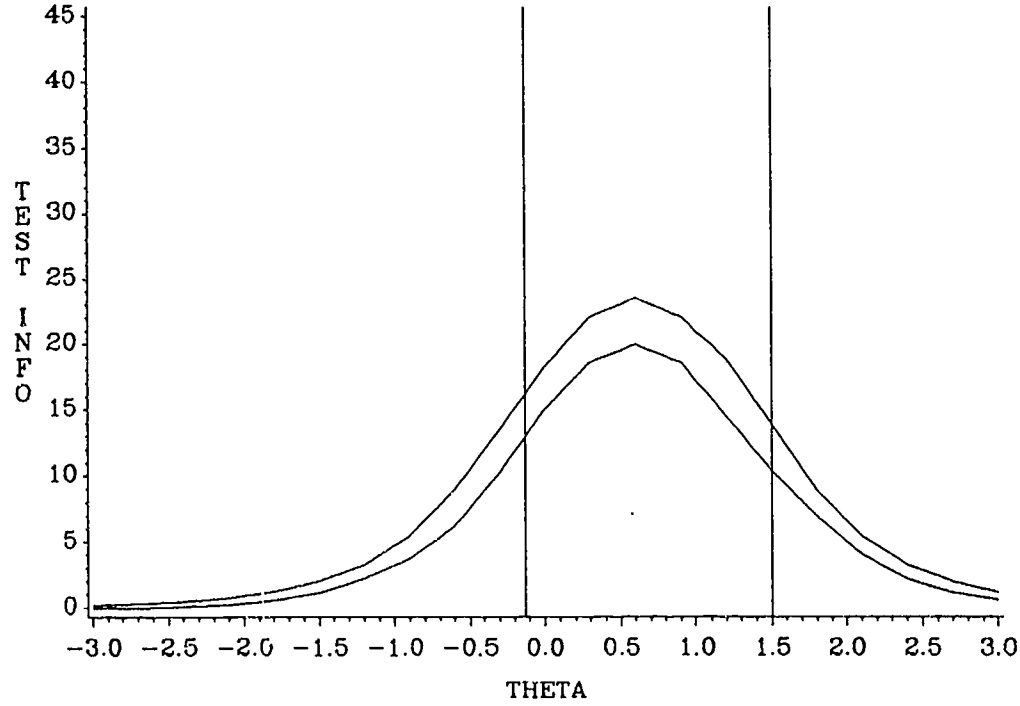
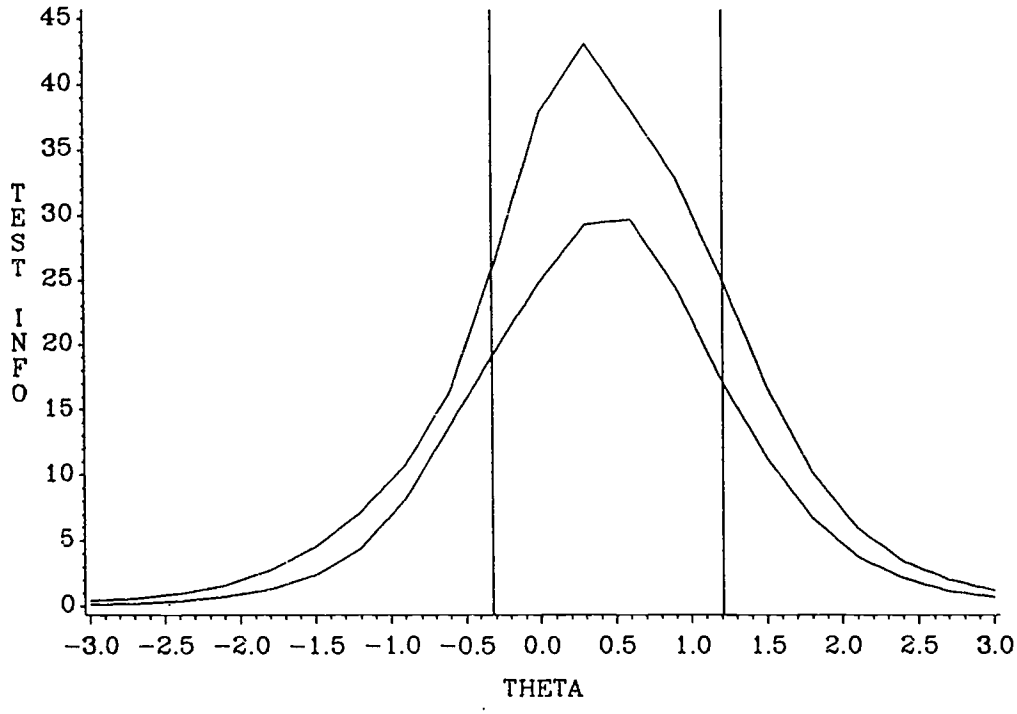


Figure 3
Target Test Information Function for Section 3



Identifying the relationship between classical and IRT item statistics. In order to facilitate the transition from the use of classical item statistics in manual test assembly to the use of IRT-based item statistics in AIS test assembly, the relationship between IRT-based item level statistics and classical item statistics was studied. The knowledge of IRT-based item level statistics was especially important for test developers, who needed to replace single items in the later (post-AIS) phases of assembly.

The relationship between classical item statistics (i.e., R-biserial correlation and Delta) and the IRT a- and b- parameters was theoretically determined by Lord (1980) under the assumptions that (1) θ is normally distributed; and (2) there is no guessing. For the TOEFL test, however, assumption (1) is met approximately by Section 1 only, whereas assumption (2) is violated in all three sections. Further, because the item information function used in the TOEFL statistical specifications is specified using all three IRT item parameters, it would seem more important to know the relationship between this function or a closely related function and the classical statistics than the relationship between each of the parameters used in specifying this function and the corresponding classical statistic.

It can be seen in equation (12) that the item information function is also a function of the ability parameter θ. This function achieves its maximum values (Infomax) at θ = θ_{max} (Thetamax). The quantity Thetamax for an item i based on the 3PL model is described in the following equation:

$$\theta_{MAX_i} = b_i + \frac{1}{Da_i} \ln \frac{1 + \sqrt{1 + 8c_i}}{2} . \qquad (13)$$

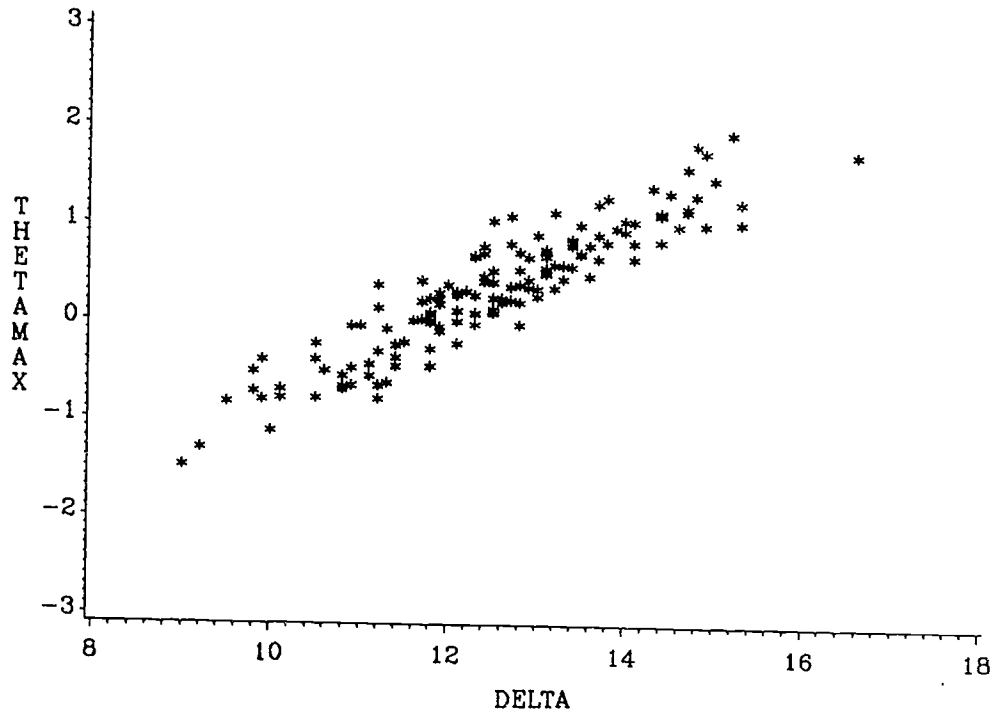
The quantity Thetamax is a function of the three IRT item parameters and can be considered as an indicator of item difficulty. For example, an item having maximum item information at θ_{max} = -1.5 can be considered to be easier than an item having maximum item information at θ_{max} = 1.5. In addition, because the test information function is a summation of the item information functions and will peak in a particular θ range if many items in the test have Thetamax values in this range, selecting items with particular values of Thetamax allows the test assembler to manipulate the height of the test information function curve in a specified θ range. From the classical perspective, Delta, a function of the proportion of examinees who answer an item correctly, is the item difficulty index. The classical item difficulty statistic Delta is defined as:

$$\Delta = 4z + 13 , \qquad (14)$$

where z is the value corresponding to 1 - P_+ in the standard Normal distribution, and P_+ is the proportion of examinees who answer an item correctly.

For each of the seven TOEFL item types, a bivariate plot of Thetamax vs. Delta was produced. As an example, the scatter plot for the Reading Comprehension item type in Section 3 is presented in Figure 4. The plot suggests a linear relationship between these two statistics.

Figure 4
The Linear Relationship between Thetamax and Delta
Section 3, Reading Comprehension



To further investigate the relationship between these two statistics, both linear and quadratic functions were fitted to the data using a regression technique. For all item types, except for Structure items in Section 2, the estimated quadratic coefficients were not significantly different from 0. This means that the relationship between the two statistics is essentially linear for these item types. For Structure items in Section 2, a small but significant quadratic coefficient was obtained ($\hat{\beta}_2$ = -0.0284, $p < 0.05$). Because the estimated quadratic function reaches its peak and starts to decrease at Delta = 35, and the TOEFL observed Delta range is between 6 and 16, a linear model was considered to be more practical for that item type.

The estimated linear regression model, using Δ to predict Thetamax, is in the following form:

$$\hat{\theta}_{max} = \hat{\beta}_0 + \hat{\beta}_1 \Delta \ , \qquad (15)$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the least squares estimators of the intercept and slope, respectively, of the regression line. The values of these two estimators for each of the seven TOEFL item types is presented in Table 1.

Table 1
Linear Regression: Using Delta to Predict Thetamax

	$\hat{\beta}_0$	$\hat{\beta}_1$	R^2	N
Section I				
Statement	-3.6455	0.3421	0.7899	95
Dialogue	-3.5084	0.3307	0.7661	80
Minitalk	-4.7785	0.4518	0.7511	84
Section II				
Structure	-4.2275	0.3671	0.7493	95
Writ. Exp.	-4.6395	0.4071	0.7281	174
Section III				
Vocabulary	-4.5527	0.3859	0.7066	159
Read. Comp.	-5.1264	0.4366	0.8514	145

The R^2 in Table 1 indicates the proportion of variation of Thetamax that can be explained by the variation of Delta, and it is in fact the square of the Pearson correlation. The R^2 value for each of the item types, ranging from 0.7281 to 0.8514, indicates that the linear relationship is moderately strong. Because the slope of the linear regression line is positive for each of the seven item types, one can expect that an item with a high Delta will also have a high Thetamax and vice versa. Knowing this relationship between Delta and Thetamax, one can have confidence that the transition from assembling TOEFL tests based on Delta specifications to tests based on IRT specifications should not be a difficult task for the test assemblers.

Developing rules to ensure the item pool can support statistical specifications. In the experimental stages of applying AIS to the TOEFL test assembly, it was observed that several AIS-assembled forms had all items within a narrow Thetamax range. Therefore, to facilitate pool management, rules (constraints) were developed to ensure that the distribution of Thetamax values for a test reflects the distribution in the item pool[1].

In order to develop these rules, an inventory of the statistical characteristics of the TOEFL item pool was conducted. The proportion of items in each Thetamax category

[1]It may be suggested that the Thetamax distribution rules are redundant with the statistical specifications (target curves). These distribution rules provide insurance to test assemblers that the item pools can support the new statistical specifications, however. As more experience is gained about the Thetamax distribution and its relationship to statistical specifications, these rules may be dropped.

in the pool and in AIS-assembled experimental tests (10 experimental tests for Section 1 and five for Section 2) was computed. Based on the comparison of the proportion of items in each Thetamax category in the pool and in the experimental tests, the upper and lower bounds for the number of items at each Thetamax category in the test were developed. These rules insure that the distribution of Thetamax in a test is close to that in the item pool. These rules were developed for the discrete items in Section 1 (Statements and Dialogues) and Section 2 (Structure and Written Expression) only. For the subpart in a section that has item sets (Extended Conversations and Minitalks in Section 1 and Reading Comprehension in Section 3), the AIS unit is a set instead of an item. Because the Thetamax distributions are different across sets, rules were not developed for these item types. Rules were also not developed for the discrete item type in Section 3 (Vocabulary), because this item type will not be used in the TOEFL test in the near future. An example of the rules to control the Thetamax distributions for Statements and Dialogues in Section 1 of a test is presented in Table 2.

Table 2
Rules to Constrain the Thetamax Distribution in a Test to Reflect the Item Pool

Statements		Dialogues	
θ_{max}	# items	θ_{max}	# items
-2.00 - -1.00	0 - 1	-2.00 - -1.00	0 - 1
-0.99 - -0.50	1 - 3	-0.99 - -0.50	0 - 2
-0.49 - 0.00	3 - 5	-0.49 - 0.00	3 - 6
0.01 - 0.50	6 - 8	0.01 - 0.50	4 - 6
0.51 - 1.00	4 - 6	0.51 - 1.00	2 - 4
1.01 - 1.50	1 - 3	1.01 - 2.00	0 - 2
1.51 - 2.00	0 - 1		

Evaluating AIS tests. Five to 10 experimental tests for each of the three TOEFL sections were assembled using the AIS procedure. Their test information function curves were evaluated without replacing any items. The Thetamax distributions of the discrete items of each test were also evaluated (except Vocabulary in Section 3). Two TOEFL final forms were also assembled using the AIS procedure, and the items in the tests were evaluated from a content perspective by the Test Development staff. Items which did not meet content requirements were revised or replaced. The descriptive statistics for both the classical and IRT-based item statistics were computed for these forms. IRT-based

preequatings were conducted on these forms, and the converted score differences on these forms were computed and plotted.

Test Development Related Methods

Statistical consistency or parallelism of test forms was but one of many important concerns in determining the feasibility of AIS test assembly. Additionally, test developers needed to investigate 1) rules which would successfully embody test content specifications and guidelines, 2) the efficiency of the AIS assembly process, 3) the nature of the AIS-generated tests, and 4) the implications of AIS assembly for pool management.

Many months were spent on the development and refinement of rules for each of the three TOEFL sections. Test development experts in TD/DC software applications who were familiar with the item classification codes used in the computerized inventorying of item pools worked closely with TOEFL Test Development Section Coordinators and staff from Statistical Analysis to create rules. When weighted appropriately, these rules form sets of AIS specifications that in turn would generate acceptable draft tests. Processing efficiency suggested that the number of rules should be as small as possible; the complexity of the language construct suggested that the numbers needed to be greater. This phase of the research was highly iterative. Since that time, in fact, the investigation and refinement of AIS rules has become an ongoing part of test development work.

Once the rules were developed, two three-section TOEFL tests, or a total of six separately timed test sections, were assembled by six different test assemblers. Each assembler received a printout of a draft test which met all the statistical and content constraints embodied in the AIS rules. Assemblers were asked to identify weak items, problems in content balance, or other areas in which the test did not meet content specifications. When problems were found, the assemblers were to revise or manually replace individual items, making sure the properties of the replacement items did not bring the test information function curve for the revised test out of the desired statistical bounds. Assemblers were asked to monitor how much time they spent during this initial assembly phase.

In the interests of efficiency, the six test sections were treated as operational forms and, as part of the routine test development process, underwent a series of reviews. These included a Test Specialist Review, a Coordinator Review, and a Planograph (or Mechanical) Review. Assemblers were asked to record how many item replacements and revisions they made after each of the reviews, including the initial Assembler's Review described above.

After the tests were completed, test assemblers were asked to respond to the following questions designed to elicit their subjective observations about the AIS development process:

What is your perception of how AIS assembly differs from traditional assembly in terms of:

- *the quality of the test*
- *the nature of the process, including reviewing and replacing items (this can include intangible aspects such as how the process felt to you)*
- *section-specific concerns*

In a large-scale testing program such as TOEFL, which tests internationally every month, active pool management is vital to guarantee valid and consistent measurement from form to form. But exactly **how** large should an AIS/IRT pool be? The assumption was that the AIS algorithm would perform better when there was an overage of items, particularly when multiple parallel forms needed to be produced concurrently. Large pools (of more than 1,000 items) usually have more depth and breadth than small pools, but can require long processing times of more than 10 hours on a 386-20E personal computer to acquire multiple forms. Minipools (of several hundred items) have quicker processing times and are more flexible, as several test assemblers can work concurrently on different forms with different minipools and can draw upon mutually exclusive sets of replacement items.

In order to investigate the relationship of pool size to the number of actual AIS test forms that could be assembled for each of the three TOEFL sections, successive AIS runs of multiple forms were carried out on pools of sizes varying from 290 to 1,432 items, until the maximum number of forms that could be successfully created was established. Then, for each pool, the ideal number of forms that could be created **without** any regard for rules was calculated. (For example, if there were no rules to meet, six 50-item forms could be created from a 300-item minipool.) This rough number provided a constant against which ratios could be calculated, indicating the degree of overage required in pools of varying sizes in different sections. Additionally, because it was hypothesized that the presence of item sets required a proportionally larger pool for successful AIS, the percent of items in sets (or nondiscrete items) in each of the sections and in each of the pools was calculated.

Results

Summary of Statistical Results

Evaluating AIS-assembled tests without replacing items. Five to 10 AIS-assembled experimental tests were evaluated in terms of satisfying the statistical specifications for each of the three sections. The experimental tests for Sections 1 and 2 were also evaluated in terms of the rules that constrain the Thetamax distribution in a test. Because similar results were obtained for the three sections, and Section 1 has both discrete items and items in sets, the results of four experimental tests in Section 1 are presented in this section[2].

Figure 5 presents test information functions for four AIS-assembled tests for Section 1. All items in the tests were selected by the AIS procedure and no items were replaced. Except at the very low and very high ends of the θ scale, the test information functions for all four tests are between the target boundary curves. The test information function for one form (Experimental Test 2) is slightly above the upper bound at $\theta < -2.0$ (The maximum violation is 0.16 at $\theta = -2.4$), and for another form (Experimental Test 4) is slightly above the upper bound at $\theta > 2.1$. (The maximum violation is 0.20 at $\theta = 2.7$). Because these θ ranges are either two standard deviations above or below the mean of the θ distribution, where less than 2.5 percent of the examinees are included, the violation is considered to be tolerable. In addition, the violation can be easily corrected by replacing one or two items with lower Infomax values.

[2]The reason for reporting only four of the 10 experimental tests is that five of the 10 tests were used to develop the Thetamax distribution rules, i.e., they were assembled before the rules were input to the AIS software. One experimental test was used as a draft operational test immediately after the initial evaluation of the fit of the target information boundary curves, and the test was revised without the original statistics being saved.

Figure 5
Test Information Functions for Four AIS-Assembled Tests without Replacing Items
Section 1

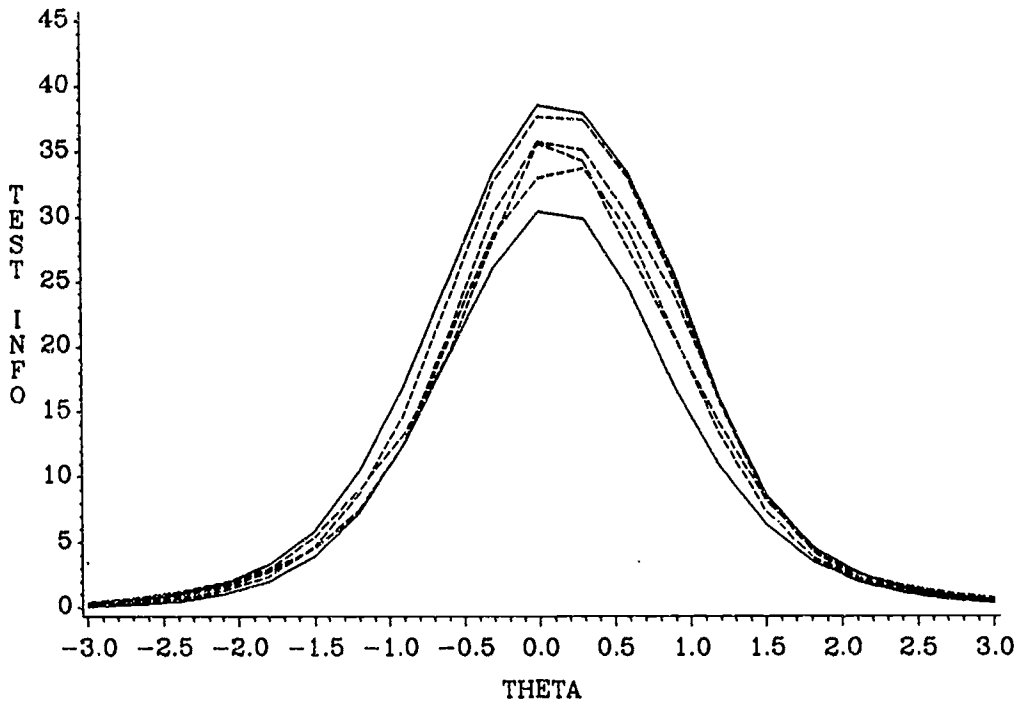


Table 3 shows that for four experimental tests, the distributions of Thetamax were within the specified rule range. For Experimental Test 4, the AIS procedure selected one more item than specified in the Thetamax range between -0.99 and -0.50 for Dialogues items. Other than that violation, the Thetamax distributions of the four tests are similar.

Table 4 presents the Thetamax distributions of the Minitalks and Extended Conversations items in Section 1 for the four experimental tests. These items are in sets, and no Thetamax distribution constraint rules were developed for these item types. Table 4 shows that the Thetamax distributions are very different among the four tests.

Table 3

Number of Items in Each θ_{max} Range for the Four Experimental Tests (Section 1)

Item Type	θ_{max} Range	Rule Range	# Items T1	T2	T3	T4
Statements	-2.00 - -1.00	0 - 1	0	0	1	1
	-0.99 - -0.50	1 - 3	3	2	3	3
	-0.49 - 0.00	3 - 5	4	5	4	4
	0.01 - 0.50	6 - 8	8	8	6	7
	0.51 - 1.00	4 - 6	4	4	4	4
	1.01 - 1.50	1 - 3	1	1	2	1
	1.51 - 2.00	0 - 1	0	0	0	0
Dialogues	-2.00 - -1.00	0 - 1	0	1	1	1
	-0.99 - -0.50	0 - 2	2	1	2	3*
	-0.49 - 0.00	3 - 6	6	6	4	5
	0.01 - 0.50	4 - 6	5	4	5	4
	0.51 - 1.00	2 - 4	2	3	2	2
	1.01 - 2.00	0 - 2	0	0	1	0

Note: * indicates the number is outside the θ_{max} range.

Table 4
Number of Items in Each θ_{max} Range for Items in Sets
Section 1

θ_{max}	# Items			
Range	T1	T2	T3	T4
-2.00 - -1.00	4	6	0	1
-0.99 - -0.50	0	0	1	1
-0.49 - 0.00	6	1	3	3
0.01 - 0.50	2	2	8	3
0.51 - 1.00	2	4	3	4
1.01 - 1.50	1	1	0	1
1.51 - 2.00	0	1	0	2

Evaluating the two TOEFL final forms assembled by AIS. The statistical properties for two TOEFL final forms are discussed in this section. These tests were initially assembled using the AIS procedure. After the tests were assembled, several items were revised by the test assemblers based on content considerations. For Section 1 of Test 1, two items were revised. For Section 1 of Test 2, one item was revised. For Section 3 of Test 1 and Test 2, two items were revised in each test[3].

Test information function curves for Sections 1 to 3 of the two TOEFL final forms are presented in Figures 6 to 8. The revised items were excluded from the computation of the test information functions because the IRT parameter estimates obtained at the time of their pretest use were no longer valid. The test information functions for the sections with revised items were proportionally increased so that they reflected the same number of items as the target curves.

Figures 6 to 8 show that, with the exception of a few trivial violations, all curves fall within the target boundaries. For Test 2, Section 3, the test information function is slightly higher than the upper target at θ values between -0.9 to -0.6. These values are outside the θ range where measurement is of the most interest. The maximum violation is 0.28.

[3]The revisions discussed here are serious enough that the revised items can no longer be considered to be the same items. The number of revised items presented in Table 7 includes items with both serious revisions and minor revisions (items are essentially unchanged).

Figure 6
Test Information Functions for AIS-Assembled Final Forms
Section 1

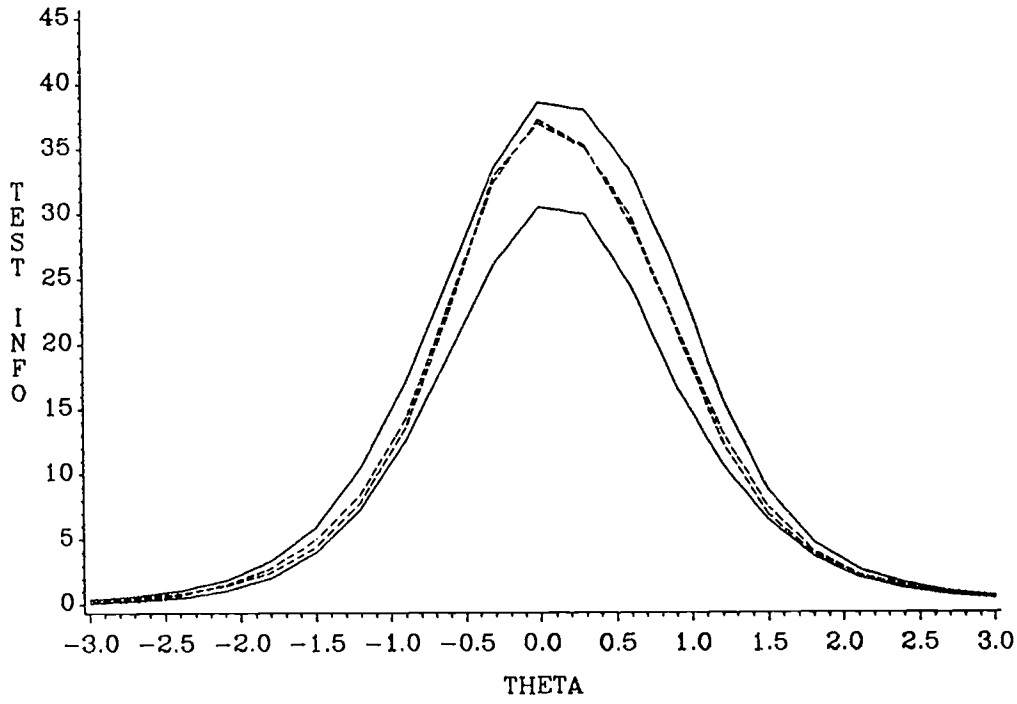


Figure 7
Test Information Functions for AIS-Assembled Final Forms
Section 2

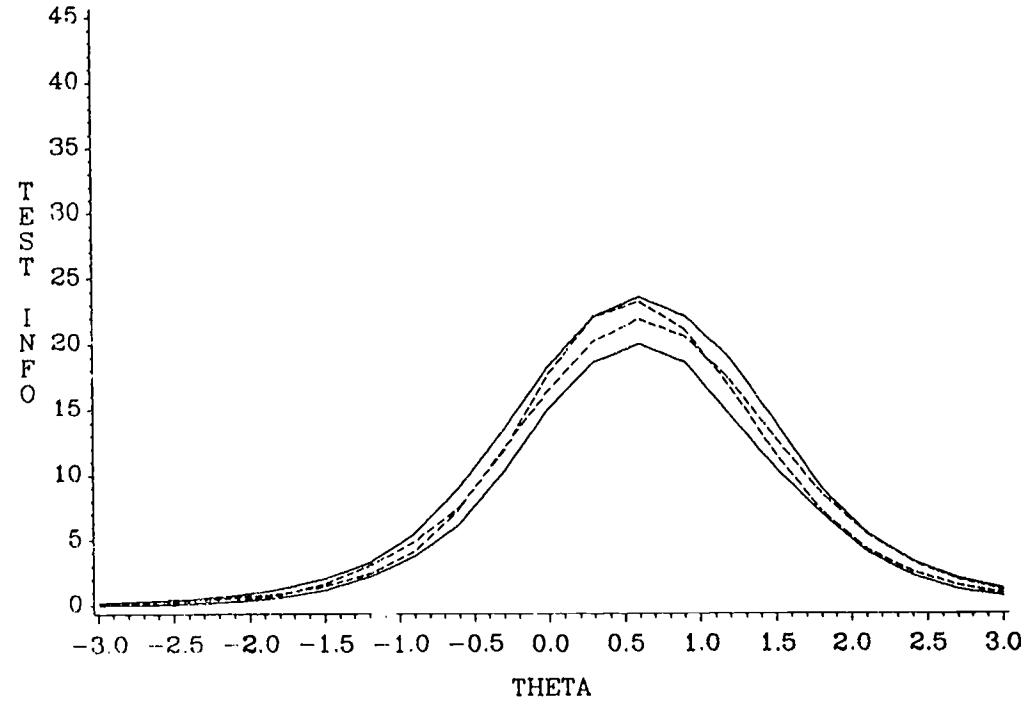
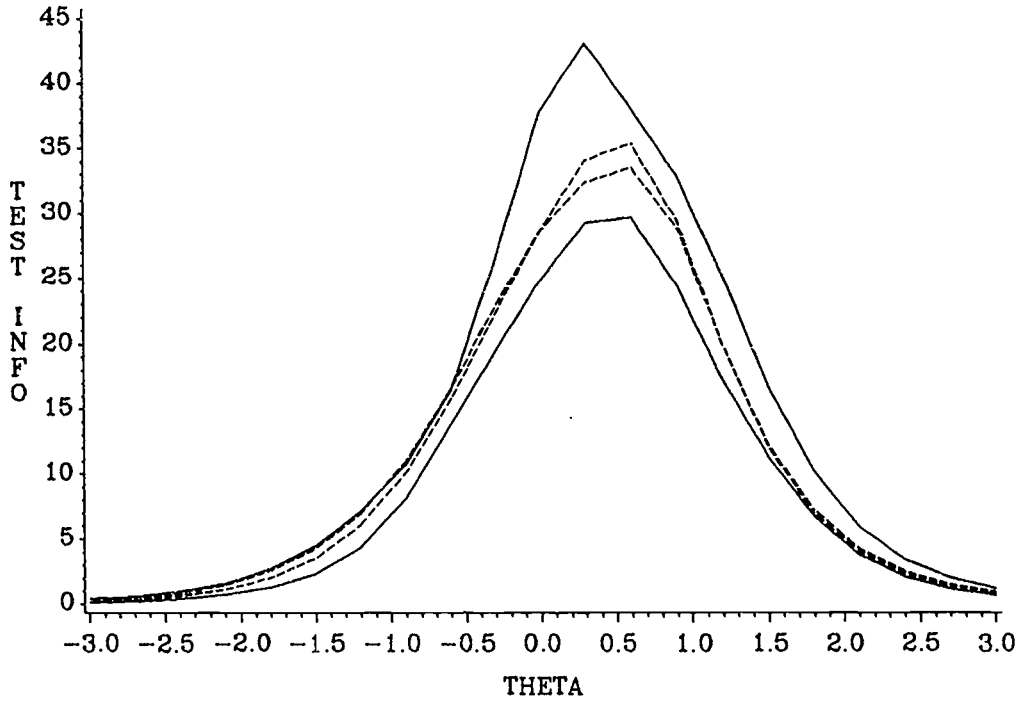


Figure 8
Test Information Functions for AIS-Assembled Final Forms
Section 3



Summary statistics for these two tests are presented in Table 5. The means for each of the item statistics, both classical and IRT-based, are sufficiently close for the two forms. For Sections 2 and 3, the test form that has the higher mean Thetamax value also has the higher mean Delta value. This confirms the positive linear relationship between these two item statistics discussed previously. For Section 1, Test 1 has a slightly lower mean Thetamax value and a slightly higher mean Delta value than Test 2. This seems to violate the predicted positive linear relationship between Delta and Thetamax. The differences are small, however. Because the linear regression equations had R^2 values between 0.75 and 0.79 for Section 1 item types, this very small reversal is acceptable.

Table 5
Summary Statistics for the Two AIS-Assembled Final Forms

Test	Section	n	$\bar{a}$	$\bar{b}$	$\bar{c}$	θ_{max}	$I(\theta_{max})$	Δ	R_{bis}
1	1	48	1.49	-0.06	0.22	0.04	1.12	10.76	0.51
2	1	49	1.56	-0.03	0.26	0.09	1.16	10.66	0.50
1	2	38	1.39	0.36	0.25	0.50	0.92	12.51	0.50
2	2	38	1.39	0.40	0.24	0.53	0.91	12.58	0.49
1	3	56	1.34	0.14	0.24	0.27	0.85	12.32	0.47
2	3	56	1.31	0.09	0.22	0.22	0.84	12.28	0.48

To evaluate the parallelism of the two forms, the differences between the rounded converted scores for these two test forms for each of the three sections were calculated and are illustrated in Figures 9 to 11[4]. It should be kept in mind that the converted scores were derived using pretest item parameter estimates; therefore, the raw-to-converted score transformation may not be as accurate as the transformation obtained for the real test administration, where item parameter estimates are obtained from a much larger sample.

[4]An IRT true score equating procedure is used to equate TOEFL forms to the base form and obtain converted scores for given raw scores in a TOEFL test form.

Figure 9
Converted Score Differences between the Two AIS-Assembled Tests (Section 1)

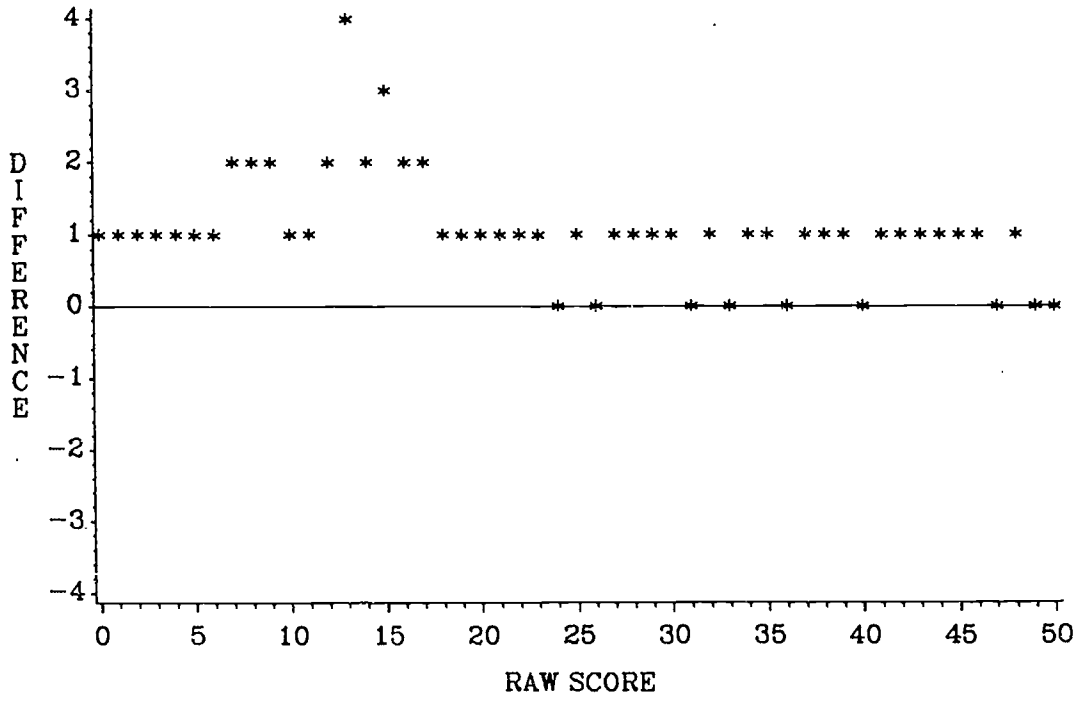


Figure 10
Converted Score Differences between the Two AIS-Assembled Tests (Section 2)

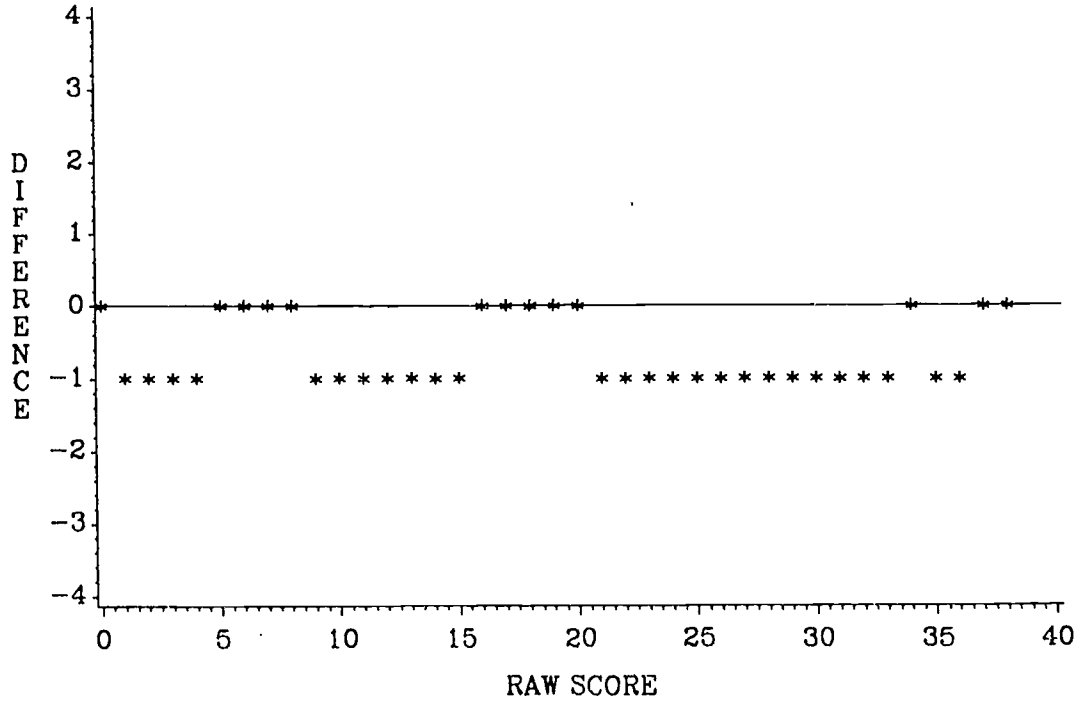
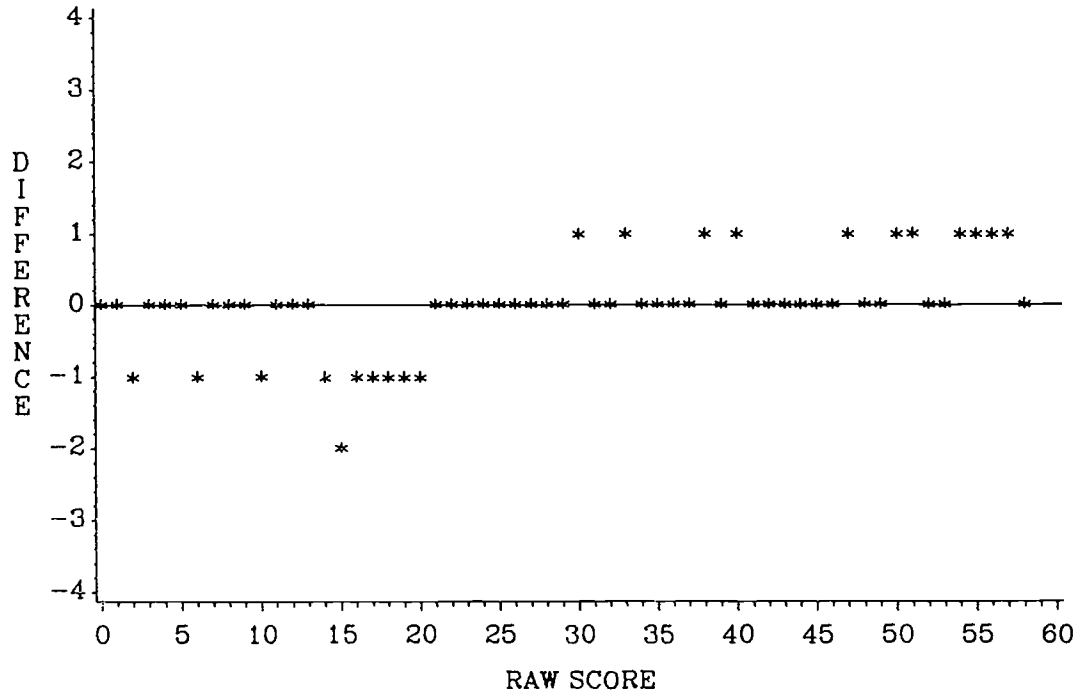


Figure 11

Converted Score Differences between the Two AIS-Assembled Tests (Section 3)



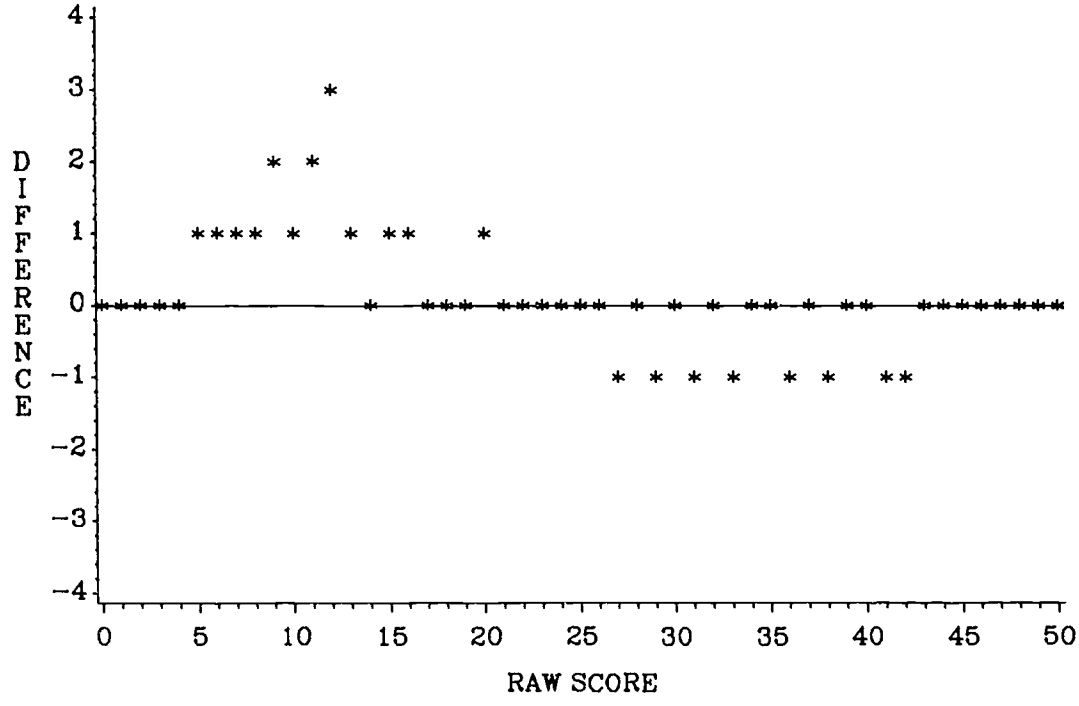
The differences in rounded converted scores for Section 2 AIS forms are within one point. For Section 3, the differences in rounded converted scores are also within one point except at raw score 15, where a two-point difference is observed. Because this score point is two standard deviations below the mean of the raw score distribution, and a very small percentage of examinees will obtain that score, this difference is acceptable.

For Section 1 at raw scores below 17, which corresponds to abilities two standard deviations below the mean, the differences in rounded converted scores between the two AIS forms are somewhat large: four points for raw score 13, three points for raw score 15, and two points for seven other raw score points. These differences likely reflect the fact that both the abilities, and hence the equating relationships at very low score levels, were not very well estimated. One of the reasons for the less precise estimation in this score range is that the target information function provides little information at two standard deviations below the mean. Again, because very few examinees will obtain scores in this range, these differences are considered acceptable. It is worth mentioning that three items in the two tests were revised. For comparison purposes, the item parameter estimates obtained from their pretest use for these items were included in the true score equating. This may also contribute to the differences in rounded converted scores seen for Section 1.

Figure 12 presents the differences in converted scores based on two recently administered TOEFL tests (Section 1) which were manually assembled. Similar patterns in score differences as seen in Figure 9 can be observed in Figure 12. Therefore, for Section 1, the degree of parallelism between the two AIS-assembled tests is similar to that observed between two test forms assembled using traditional procedures.

Figure 12

Converted Score Differences between Two Manually Assembled Tests (Section 1)



Test Development Related Results

The development of AIS content-related rules. The number and nature of the rules created for each section varied to a great extent. The AIS specifications for Section 1 included 120 weighted rules; for Section 2, 87 weighted rules; and for Section 3, 49 weighted rules. The number of rules required was related to many factors, including the number and type of item formats within a section, the type of categorization of language (e.g., structural, functional, subject matter, gender), item difficulty, key distribution, etc. Weights and upper and lower bounds were initially set based on the requirements of content and statistical specifications and later adjusted based on the results of pool inventories and trial runs of the AIS software. In the Appendix, five rules from the Section 1 specification are presented.

The AIS assembly process. Even with the learning curve for dealing with IRT statistics, AIS assembly generally proved to be more efficient than the traditional mode of assembly, as can be seen in Table 6. This table presents a comparison of the time required by test assemblers for the initial AIS assembly phase (beginning when staff received the printout of the draft test and ending when the test was deemed ready to be handed over for the Test Specialist Review) and the time required for traditional test assembly (also ending just before the Test Specialist Review). Because assembly time varied to such an extent in initial efforts, especially for Section 3, a range of hours is reported instead of the mean. This variance is attributed to differing work styles and speeds of individual staff, which assumably would manifest itself in both modes of assembly, as well as to the luck of the draw in an AIS test, i.e., the varying number of items that need to be revised or replaced in any given draft test and the richness of the pool from which replacement items must be chosen.

Table 6
Test Construction Time Required by Assembler
(Before Test Specialist Review)

	Traditional Method	AIS
Section 1	7 - 10 HRS.	3 - 7 HRS.
Section 2	4 - 6 HRS.	2 - 4 HRS.
Section 3	12 - 14 HRS.	6 - 21 HRS.

For Sections 1 and 2, AIS assembly took roughly half the time that traditional assembly did. For both these sections, the longest AIS assembly time was equal to the shortest assembly time in the traditional mode. For Section 3, however, the shortest time, six hours, was six hours shorter than the shortest traditional assembly time. On the other hand, the longest AIS time, 21 hours, was seven hours longer than the longest traditional

assembly time. When discrete vocabulary items in Section 3 needed to be replaced because the words had been tested in recent months, or when whole sets in Section 3 needed to be replaced because of content overlap with other sets or because of item quality, the cumbersome nature of mixed assembly modes (i.e., manually replacing appropriate IRT items originally selected by machine) caused the AIS/IRT assembly process to be less efficient than the traditional mode of item selection.

The data in Table 7 further demonstrate how content changes varied from section to section in the different phases of the AIS assembly process. The number of content item revisions and replacements made in response to the Test Assembler's Review of the draft test reflects how the content of the two AIS-assembled tests in this study needed to be modified in order to meet the test specifications for each section. These findings are consistent with the ranges of time reported in Table 6: The fewest changes were made in Section 2; the most changes were made in Section 3. It should be noted, however, that this trend is probably partly due to the relative size of each section, with 50 scored items in Section 1, 38 scored items in Section 2, and 58 scored items in Section 3.

Table 7
Number of Changes in Content of Two AIS TOEFL Tests

		TA		TSR		Coordinator		Plano		Total	
		RV	RP	RV	RP	RV	RP	RV	RP	RV	RP
Section 1	T1	2	6	1	0	3	0	1	0	7	6
	T2	1	2	3	1	6	0	0	0	10	3
Section 2	T1	0	2	2	0	2	2	0	0	4	4
	T2	0	1	0	2	0	2	0	0	0	5
Section 3	T1	6	8	2	0	6	1	1	0	15	9
	T2	3	15	2	1	4	2	0	0	9	18

Note: TA indicates test assembler review; TSR indicates test specialist review; RV indicates the number of revisions; RP indicates the number of replacements.

The number of changes in content made in response to each of the subsequent three reviews is also documented. It was hypothesized that, beginning with the Test Specialist Review, the number and type of revisions and replacements would be comparable to those made in manually assembled tests. Although the present study did not track revisions and replacements in tests produced in the traditional assembly mode, staff reported that, based on previous experience, the number of changes in an AIS/IRT test was about the same. It

is important to note as well that staff also reported the amount of time required to make manual replacements in order to meet IRT specifications in the later phases of AIS/IRT assembly sometimes took longer than in the traditional assembly mode involving classical statistics. In general, however, assembly time diminished, and given the concern on the part of many staff and others in the field that computer-assembled IRT tests risk sacrificing test content and test validity (Linn, 1990), it was gratifying to be able to document the continuing role of expert review and quality at the same time efficiencies had been achieved by exploiting technology.

Results of survey. Staff response to the use of AIS was generally positive. Individual comments often reflected the obstacles inherent in the application of the AIS model to a particular TOEFL section. The richness (or lack thereof) of the item pool and the quality of individual items or item sets were cited by many as critical factors contributing to the quality of AIS tests as well as to the efficiency of the process.

The advantages of AIS assembly are:

- *It is usually less time-consuming.*
- *It facilitates the regular turnover and review of items in the pool and thus facilitates timely pool maintenance.*
- *It serves as a job aid by balancing keys, gender, and other criteria which are tedious to tally manually.*
- *It fosters a lack of ownership of work and a distancing from the subject material in items (fostering objectivity).*

The disadvantages of AIS assembly are:

- *It fosters a lack of ownership of work and a distancing from the subject material in items (fostering engagement).*
- *The nature of the cognitive tasks performed by the assembler are less demanding and less creative.*
- *To some assemblers, AIS tests do not seem as good as they would have had the items been selected additively using individual judgment.*

Most respondents were sanguine about the changing role of the assembler in AIS tests, believing that the advantages noted above warranted the implementation of the technology. As one individual wrote, "I can say that I surely would have made at least a few different item choices had I used traditional methods of assembly. There can be, ironically, a sense of sameness about some of the items randomly selected by AIS... [But] one still has the autonomy to reject, replace, move items, etc."

One aspect of the AIS assembly process commented on most frequently was the occasional difficulty of finding appropriate replacement items which would also meet the IRT target curves. With all discrete items, meeting IRT specifications in Section 2 was the

easiest. For Sections 1 and 3, finding items that met the statistical specifications proved difficult and time-consuming when an entire set needed to be replaced. Another area for improvement cited was the inefficiency that resulted when constraints inherent in the test content specifications were not reflected in the AIS rules. Examples of these were content classifications that are abstract or vaguely formulated and thus subject to interpretation, and vocabulary word pairs that had been tested in recent months and were thus not appropriate for use. Sometimes only a few items out of a large item set were selected, and this minisampling of items did not appropriately cover the passage content.

Pool maintenance. Table 8 presents the results of the AIS runs for parallel forms on item pools of varying sizes. The Ideal/Actual efficiency ratio evaluates, in very broad terms, the relative efficiency of AIS on a given item pool, making the assumption that there are no significant imbalances or deficiencies in that pool. In terms of efficiency, the ideal AIS pool, in which no extra items would be required, would yield a 1:1 ratio. In fact, in the practical and complex world of item development in large testing programs, some overage is necessary and even desirable. The ideal/actual ratio should thus be viewed as a performance test of how the contents of a given pool translate into multiple parallel forms. It is **not** intended to replace a rigorous inventory of items in the various cells of the classification matrices. It also cannot compare the number of machine-selected forms to the number that could be created manually.

Table 8
Relationship of Pool Size to Number of Actual AIS Forms

	Test Spec		AIS Spec	Pool		Parallel Forms		
	# Items	% of Items in Sets	# Rules	Pool Size[1]	% of Items in Sets	Ideal #	Actual #	Efficiency Ratio[2]
Section 1	50	30%	120	343	48%	6	3	2.0 : 1
				495	28%	9	4	2.3 : 1
				838	36%	16	7	2.3 : 1
Section 2	38	N/A	87	320	N/A	8	4	2.0 : 1
				340	N/A	8	5	1.6 : 1
				640	N/A	16	9	1.7 : 1
Section 3	58	50%	49	290	57%	5	2	2.5 : 1
				1423	47%	24	10	2.4 : 1

Note: [1]Including number of items, not including stimulus material.
[2]Efficiency Ratio is Ideal/Actual, and does not factor in the percentage of items in sets.

It was hypothesized that the AIS algorithm would have a higher efficiency ratio 1) on larger item pools and/or 2) when fewer items were required in a test. The data in Table 8 support these hypotheses. It is possible that the efficiency ratio would be higher when proportionately fewer items are linked to item sets and/or there is a small number of AIS rules. In this study, it was not possible to control adequately for those conditions. Section 2, composed entirely of discrete items, showed the best efficiency ratio of all sections. In this study, Section 3 required the largest overage of items despite the fact that it had the fewest number of rules to meet.

Based on these data, pools with proportionately more sets may need to contain more overage than other pools. It is not clear to what extent pool size is a contributing factor in the functioning of the algorithm.

Conclusions and Discussion

Statistical Conclusions and Discussion

In order to apply the automated item selection (AIS) procedure to TOEFL test assembly, IRT-based statistical specifications were developed; the relationship between a particular IRT-based item statistic and the classical Delta item statistic was investigated; and rules to ensure the item pools could support the statistical specifications were developed. In addition, the statistical consistency (parallelism) of the AIS-assembled tests was evaluated in terms of the similarities of the test information function curves and the closeness of the rounded converted scores. The results of the study indicate that the TOEFL pool supports AIS assembly, and the statistical quality of the tests can be improved as a result of this application.

Because AIS uses IRT-based statistical specifications, which are precisely defined in functional form, the degree of parallelism among tests is better controlled. The present study found that the rounded converted score differences between two AIS-assembled test sections were within one point at all but one raw score point. This finding indicates that the statistical consistency (parallelism) of the tests assembled using AIS may be superior to the consistency of tests assembled using classical statistical specifications.

The results of the study also provided strong evidence that AIS-assembled tests can successfully meet the IRT-based statistical specifications. Before the IRT-based statistical specifications were used in practice, it was suspected that the more complicated IRT specifications, compared with the simpler Delta specifications, might be difficult to meet. The results of the study showed that the test information functions for draft tests assembled by the AIS algorithm were usually within, or sufficiently close to, the target information boundaries. The replacement of a few items based on content considerations will have only a small impact on the curves in general.

The results of the study also suggested several issues worthy of further investigation. The rules developed to ensure that the item pool is able to support the statistical specifications were occasionally violated while other constraints (rules) were met. One explanation for this might be that the violated rules were weighted by one, and other rules, such as the upper and lower bounds of the target curves and some content rules, were weighted by five or more. Based on the limited experience of applying AIS to the TOEFL test, it is not clear whether the weights on certain of the rules are appropriate. If violations of the rules for the number of items within certain Thetamax ranges are consistently observed, the weights for these rules may need to be adjusted. In addition, these rules may need to be periodically modified because new pretested items are continually added to the pool and others are removed. Therefore, the statistical properties of the pool may change. How often an inventory of the pool needs to be conducted to update the rules can be decided upon once more experience is gained.

The present study investigated only the Thetamax distributions of discrete items in the item pool. However, after July 1995, all items in Section 3 of the TOEFL will appear in sets. Therefore, in the future, pool inventories should also be conducted for sets in addition to discrete items. More information about item statistic distributions within a set and between sets will facilitate pool management.

This study investigated only one IRT-based item statistic, Thetamax. Another important IRT-based item statistic, Infomax, might need to be investigated further. For example, an item with very low Infomax is certainly not desirable. Estimating the bivariate distribution of Thetamax and Infomax for items in the item pool might be informative.

Test Development Conclusions and Discussion

Just as Statistical Analysis staff observed an improvement in the precision of measurement in the AIS-assembled tests, Test Development staff observed visible gains in efficiency in the assembly of final forms in Sections 1 and 2 and the potential for time gains in Section 3. Offsetting these efficiencies is the time required for online pool maintenance functions, including an increased need to carry out quality control checks on item computer codes to guarantee the integrity of the data, and the frequent physical transfer of data from one pool to another as items are downloaded or regrouped to form new minipools. Above all, the optimization of AIS technology requires Test Development staff to focus more than ever on 1) the quality control of raw item development, 2) the rigorous culling of the pretest item pool, and 3) the building of a calibrated item pool especially tailored to facilitate the workings of AIS, e.g., by focusing on item categories where constraints are routinely not met.

Because the concept of AIS pool management is still relatively new to the TOEFL program, additional data will need to be collected before a clear pattern can be established. To learn more about the optimum size and composition of an AIS pool, a research study is needed in which the overall AIS pool efficiency ratio for the assembly of multiple forms is compared with patterns of pool deficiencies identified by inventories. Another recommended action is for Test Development staff to systematically examine the rules, or constraints, which have been violated after the threshold AIS run and to develop extra items in the categories that were short. Rules will need to be further refined, especially those for Section 3. One area of recommended research is to explore whether AIS constraints could include lexical searches.

Additionally, Test Development staff should explore new ways of performing lexical searches on the online pool so that inappropriate items will not be acquired into AIS draft tests. The principle underlying all of these recommended steps is to shift the focus of effort and quality control to phases in the test development process that occur **before** AIS runs take place.

Finally, the accumulation of experience in using the AIS procedure to assemble the present paper-and-pencil TOEFL tests should provide valuable information for assembly of computer linear or computer adaptive versions of the TOEFL test, should the program choose to move in that direction.

Appendix

Sample AIS Specification for TOEFL Section 1
(Four Forms Selected)

Constraint Name	Weight	Lower Bound	Upper Bound	Items Found	Items Selected Form 1	Form 2	Form 3	Form 4
LCSART	1.0	0	4	4	1	0	0	1
LCSBUS	1.0	0	4	4	0	0	1	0
LCSCLO	1.0	0	4	3	0	0	0	0
LCSCOM	1.0	0	4	1	0	1	0	0
LCSED	1.0	1	4	29	3	4	3	3

Note: LCSART represents a Statement item on the subject of the arts. The other rule names use a similar convention.

References

Baker, F. B., Cohen, A. S., and Barmish, B. R. (1988). Item characteristics of tests constructed by linear programming. Applied Psychological Measurement, 12, 189-199.

Cowell, W. R. (1982). Item response theory preequating in the TOEFL testing program. In P. W. Holland and D. B. Rubin (Editors) Test equating. New York: Academic Press.

de Gruijter, D. N. M. (1990). Test construction by means of linear programming. Applied Psychological Measurement, 14, 175-181.

Hicks, M. (1983). True score equating by fixed b's scaling: A flexible and stable equating alternative. Applied Psychological Measurement, 7, 255-266.

Hicks, M. (1984). A comparative study of methods of equating TOEFL test scores (Research Report 84-20). Princeton, New Jersey: Educational Testing Service.

Linn, R. L. (1990). Has item response theory increased validity of achievement test scores? Applied Measurement in Education, 3, 115-141.

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

Stocking, M. L., Swanson, L., and Pearlman, M. (1991). Automated item selection using item response theory (Research Report 91-9). Princeton, New Jersey: Educational Testing Service.

Swanson, L., and Stocking, M. L. (1993a). A model and heuristic for solving very large item selection problems. Applied Psychological Measurement, 17, 151-166.

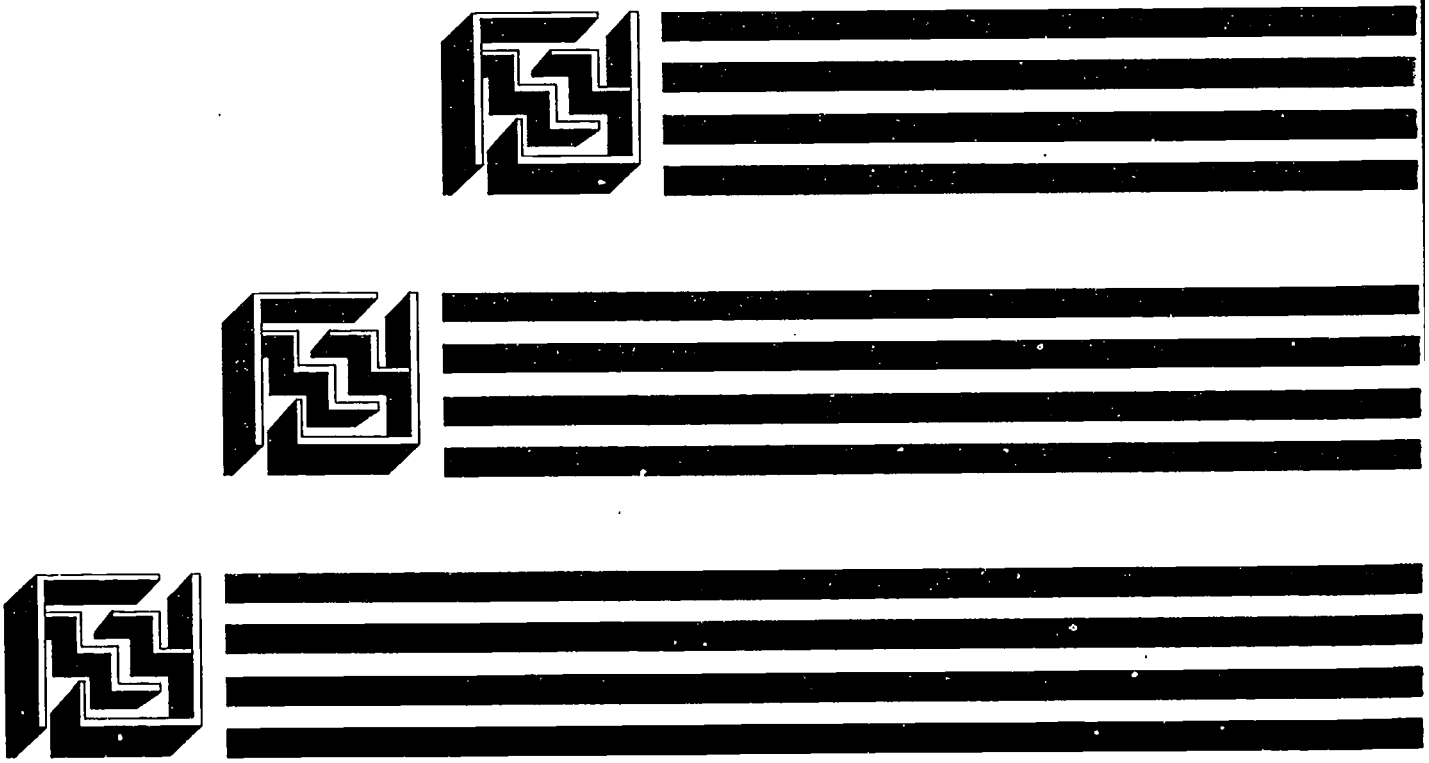
Stocking, M. L., Swanson, L., and Pearlman, M. (1993b). Application of an automated item selection method to real data. Applied Psychological Measurement, 17, 167-176.

Theunissen, T. J. J. M. (1985). Binary programming and test design. Psychometrika, 50, 411-420.

Theunissen, T. J. J. M. (1986). Some applications of optimization algorithms in test design and adaptive testing. Applied Psychological Measurement, 10, 381-389.

van der Linden, W. J., and Boekkooi-Timminga, E. (1989). A maximin model for test design with practical constraints. Psychometrika, 54, 237-248.

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



57906-06442 • Y25M 5 • 275590 • Printed in U.S.A.