ED 395 942 TM 025 009

AUTHOR	Hicks, Marilyn M.
TITLE	The TOEFL Computerized Placement Test: Adaptive Conventional Measurement. TOEFL Research Reports, Report 31.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-89-12
PUB DATE	Jan 89
NOTE	40p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Adaptive Testing; Adults; Algorithms; *Computer Assisted Testing; Computer Literacy; *Difficulty Level; English (Second Language); *Equated Scores; Item Response Theory; Scaling; *Scoring; Simulation; Student Placement; Test Construction; Test Validity; True Scores
IDENTIFIERS	*Test of English as a Foreign Language

ABSTRACT

Methods of computerized adaptive testing using conventional scoring methods in order to develop a computerized placement test for the Test of English as a Foreign Language (TOEFL) were studied. As a consequence of simulation studies during the first phase of the study, the multilevel testing paradigm was adopted to produce three test levels varying in difficulty, with a basic testing algorithm that routed examinees through item blocks or testlets that permitted backtracking in order to review and change items. The ability to control important facets of test construction was illustrated. Resulting test levels were equated to the established scale via item response theory true score equating, and some desirable properties of the obtained score scales were identified, namely overlapping scales at the boundaries of the test levels and limits on the scores obtainable on each of the levels. Data from a preliminary validation study with 152 examinees indicated that the test was functioning satisfactorily. These examinees indicated that, although 72% had never been exposed to a computer before, 59% preferred the computerized test version or were neutral. The apparent lack of computer experience of English-as-a-Second-Language students must be considered in computerized test development. (Contains 3 figures, 4 tables, and 13 references.) (Author/SLD)

TEST OF ENGLISH AS A FOREIGN LANGUAGE

Research Reports

REPORT 31
JANUARY 1989

The TOEFL Computerized Placement Test: Adaptive Conventional Measurement

Marilyn M. Hicks

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 over 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 TOEFL 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. Local research may sometimes require access to TOEFL data. In such cases, the program may provide this data following approval by the Research Committee All TOEFL research projects must undergo appropriate ETS review to ascertain that the confidentiality of data will be protected.

Current (1988-89) members of the TOEFL Research Committee include the following.

Patricia L. Carrell (Chair)	Southern Illinois University
Lily Wong Fillmore	University of California at Berkeley
Fred Genesee	McGill University
Russell G. Hamilton	Vanderbilt University
Frederick L. Jenks	Florida State University
Harold S. Madsen	Brigham Young University

The TOEFL Computerized Placement Test: Adaptive Conventional Measurement

by

Marilyn M. Hicks

Educational Testing Service
Princeton, New Jersey 08541

RR-89-12

ACKNOWLEDGMENTS

Appreciation for suggestions and assistance in the execution of this project is extended to Neal Kingston for recommending the method of the simulations studies in the first phase of the study; Jennifer Kole for the extensive programming required for these studies; and the staff associated with EASIS for programming the test. Grateful appreciation is also extended to Eli Hinkle of the University of Toledo and William Gaitskill of UCLA for their participation in the field trials of the test, and for their helpful comments and suggestions.

ABSTRACT

This study focused on methods of computerized adaptive testing using conventional scoring methods in order to develop a TOEFL computerized placement test. Some advantages of adaptive conventional measurement were illustrated, such as providing the user with an alternative, efficient test with most of the same properties of the conventional test. Test specifications and structure very closely parallel the full-length test, yielding scores on the (familiar) reported score scale, thus providing the user with comparable interpretation within the limits of differences in score precision.

As a consequence of simulation studies conducted during the first phase of this study, the multilevel testing paradigm was adopted for this development. Its implementation produced three test levels varying in difficulty, each approximately half the length of the regular TOEFL. The basic testing algorithm routed examinees through item blocks or testlets that permitted backtracking in order to review answers and change them, an option not easily implemented in standard computer adaptive testing. The ability to control important facets of test construction, as well the degree of measurement effectiveness, using this testing method was illustrated. Resulting test levels were equated to the established scale via IRT true score equating, and some desirable properties of the obtained score scales were described, namely, overlapping scales at the boundaries of the test levels and limits on the scores obtainable on each of the levels.

Data from a preliminary validation study were presented that indicated the test was functioning satisfactorily. Responses to a questionnaire administered to the sample of examinees revealed that 72% of the group had never been exposed to a computer before this testing experience. Nonetheless, 59% indicated that they preferred the computerized test to the paper-and-pencil version or were neutral. The apparent lack of computer experience among ESL students will need to be considered in any computerized test development for this group of examinees.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Adaptive Conventional Measurement. Most of the current computerized adaptive testing research and development derives from latent trait theory. In this testing mode, the examinee is branched through the test on the basis of statistical information associated with each individual item (Fisherian information). Yet there may be many instances when an adaptive conventional test, in which the examinee is branched using conventional scoring methods, would be appropriate or even preferred. This study, undertaken in two phases, focused on methods of computerized adaptive testing using conventional scoring methods in order to develop a TOEFL computerized placement test. Major advantages of the resulting adaptive conventional instrument are that test specifications and test structure very closely parallel the full-length test, yielding scores on the reported score scale, thus providing the user with comparable interpretation on both the computerized and paper-and-pencil tests within the limits of differences in score precision. (Note that in IRT based computerized adaptive measurement, scores are estimated abilities.)

The adaptive test developed in this investigation produced scores with greater precision at the extremes of the ability distribution, and did this more efficiently in terms of testing time. Although branching decisions did not depend on scoring methods associated with latent trait theory, test construction and equating were facilitated through the use of relative efficiency curves and IRT true score equating. Nonetheless, all the methods employed in this development can be used in the absence of any IRT data; formulas for relative efficiency curves from conventional data are given in Lord (1980), and conventional equating methods are also applicable.

Conventional scoring methods were considered more appropriate for an adaptive version of TOEFL since IRT item parameters vary among language subgroups in the testing population. For purposes of score equating, item parameters are estimated on the TOEFL testing population as a whole, and have been shown to produce equating results that are robust to subgroup differences (Hicks, 1984). Observed subgroup variations in item parameter estimates precluded the use of IRT based branching methodologies.

Previous Research. The current development is an outgrowth of the extensive research of two-stage and multilevel testing, among others, which began in the 1970s, primarily at the University of Minnesota (Weiss, 1975). These methodologies were generally characterized by the administration of item sets such that the examinee would be given the most appropriate test level in terms of difficulty. In two-stage testing, for instance, the initial stage consisted of a routing test that attempted to determine the examinee's probable level of ability, and the second stage concluded with the administration of the appropriate test level given the results of the routing test. Many of these investigations usually consisted of two fixed-length tests. Lord has investigated this testing problem extensively, presenting many strategies for its solution (see Chapter 9, Lord, 1980).

During the rapid dissemination of sophisticated IRT measurement techniques in the years that followed, computer adaptive testing development utilized the statistical power of latent trait theory, which enabled branching decisions on the basis of a single item administration, and the interest in developing conventional adaptive measures became limited or virtually nonexistent. Indeed, many of the problems encountered in the research in two-stage testing and its variants also contributed to reduced interest in it as a viable method of testing; the major problem was the high rate of examinee misclassification in terms of the appropriate test level.

All misclassifications in a study of two-stage testing by Betz and Weiss (1973) resulted from misrouting of low-level examinees to test levels that were too difficult for them as a result of lucky guessing. Their routing test consisted of 10 items of mean difficulty .62, approximately normally distributed, with assignment to the test levels based on degrees of number correct. This rather straightforward approach produced 5% misclassifications.

Earlier studies of two-stage testing (Cleary, Linn & Rock, 1968a, 1968b) yielded misclassifications as high as 20%. Cleary et al. investigated several routing procedures, among them a test with a rectangular distribution of difficulties, a two-phase routing process, and a sequential testing procedure using the sequential probability ratio test, all based on 20 items. Most of these studies routed to four discrete test levels, providing greater opportunity for misclassification. That is, none of them included overlapping test levels at the boundaries of each level that could absorb many of the borderline cases. A simulation study, undertaken during the initial phase of this investigation, defined test levels that overlapped with the expectation that the misclassification rate would be reduced.

Procedures for designing multilevel tests were given by Lord (1974) in which he concluded that a three-level multilevel test discriminated as well as a four-level test, and better than a test with fewer than three levels. Overlapping score scales were an inherent feature of the multilevel design. Marco (1977) investigated multilevel testing in a paper and pencil mode using SAT mathematical items and also demonstrated the flexibility of IRT equating by pre-equating the shorter levels to the SAT scale. Since the examinees were required to branch themselves based on "guesstimates" of their performance, the not-unexpected result was a high proportion of branching errors, particularly among examinees at low ability levels. On the assumption that computerized administration of multilevel te ting could eliminate this source of error, this test design was also investigated during the initial phase of this study.

Results of Phase 1. During the first phase of this study, several methodologies for assigning individuals to overlapping levels of a test were evaluated in order to identify a testing strategy that minimized classification errors (Hicks, 1985). Various routing and branching tests were constructed using the operational and pretest items on one of the subtests at a large TOEFL administration. The sample consisted of all 1327 examinees taking the subtest. Computer programs were developed that simulated test administrations of several varieties, including two-stage and multilevel testing. Scores on the simulated adaptive tests were equated to the TOEFL scale via true-score IRT equating (Lord, 1980, chapter 13) and compared with scores obtained on the full length version. Examinees were assigned "t ue" test levels based on their scores on the full-length TOEFL, and classiication error was determined by the extent to which they failed to be assigned to these levels in any of the simulated testing schemes.

The results of the simulations demonstrated that it was possible to improve the rate of correct classification in two-stage testing over those reported in the literature if overlapping levels for the measurement tests were constructed, and if some form of sequential item presentation was used at the routing stage (in contrast to a fixed routing test). The routing test consisted of 10 items in these simulations. The rates of misclassification ranged from .90% to 1.96% for three overlapping levels. This contrasted with the 5%-20% range of misclassification rates cited in the literature for routing tests based on 10-20 items (usually routing to four discrete levels).

The most effective and satisfactory of all the methods investigated during Phase 1 was the computer administered multilevel test. The correlations between total scores on the multilevel tests and the regular TOEFL (.94-.95), and the close correspondence of the summary data (see Table 2, which lists data for two sections only) indicated that this method of testing assigned examinees to appropriate levels with a high degree of accuracy. While the branching criteria (i.e., the cutpoints) affected t' distribution of examinees among the levels, they did not impact on the overall results appreciably, demonstrating that the overlapping test levels can provide adequate measurement at adjacent ability levels. Indeed, the methodology easily handled examinees at the boundaries of the levels, the most difficult to classify in other methods.

The hierarchical administration of testlets recently recommended by Wainer and Kiely (1987) is substantively analogous to the multilevel testing paradigm adopted in this development. After considering some of the disadvantages of the sequential administration of individual items in computerized adaptive testing, Wainer and Kiely have recommended a hierarchical presentation of testlets as a better adaptive testing strategy. Distinct advantages of such methods over current adaptive testing procedures include the ability to control for many important aspects of test development, such as contextual effects and content. Wainer and Kiely also acknowledged that Fisherian information based on the item parameters has proved to be insufficient for effective test construction in current adaptive applications.

Phase 2 Objectives. Based on the foregoing results, computerized multilevel tests were developed for TOEFL Section 2 (Structure and Written Expression) and TOEFL Section 3 (Vocabulary and Reading Comprehension) during the second phase of the study. The testing strategy produced three levels of the test, each equated to the TOEFL scale. Test construction procedures described below illustrate the control that was exerted over important aspects of the tests, such as assuring that each test level was developed according to similar content and statistical specifications, at the same time maintaining equivalent measurement efficiency across all levels. Some desirable features of the score scales obtained in this mode of testing are discussed. This report primarily describes the TOEFL computerized tests developed using the multilevel paradigm, and presents some preliminary validation data, as well as the reactions of ESL staff and students to the computerized testing experience.

SPECIFICATIONS OF THE COMPUTERIZED TESTS

Branching Algorithm. Three levels of Sections 2 and 3 (Structure and Written Expression, Vocabulary and Reading Comprehension) of TOEFL were generated utilizing five item blocks (or testlets) of varying difficulty as follows:

$b < -.85$	$-.84 < b < -.25$	$-.24 < b < .25$	$.26 < b < .85$	$b \geq .86$.
(A)	(B)	(C)	(D)	(E)

TOEFL b-parameters, the IRT index of difficulty, range from approximately -2.5 to +2.5; items of middle difficulty are in the range indicated by difficulty level C. Item parameters used to construct the computerized tests were those from the existing item pool, estimated on the paper-and-pencil versions, a necessary point of departure in developing the computerized tests. The questions associated with the statistical validation of the computerized version (i.e., the correspondence of constructs and item and score data between the paper and pencil and computerized tests) are discussed on page 26.

In general, the number of items in each item block was determined by the desired number of test levels and total number of items. For Section 2, each item block consisted of six items. Starting at level C, the examinee was branched up to the next unattempted level if four or more C-level items were answered correctly, and branched down to the next unattempted level for three or fewer correct responses. Testing was completed after three item blocks were administered, resulting in three test levels: CBA (Test Level 1), CDB or CBD (Test Level 2), and CDE (Test Level 3). Examinees were also presented two additional items from the item blocks not administered in order to facilitate the equating. Each item block consisted of two structure items and two written expression items in roughly the same proportions as on the regular test. The total number of items administered in Section 2 was 20, or one halt the total number of items on the regular TOEFL.

For Section 3, each item block consisted of 10 items: the examinee was branched up if six or more items were answered correctly, and down for five or fewer correct responses. Four additional items were administered to improve the equatability of the levels. A total of 34 items was administered in Section 3 (the full-length test contains 60 items), and each block consisted of five vocabulary items and five reading comprehension items reflecting the same proportions as on the regular test. The cut point for these branching criteria was based on the average proportion of correct responses corresponding to the mean score for five recent TOEFL forms, which was determined to be .675 for Section 2 and .60 for Section 3.

These results are summarized as follows:

	Section 2	Section 3
Total No. Items	20	34
No. Items in Each Content Category	Struct. 6 (15)* Wr.Expr. 12 (25)*	Vocab. 15 (30)* R.Comp. 15 (30)*
No. Equating Items	2	4
No. Items in Each Block	6	10
Branching Criterion	.675	.60

*Number in parentheses = number of items on the regular TOEFL.

The three test levels resulting from the hierarchical administration of the item blocks were

```
                              Item Blocks

                  A          B          C          D          E
Test Level 1      [=====================]          *          *
Test Level 2      *          [=====================]          *
Test Level 3      *          *          [=====================].
```

The asterisks represent the few additional items administered for equating purposes.

Backtracking. Green (1988) noted that the computer cannot conveniently allow examinees to review questions, change their answers, or postpone trying to answer questions. While this may be a difficulty with sequential item administrations, backtracking is easily accommodated when the test is administered in the form of item blocks. In this development, the examinee had the option to review and change answers at the end of the administration of each block of items or testlet.

Maximizing Discrimination at Each of the Test Levels. The item pool consisted of 160 items for Sections 2 and 3--60 Section 2 items and 100 Section 3 items. From this pool, in which the proportions of item types found in the regular TOEFL were maintained, two testing sequences were constructed, Test A and Test B, each sequence yielding three test levels for each of the sections. The IRT index of discrimination, the a-parameter, ranges from zero to 1.5 for TOEFL items, but was constrained to be greater than or equal to 1.00 for inclusion in the pool in order to increase measurement efficiency at each of the levels. Table 1 on page 9 presents the means and standard deviations of the a- and b-parameters for each difficulty level for the two testing sequences developed for this project.

Within the limits of the pool, an attempt was made to equalize the distribution of the a-parameters across item blocks and to match a's and b's across testing sequences (Tests A and B). The effect of this type of test construction was to maintain the same level of measurement effectiveness across the levels--that is, across the whole spectrum of scores--and to produce comparable test levels no matter what testing sequence was used (Test A or B). This is not the case for the regular TOEFL, where score precision varies and is greatest at the middle range of ability. Since the correlation between a's and b's on the regular TOEFL is comparatively low (.25), this constraint did not impact on the dimensionality of the computerized test. Indeed, as explained above, each examinee was administered a few "equating" items from item blocks to which he or she was not exposed, which had the effect of exposing each examinee to the full spectrum of difficulty of the test in proportion to his or her ability.

Interpreting TOEFL Relative Efficiency Curves. The relative efficiency curve (REC), the ratio of the information curve for the current test to that for a comparison test (Lord 1980, chapter 6), provides information regarding the level of relative measurement effectiveness at each scaled score. The information curve is a ratio of the slope of the test characteristic curve to the standard error of measurement (both terms squared and expressed as functions of ability). Variations in the slopes of the test characteristic curve are due, in large part, to variations in the a-parameters over ability. Since, for TOEFL, the relationship between ability and item discrimination is minimal and comparable across forms, the greatest effect on the information curve is the standard error of measurement at each ability level. Thus, tne RECs (for TOEFL) are reflecting the relative measurement error at each ability for the two tests being compared.

RECs for the three test levels of both testing sequences are given in Figures 1 and 2. For TOEFL, the comparison test is the form to which the current test is being equated. The points on the abscissa corresponding to the intersection of the curve with the horizontal line (equal to 1.00 on the ordinate) indicate where the current test and base form share the same level of measurement effectiveness. In terms of the discussion above, this is where the two versions of TOEFL have relatively the same measurement error. When the REC is greater than 1.00, the current test--the test constructed for the computer in this instance--is exhibiting smaller measurement error than the comparison test (a version of the regular test in this example). It can be noted that the curves do not extend below scores at about 30, which corresponds (approximately) to the chance level on TOEFL; scores below this point cannot be derived from the test characteristic functions, but are calculated from a line relating the c-parameters of the two tests.

The REC for Level 2 is typical of those observed for the regular TOEFL 2 and reflects the fact that the test is constructed for maximum discrimination over the middle range of language proficiency; it is peaked there. Levels 1 and 3, however, yield comparable measurement effectiveness in the multilevel testing paradigm. This is mainly due to the fact that more items are clustered at these levels than on a regular TOEFL. An important implication of the increased efficiency of measurement at the extremes of ability is that there will be less comparability between scores on the regular and computerized TOEFL at these points, but the reduced comparability is due to the increased accuracy of the computerized version at these levels of ability.

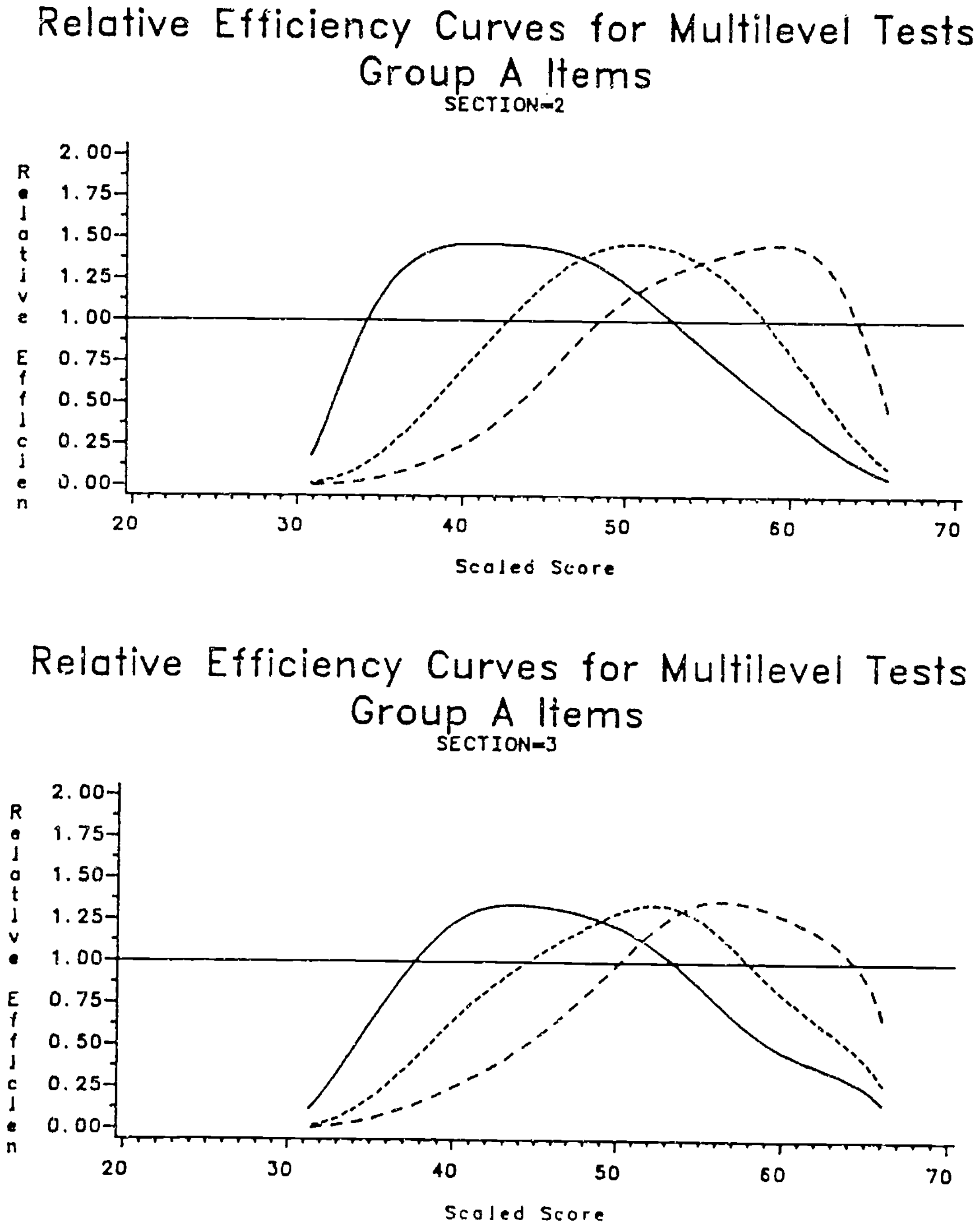
The use of RECs is critical to the construction of optimal test levels, but their use is not restricted to IRT data. Lord (1980) has provided formulas for their computation based on conventional test data.

Table 1

Means and Standard Deviations of a- and b-parameters by Difficulty Level

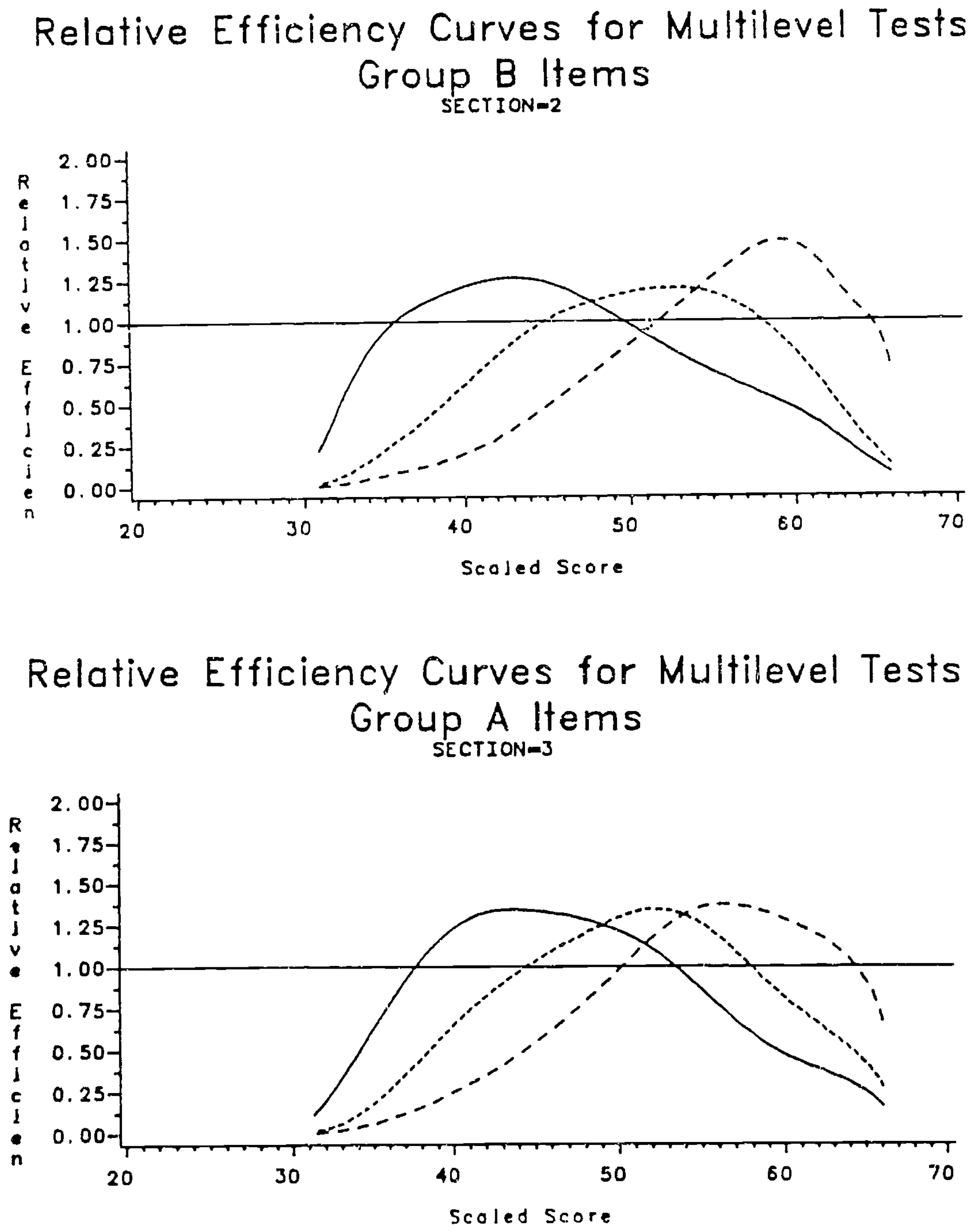
	Section 2		Section 3	
Level	Test A	Test B	Test A	Test B
		b-parameters		
A	-1.31	-1.33	-1.19	-1.19
	.25	.23	.33	.29
B	- .51	-.55	-.58	-.52
	.19	.20	.22	.16
C	.04	-.02	.05	.04
	.16	.12	.17	.17
D	.59	.58	.58	.58
	.18	.19	.18	.16
E	1.34	1.45	1.24	1.23
	.29	.48	.24	.18
		a-parameters		
A	1.30	1.11	1.23	1.25
	.15	.13	.16	.16
B	1.21	1.25	1.33	1.28
	.19	.20	.15	.19
C	1.47	1.16	1.21	1.36
	.05	.10	.10	.17
D	1.33	1.41	1.25	1.18
	.13	.14	.18	.16
E	1.37	1.50	1.24	1.23
	.13	.00	.19	.19

Figure 1. Relative efficiency curves, Test A.


Relative Efficiency Curves for Multilevel Tests
Group A Items
SECTION=2

Relative Efficiency Curves for Multilevel Tests
Group A Items
SECTION=3

Figure 2. Relative efficiency curves, Test B.



Relative Efficiency Curves for Multilevel Tests
Group B Items
SECTION=2

Relative Efficiency Curves for Multilevel Tests
Group A Items
SECTION=3

Sections 2 and 3 of the standard TOEFL contain 40 and 60 items, respectively; the scaled score ranges for these sections is 20 to 68 for Section 2 and 20 to 67 for Section 3, the scales to which the test levels were equated. The range of maximal discrimination (>1) for each test level was:

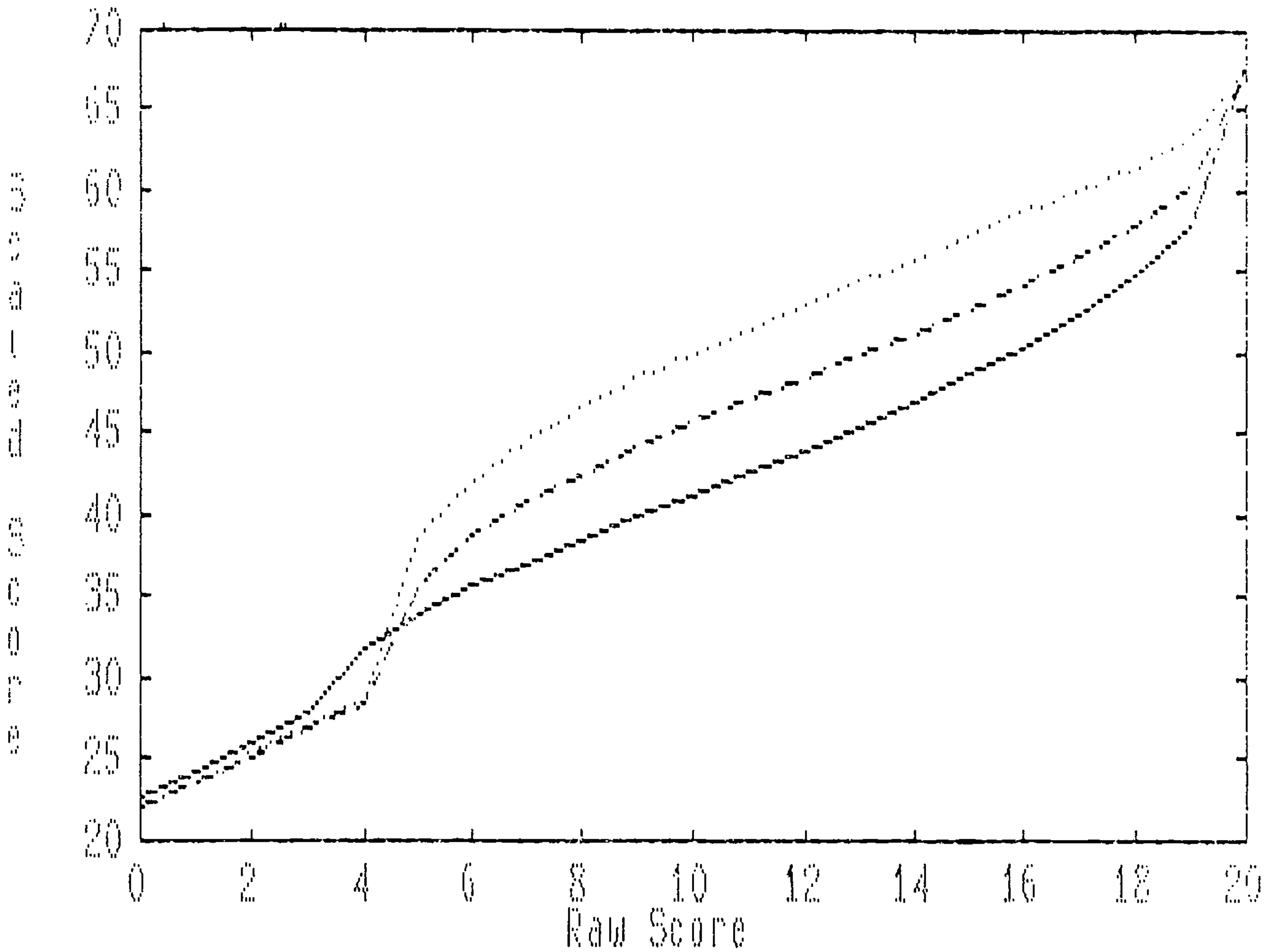
	Test A		Test B	
Test Level	Sec. 2	Sec.3	Sec.2	Sec.3
1	34 - 53	37 - 54	35 - 50	36 - 55
2	43 - 58	44 - 58	45 - 58	44 - 59
3	48 - 64	50 - 64	52 - 65	50 - 64.

A basic feature of adaptive measurement is the more equitable distribution of the precision of measurement over the score scale. From the relationships among the curves in Figures 1 and 2, it can be seen that this has been achieved in this testing algorithm.

Scoring the Test Levels. The objective of the testing methodology in this development was to branch examinees through item subsets of appropriate difficulty such that one of three possible test levels was administered. The raw scores obtained on the test levels were equated to the TOEFL scale since each of the levels varied significantly in difficulty. Utilizing true score IRT equating methods (Lord, 1980, chapter 13) in which abilities rather than scores are equated, shorter tests can be scaled to longer measures, a process that is problematic with conventional equating methods. In addition, the flexibility accorded by pre-equating, where the equating functions are derived from existing item parameters and do not require examinee data, facilitated the scoring process. The equating curves for Section 2, Test A, are presented in Figure 3 to illustrate the continuity of the conversions across test levels. A raw score of 10, for instance, converts to 41 on Test Level 1, to 46 on Level 2, and 50 on Level 3, reflecting the differences in the difficulty of the three tests.

Even though the equating curves indicate that a perfect score on Test Level 1 converts to the maximum scaled score, this maximum cannot be obtained on a Level 1 test. This is so because of the constraints of the minimum and maximum scores possible inherent in the branching algorithm. Only the Level 3 test produced the top converted score.

Figure 3. Equating curves, Section 2, Test A.



The minimum and maximum possible raw and converted scores for Section 2 were as follows:

	Raw	Converted Score Range	
	Score Range	Test A	Test B
Level 1	0-15	23-49	21-48
Level 2	4-17	29-56	27-56
Level 3	8-22	47-68	46-68.

For Section 3, the minimum and maximum possible raw and converted scores were:

	Raw	Converted Score Range	
	Score Range	Test A	Test B
Level 1	0-24	22-49	23-50
Level 2	6-29	29-58	30-58
Level 3	12-34	46-67	45-67.

The constraints of the testing algorithm effectively place limits on the scores obtained in any level; even though the equating function provides converted scores that may be meaningless in terms of the actual ability on a test level (for instance, a raw score of 20 on Level 1 converts to the top score), such scores cannot be achieved. Comparisons of scaled scores obtained on the computerized and regular tests should be interpreted in this light. For instance, examinees taking Level 1 and Level 2 tests are restrained from lucky guessing on difficult items and, in isolated cases, may be prevented from attempting difficult items that they actually know, which might conceivably occur in the vocabulary subsection; this would, in turn, depress the computerized test score relative to the regular test.

The overlapping converted score ranges avoid misclassification problems generally encountered in testing methodologies of this type with the result that the levels themselves cannot be used for classificatory purposes. That is, individuals on the borderlines of the test levels can be tested efficiently on either level. To illustrate the measurement properties of tests constructed in this fashion, the overlapping scales for Section 3 , Test A, are shown below by the double lines. The range of scores where the relative efficiency is greater for the current test relative to the base form is indicated by the single lines enclosed in brackets.

Limits of the Converted Score Ranges and Ranges of Relative Efficiency >1
Three Test Levels, Section 3

```
22            [-----------49
=============================  Test Level 1

       29            [------------ 58]
       ===============================  Test Level 2

                         46  [-------------] 67
                         ======================    Test Level 3
```

The wider ranges of scores for Levels 1 and 2 are due to below-chance level scores that can occur at these levels, but not at Level 3. In any case, the diagram illustrates the overlapping scales at the boundaries of the test levels, and that efficient measurement is maintained over their intersections.

Comparisons of Multilevel and Regular TOEFL Scaled Scores. There are at least two reasons for observed differences in the scaled scores obtained from a computerized multilevel test and the full-length test: (1) the inability of lower-level students to make lucky guesses on difficult items in the computerized version and (2) the reduced error of measurement on the multilevel test for extremes of the score scale where the examinee is administered more items at these levels than would be the case on the regular test. These results are demonstrated and implied in Table 2, which presents data from the simulation study. Listed there are means and correlations for two simulations based on different branching criteria. Criterion 1 was similar to the one used in the development of the current tests; the second criterion was easier, requiring fewer correct responses per item block to advance. The data in Table 2 indicate that the simulation tended to result in lower Section 3 scores on the multilevel test than on the regular TOEFL for the least able examinees, and that the correlations between these sets of scores were among the lowest, ranging from .73 to .84.

Data available for a small group of beginning students (n=22) who took the computerized test and for whom TOEFL section scores were reported are given below:

Table 2

Means and Correlations for Multiievel Simulations
Two Routing Criteria (Cutpoints)

	Section 2 Structure and Written Expression			Section 3 Vocabulary and Reading Comprehension		
	TOEFL	Multi	r	TOEFL	Multi	r
Level 1						
Crit. 1	41	41	.84	41	40	.79
Crit. 2	39	38	.79	39	37	.73
Level 2						
Crit. 1	49	49	.79	49	49	.83
Crit. 2	46	46	.80	47	46	.83
Level 3						
Crit. 1	58	59	.85	57	58	.85
Crit. 2	55	56	.88	55	56	.88

	Regular TOEFL	Computerized TOEFL	r
Sec.2 Mean	47.71	46.73	.86
SD	(8.0)	(9.1)	
Sec.3 Mean	47.86	45.82	.72
SD	(7.1)	(7.7)	

The patterns of means and the range of correlations are comparable to those obtained in the simulation and appear to be what might be expected from this relatively low (Levels 1 and 2) scoring group. Indeed, the correlations are in the range of those observed in comparable studies. Olsen (1986) reported correlations ranging from .76 to .79 between equated scores for computerized adaptive and paper-and-pencil mathematics tests administered to sixth graders; correlations between .83 and .89 were observed for third graders.

PROGRAMMING THE TEST

Programming of the test was implemented through EASIS, an ETS test authoring system. The test is housed on a floppy disk and can accommodate 125 examinees. Each examinee record consists of the usual data collected in a regular TOEFL administration (e.g., native language and country codes, gender). After the TOEFL logo is presented, a frame follows requesting that the proctor indicate a choice of Test A or Test B. The examinee is then paced through a keyboard familiarization segment. Part of this familiarization process consists of practice with the F1 and F2 keys, which are used in the reading comprehension section to page forward and backward through long passages. This feature was included experimentally in response to the observation that computerized language testing did not adequately assess reading comprehension since short passages were required to accommodate a single frame. The ability to backtrack, as described above, was incorporated into the testing procedures.

After these preliminaries, the examinee is routed to instructions for Section 2, which parallel those presented in the regular test. After this introduction, the test-taker is branched through the test. Section 3 is then administered in the same fashion, and a printout of the raw and converted scores completes the test administration.

The test was programmed for an IBM PC or compatible with DOS 2.0 or higher. A Hercules graphics or CGA board is also required.

FIELD TEST RESULTS

Preliminary Field Test. A preliminary field test at Brigham Young University included the administration of a shortened version of Section 1, Listening Comprehension, along with Sections 2 and 3. The Listening Comprehension test was not adaptive, but consisted of 25 items equated to the TOEFL scale. Examinees listened to test questions on a cassette and were required to pace themselves through the test by pressing keys on the computer in response to the verbal instructions on the tape. If an examinee experienced difficulty with the keys, the pace was lost. While such an administration might be feasible for those with computer experience (no difficulties at all were encountered by statistical assistants who tried this section), it proved to be extremely difficult for the inexperienced. As will be shown below, most of the testing sample had never touched a computer before. The administration of Listening Comprehension proved to be so problematic that it was decided to eliminate it from the final version of the test.

During the preliminary field trials, the tests were timed, but the timing instructions and time reminders added to test anxiety, among other things; thus, it was decided to remove timing from the final research version. In addition, it is not clear that the same time constraints should hold for both computerized and paper-and-pencil tests (Green, 1988). No problems associated with the use of the F1 and F2 keys were reported at this administration.

Final Field Test Samples. After completing the final version of the test, it was planned to administer the test to 250 examinees in order to develop some validating information as well as to procure feedback on examinee reaction to computerized testing. In the fall of 1986, requests to participate in such a field test were sent to 20 institutions but no responses were received. A second mailing at the beginning of 1987, in which examinees were offered a payment of $15 to take the test, did elicit one response, from the University of Toledo. These examinees were volunteers who responded to a sign posted by the test administrator. A second group of examinees, students enrolled in ESL classes at UCLA Downtown, was secured in late fall 1987, and were mostly beginning students. The total number of examinees obtained from the two institutions was 162; 90 from the University of Toledo and 72 from UCLA.

The advantages of administering a test by computer are many, but their limitations include the need for a bank of computers for efficient testing. In addition, in a field testing situation substantial time on the part of the test administrator or proctor is also required. Neither of these were apparently in abundant supply among the ESL departments as, understandably, busy staff cannot take the time to monitor these individually administered tests (the average time for the test administration was 45 minutes, but many examinees took much longer). Large samples from computerized testing required for item analyses, among other things, must be collected either by special administrations monitored by research perfonnel or slowly over time. Hardware requirements for the test (IBM compatible PC and Hercules board) may have also eliminated some potential participants.

Results of the Questionnaire. The questionnaire administered to the examinee sample and the percentages responding to each option are given below. All of the UCLA examinees responded since it was required that they fill it out immediately after taking the test. On the other hand, the questionnaire was mailed to the Toledo sample after the testing was completed and the data received, resulting in a response rate of 54%, thus, the results are based on a total of 121 examinees.

1. Have you ever used a computer before taking the test?

Yes	28%
No	72%

2. Did you find the instructions easy to understand?

Easy	48%
Somewhat easy	25%
Difficult	27%

3. Do you prefer taking a test by computer to taking a paper-and-pencil test?

Prefer paper and pencil test	41%
Doesn't make any difference	25%
Prefer a test by computer	34%

4. In the reading section you had to press two special keys in order to page forward and backward on some long reading passages. Was this confusing for you?

Not at all	48%
Somewhat confusing	33%
Very confusing	19%

5. How difficult did you find Section 2, Structure and Written Expression?

Very easy	3%
Somewhat easy	10%
Just right for me	24%
A little difficult	29%
Very difficult	34%

6. How difficult did you find Section 3, Vocabulary and Reading Comprehension?

Very easy	5%
Somewhat easy	7%
Just right for me	15%
A little difficult	37%
Very difficult	36%

7. Do you think the score you received on the TOEFL computerized test is an accurate estimate of your English language proficiency?

Yes	65%
No	35%

The most surprising result of this brief survey was the large percentage of the sample who had never used a computer before--almost three-quarters of the group. Given this result, it is also of interest that only 41% indicated a preference for a paper-and-pencil test; for 59% of the group the method of testing was immaterial or a computerized test was preferred. These results were somewhat unexpected given the anxiety and emotional reactions that were reportedly observed during testing at Toledo, in particular. While only 19% noted that the F1 and F2 keys were confusing, the test administrators indicated that many examinees experienced great difficulties with them.

Large percentages of the examinees found both tests a little, or very difficult, which may be due to the low ability levels represented in these samples. For the relatively few examinees at the middle and upper levels of ability, all found the test to be easy to just right; this probably reflects the fact that beginning students would find a comparatively easy test somewhat difficult. These results might be contrasted with the 65% who found the scores to be an accurate estimate of their language proficiency; however, responses to this question are probably of low validity in this instance since most of the examinees had never taken TOEFL before and were not sufficiently informed about the TOEFL scale relative to their own performance.

A final question requested TOEFL score information if available. Only 17 examinees at Toledo and 5 at UCLA had TOEFL scores. The TOEFL scores from Toledo were self reported. TOEFL score data from the UCLA sample were from a recently administered Institutional TOEFL (a regular TOEFL administered under institutional auspices), and were provided by the test administrator.

Score Data from the Final Field Test. Means and standard deviations for the sections and test levels are given in Table 3. The rather large differences in means between Sections 2 and 3 in Table 3 indicated that some problems were encountered in Vocabulary and Reading Comprehension. In this group, Section 3 means are depressed compared to those of Section 2, while typically these means for domestic groups are very comparable. Lower Section 3 means relative to those of Section 2 were expected with low scoring examinees based on the simulation data (Table 2), but it is likely that some scores may have been contaminated by difficulties with the paging keys. When one considers that most of the examinees were confronting a computer for the first time, and taking a test at that, the difficulties imposed by having to manipulate additional, nonmeaningfully coded keys were probably too taxing for some examinees.

ESL students at UCLA are placed in class levels that range from 100 to 106, and are commensurate with ESL proficiency as measured by a screening test. Correlations between Institutional TOEFL scores and these levels were obtained for 17 examinees (some of whom were not included in the computerized test sample). These were found to be .87 for Listening Comprehension, .68 for Section 2, and .73 for Section 3. For the total TOEFL score, the correlation with class level was .84. The correlations with class level for the scores on the computerized test appear to be comparable with those obtained from Institutional TOEFL scores, as shown in Table 4. While for this small sample Listening Comprehension predicts class level at UCLA better than the other sections, the computerized TOEFL apparently does as well as the regular TOEFL for the other sections.

Reactions of Test Administrators. Staff members at the University of Toledo and UCLA were asked to evaluate the test and the testing situation. Most reacted positively to the former and expressed reservations about the latter. A faculty member at Toledo said the program was good but that a student's computer literacy would markedly affect the score, and recommended a dry-run test on the computer before taking the actual test. She reacted positively to instant score reports, and suggested that instructions for use of the computer might be provided in different languages. She felt the paging back and forth in the reading section was disruptive in many ways, even affecting train of thought.

Another University of Toledo faculty member felt that for a student without previous computer experience, the computerized test is not a valid measure of English and, in fact, assesses other factors, such as computer knowledge, familiarity with a typewriter keyboard, ability to overcome fear, to reason and to remember in an unfamiliar environment, and even manual dexterity. She expressed the view that a student unfamiliar with computers will be intimidated by them, and that this could range from resentment about being put in an unusual inequitable situation to almost paralyzing terror. It would be more difficult for the student to concentrate, which would be reflected in a lower score. Because of unfamiliarity with the keyboard, the student would spend relatively more time on each question, searching for keys, making mechanical errors in the attempt to correct them using time required for answering questions; also nervousness would increase and affect performance even more. Minor unaccustomed factors would also affect performance--for example, the glare of the screen, the inability to put

Table 3.

Means and Standard Deviations for Two Groups of Examinees:
TOEFL Sections 2 and 3 and Test Levels

	TOEFL Section 2	Computerized Test Level	TOEFL Section 3	Computerized Test Level
Toledo	49.42	2.05	46.42	1.9
N=90	(7.4)	(.6)	(7.7)	(.8)
UCLA	44.18	1.65	42.76	1.6
N=72	(7.3)	(.6)	(8.0)	(.7)

Table 4

Correlations of Institutional TOEFL and Computerized TOEFL Scores
with UCLA ESL Class Level

	Institutional TOEFL n=17	Computerized TOEFL n=72
Sec. 2	.68	.74
Sec. 3.	.73	.69

the chair in just the right position, the noise of the computer signals, the movements of the test administrators as they tried to help those having problems with the computers--all in notable contrast to the controlled quiet and calm of the regular TOEFL-taking atmosphere.

These comments give a flavor of the atmosphere that probably pervaded the testing process, and from verbal reports from the test administrators, it is not difficult to believe that some examinees were extremely threatened by the experience. In the total sample of examinees, only 28% had previous experience with computers, and it is likely that considerably more American students have had this experience by the time they reach the college level. But these reports make the results all the more impressive, since except for the possible contamination of some Section 3 scores by the difficulties with the F1 and F2 keys, the scores were quite reasonable, and for the few cases with TOEFL scores, the computerized test tended to produce expected results. With the exception of the F1 and F2 keys, the examinee need only press the A, B, C, D, Enter, and Backspace keys to take the test; thus, typing skill is not seriously tried in this process. Since the tests were not timed, examinees were not penalized for difficulty in finding keys.

The observation that those with computer experience (or those without it who might tend to adapt easily to it) would probably possess an advantage is certainly reasonable. Yet with proper orientation, a computerized test may prove to be valid for testtakers without previous computer exposure. As computers become more widespread in the instructional environment, future ESL students will probably bring computer skills with them in greater numbers.

A faculty member at UCLA provided many helpful technical comments regarding the test presentation, which included suggested improvements in the keyboard familiarization phase. He also suggested that the language and country codes should be presented on the screen rather than having the student secure this information prior to the test. The test was developed in color and in monochrome, and, apparently depending on the machine, the color combinations were not standard. In general, the monochrome version was preferred.

DISCUSSION

Implications for Further Development and Test Use. The psychometric evaluation of these preliminary data indicated that the test functioned well, and, when interpreted properly, the scores can provide a satisfactory estimate of regular TOEFL performance. Because the computerized test provides more accurate measurement at the tails of the score scale, it might be argued that the regular TOEFL is less fair for these examinees. Test administration concerns, however, do not make the use of the computerized version a practical alternative to the regular TOEFL given the volume of the TOEFL test population. The paper-and-pencil version will remain the most efficient way of testing for the foreseeable future, but the computerized test can exist as a useful alternative in special situations.

One of the most important outcomes of this initial study of computerized testing of ESL students, with implications for any further development in this area, is the widespread lack of computer experience, far more widespread than would be expected among American examinees. The major question appears to be whether administering a computerized test to a population of examinees with little or no computer experience, yields a fair estimate of performance. While high levels of anxiety were apparently registered at the testing centers, the surprising number of positive or neutral response to computerized testing by those who had never touched a computer before suggests that this might be possible with proper preparation.

If a computerized test is to be developed for new ESL students, a familiarization disk should accompany the test, something like the samples of test questions included in the TOEFL Bulletin of Information. Prior practice is even more necessary in this mode of testing.

Advantages of an Adaptive Conventional Test. An adaptive version of a conventional instrument provides the user with an alternative, efficient test with most of the properties of the conventional instrument. The score scale and test specifications are maintained to a reasonable extent, and interpretation of the results from either version are comparatively similar. The test user will need to bear in mind that score differences between the two versions for Level 1 and Level 3 examinees are to be expected by virtue of their different measurement properties.

The ability to control important facets of test construction and measurement effectiveness in the testing paradigm of this study was illustrated. The advantages of computerized adaptive testing based on a hierarchal administration of testlets as they impact on control of test specifications and contextual effects were described by Wainer and Kiely (1987) who recommended them for computerized test development. In the current development, the option to backtrack and change answers, possible in paper-and-pencil tests but problematic with standard computerized adaptive algorithms, was easily implemented in this mode of item presentation. The testing algorithm, along with the equated scores, were shown to produce score scales for each of the test levels with desirable properties, namely, overlapping scales at the boundaries of the levels and limits on the scaled scores obtainable on each of the levels.

Future Research. The test disk produced by this study could be a potential source of future research in English language proficiency. It can enable the investigation of diverse item types with greater breadth and depth of language assessment, evaluating these items in the presence of existing TOEFL items by simply adding a module to the disk. It is possible that in the decades to come, widespread computerized testing will be a practical al ernative, or even the testing method of choice. Rather than delivering a facsimile of the current TOEFL, more sophisticated language assessment will be possible, and this disk, or one like it, can serve as a connection between the current test and its scale, and new measurement approaches.

Recommendations for the Use of the Current Computerized Test. A complete evaluation of the TOEFL Computerized Placement Test would include a comparison of item data based on computerized administrations with those from paper-and-pencil tests. The sample required for this statistical validation would be formidable. In another ETS computerized test development project, ETS research staff conducted these administrations on site. To develop the minimum required sample of 1,000 examinees per item for IRT parameter estimation for Test A alone would require 3,000 administrations of the test, the sample equally distributed among the levels. To estimate parameters for all the items in this small pool (Tests A and B) would require 6,000 computerized test administrations. Since a relatively small proportion of the examinees are likely to test at Level 3, even 6,000 administrations would probably not suffice; a more practical estimate is 7,500 - 8,000 in order to collect sufficient data for parameter estimation.

This objective would seriously impede practical use of the test for some time; however, there is evidence that scales of difficulty hold across the testing modes. A factor analytic study of a regular battery of the Armed Services Vocational Aptitude Battery (ASVAB) and a computerized adaptive version determined that the factor structure was the same for both versions of the test (Green, 1988). The expected correspondence between computerized and regular TOEFL section scores observed even for the limited sample in this study suggests that this may also be the case for the current test. Given the foregoing, the test might be used operationally as a placement test under the following conditions:

1. The user is informed that the test has been constructed using item statistics obtained in a paper-and-pencil mode, and that previous studies have indicate' that these statistics hold across testing modes. Until it is possible to collect all the relevant data, this is assumed to be the case for the TOEFL computerized test.

2. Until the correspondence between paper-and-pencil and computerized item data is established, or actual computerized item data are available, a point estimate of a TOEFL score would not be reported. Instead, the user would be provided with a score interval around the equated score, the limits of which would be plus and minus one standard error of measurement (one standard error of measurement is approximately 2 points,. Reporting a score band would discourage possible use of the computerized test as a substitute for TOEFL, and focus on its purpose as a placement or screening instrument.

Test use of this kind would serve at least two important purposes: the test as a useful screening instrument, yielding conservative estimates of TOEFL performance, would be immediately available to users, and the large amount of data necessary to establish it as a test in its own right (in terms of point estimates of English proficiency) could be collected.

REFERENCES

Betz, N. E. & Weiss, D. A. (1973). *An empirical study of computer-administered two-stage ability testing*. (Research Report 73-4). Minneapolis: University of Minnesota, Department of Psychology.

Cleary, T. A., Linn, R. L. & Rock, D. A. (1968a). An exploratory study of programmed tests. *Educational and Psychological Measurement*, *28*, 345-360.

Cleary, T. A., Linn, R. L. & Rock, D. A. (1968b). Reproduction of total test score through the use of sequential programmed tests. *Journal of Educational Measurement*, *5*, 183-187.

Green, B. F. (1988). Construct validity of computer based tests. In *Test Validity*. H. Wainer & H. I. Braun (Eds.). NJ: Lawrence Erlbaum Associates.

Hicks, M. M. (1984). A comparative study of methods of equating TOEFL test scores. RR-84-20. Princeton, N.J.; Educational Testing Service.

Hicks, M. M. (1985). *Computerized multilevel testing; a rapid screening methodology*. Seventh Annual Language Testing Research Colloquium, April, 1985.

Linn, R. L., Rock, D. A., & Cleary, T. A. (1969). The development and evaluation of several programmed testing methods. *Educational and Psychological Measurement*, *29*, 129 - 146.

Lord, F. M. (1974). Practical methods for redesigning a homogeneous test, also for designing a multilevel test. Research Bulletin 74-30. Princeton, NJ Educational Testing Service.

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

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, *14*, 139-160.

Olsen, J. B. (1986). Comnarison anc' equating of paper administered, computer administered and computerized adaptive tests of achievement. Paper presented at the American Educational Research Association Meeting, San Francisco, April 1986.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. Journal of Educational Measurement, 24, 185-201.

Weiss, D. J. (1975). Adaptive testing research at Minnesota - overview, recent results and future directions. In D. J. Weiss (Ed.), Proceedings of the first Conference on Computerized Adaptive Testing. Personnel Research and Development Center. U.S. Civil Service Commission.

TOEFL Research Reports currently available...

Report 1. *The Performance of Native Speakers of English on the Test of English as a Foreign Language* John L D Clark November 1977

Report 2. *An Evaluation of Alternative Item Formats for Testing English as a Foreign Language* Lewis W Pike June 1979

Report 3. *The Performance of Non-Native Speakers of English on TOEFL and Verbal Aptitude Tests* Paul J Angelis Spencer S Swinton, and William R Cowell October 1979

Report 4. *An Exploration of Speaking Proficiency Measures in the TOEFL Context* John L D Clark and Spencer S Swinton October 1979

Report 5. *The Relationship between Scores on the Graduate Management Admission Test and the Test of English as a Foreign Language* Donald E Powers December 1980

Report 6. *Factor Analysis of the Test of English as a Foreign Language for Several Language Groups.* Spencer S Swinton and Donald E Powers. December 1980.

Report 7. *The Test of Spoken English as a Measure of Communicative Ability in English-Medium Instructional Settings* John L. D Clark and Spencer S Swinton December 1980.

Report 8. *Effects of Item Disclosure on TOEFL Performance* Gordon A. Hale, Paul J. Angelis, and Lawrence A Thibodeau. December 1980

Report 9. *Item Performance Across Native Language Groups on the Test of English as a Foreign Language* Donald L Alderman and Paul W Holland August 1981

Report 10. *Language Proficiency as a Moderator Variable in Testing Academic Aptitude* Donald L Alderman November 1981

Report 11. *A Comparative Analysis of TOEFL Examinee Characteristics. 1977-1979.* Kenneth M Wilson. July 1982

Report 12. *GMAT and GRE Aptitude Test Performance in Relation to Primary Language and Scores on TOEFL* Kenneth M Wilson July 1982.

Report 13. *The Test of Spoken English as a Measure of Communicative Ability in the Health Professions Validation and Standard Setting* Donald E. Powers and Charles W Stansfield January 1983.

Report 14. *A Manual for Assessing Language Growth in Instructional Settings* Spencer S Swinton February 1983

Report 15. *Survey of Academic Writing Tasks Required of Graduate and Undergraduate Foreign Students* Brent Bridgeman and Sybil Carlson September 1983.

Report 16. *Summaries of Studies Involving the Test of English as a Foreign Language, 1963-1982.* Gordon A Hale. Charles W Stansfield. and Richard P Duran. February 1984

Report 17. *TOEFL from a Communicative Viewpoint on Language Proficiency A Working Paper.* Richard P. Duran. Michael Canale, Joyce Penfield. Charles W Stansfield. and Judith E Liskin-Gasparro February 1985.

Report 18. *A Preliminary Study of Raters for the Test of Spoken English* Isaac I Bejar. February 1985.

Report 19. *Relationship of Admission Test Scores to Writing Performance of Native and Nonnative Speakers of English* Sybil B. Carlson. Brent Bridgeman, Roberta Camp. and Janet Waanders. August 1985.

Report 20. *A Survey of Academic Demands Related to Listening Skills* Donald E. Powers. December 1985.

Report 21. *Toward Communicative Competence Testing. Proceedings of the Second TOEFL Invitational Conference* Charles W Stansfield. May 1986

Report 22. *Patterns of Test Taking and Score Change for Examinees Who Repeat the Test of English as a Foreign Language.* Kenneth M. Wilson. January 1987.

Report 23. *Development of Cloze-Elide Tests of English as a Second Language.* Winton Manning. April 1987.

Report 24. *A Study of the Effects of Item Option Rearrangement on the Listening Comprehension Section of the Test of English as a Foreign Language.* Marna Golub-Smith. August 1987.

Report 25. *The Interaction of Student Major-Field Group and Text Content in TOEFL Reading Comprehension* Gordon A Hale. January 1988.

Report 26. *Multiple-Choice Cloze Items and the Test of English as a Foreign Language.* Gordon A. Hale. Charles W Stansfield. Donald A Rock. Marilyn M. Hicks. Frances A. Butler. and John W. Oller, Jr March 1988.

Report 27. *Native Language. English Proficiency, and the Structure of the Test of English as a Foreign Language* Philip K Oltman. Lawrence J. Stricker. and Thomas Barrows. July 1988.

Report 28. *Latent Structure Analysis of the Test of English as a Foreign Language* Robert F Boldt. November 1988.

Report 29. *Context Bias in the Test of English as a Foreign Language* William H Angoff January 1989

Report 30. *Accounting for Random Responding at the End of the Test in Assessing Speededness on the Test of English as a Foreign Language* Charles Secolsky January 1989

57906-00345 • S29P2 • 275632 • Printed in U S A