DOCUMENT RESUME

ED 386 481 TM 024 038

AUTHOR Schedl, Mary; And Others
TITLE An Investigation of Proposed Revisions to Section 3 of the TOEFL Test. TOEFL Research Report 47.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-94-42
PUB DATE Mar 95
NOTE 88p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS *Adult Students; *English (Second Language); *Language Tests; Psychometrics; *Reading Comprehension; Reading Tests; *Test Construction; Test Items; Test Length; Test Reliability; Test Validity; Vocabulary Development
IDENTIFIERS Revision Processes; *Speededness (Tests); *Test of English as a Foreign Language

ABSTRACT

The Test of English as a Foreign Language (TOEFL) program is exploring a change in Section 3 of the TOEFL test that would replace the vocabulary subpart with additional reading comprehension questions. This study investigated the proposed revision in terms of the length and timing that would be necessary to address concerns of test speededness of the section. Several psychometric issues relating to the proposed revision were also investigated. Responses of 33 students at each of 42 institutions teaching English as a second language were used. Results support the implementation of a revised TOEFL Section 3 consisting of five reading passages with a total of 50 items. Results also suggest that a total testing time of no less than 55 minutes should be allowed for the revised TOEFL Section 3. Additional psychometric analyses indicate that the current TOEFL score scale can be maintained with the revised Section 3 and that the proposed revisions will not appreciably affect the reliability and validity of Section 3. Appendices include: Preliminary Dimensionality Analyses of Section 3 of the TOEFL test, supplemental information, and an experimental TOEFL section. Fifteen figures and 10 tables present analysis results. (Contains 59 references.) (SLD)

TOEFL

TEST OF ENGLISH AS A FOREIGN LANGUAGE

Research Reports

REPORT 47
MARCH 1995

An Investigation of Proposed Revisions to Section 3 of the TOEFL Test

Mary Schedl
Neal Thomas
Walter Way

ETS
Educational Testing Service

An Investigation of Proposed Revisions to Section 3 of the TOEFL Test

Mary A. Schedl
Neal Thomas
Walter D. Way

Educational Testing Service
Princeton, New Jersey

RR-94-42

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

Abstract

The TOEFL testing program is currently exploring a change in Section 3 of the TOEFL® test that would replace the vocabulary subpart with additional reading comprehension questions. This change has been proposed by internal test development specialists and is supported by external experts in the field of English as a second language. The purpose of this study was to investigate the proposed revision to Section 3 in terms of the length and timing that would be necessary to address concerns of test speededness of the section. The study was carried out using an experimental design with test length and testing time defined as independent variables, and examinee test performance defined as the dependent variable. In addition, several psychometric issues relating to the proposed revision to Section 3 were investigated as part of the study.

The results of the study supported the implementation of a revised TOEFL Section 3 consisting of five reading passages with a total of 50 items. The results of the study also suggested that a total testing time of no less than 55 minutes should be allowed for the revised TOEFL Section 3. Additional psychometric analyses indicated that the current TOEFL score scale can be maintained with the revised Section 3, and that the proposed revisions will not appreciably affect the reliability and validity of Section 3 of the TOEFL test.

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 express their gratitude to the following current and former ETS staff members who contributed to the study:

Pat Carey for coordinating statistical analysis and data management of the study, and Dan Brown for helping with the statistical analysis.

Stella Cowell, Lois Hankinson, Ellen Kohler, and Barbara Suomi for helping with the planning and recruiting of institutions that participated in the study.

Dolores Carrano and Mike Heisler for helping with the requisitioning of materials for the study. Mike Heisler also put together the scanning procedures.

Nicholas Longford for providing the VARCL program used in the analysis.

Isaac Bejar, Susan Chyn, and Kentaro Yamamoto for helpful reviews of earlier drafts of the report.

Cindy McMahon for secretarial support.

Table of Contents

List of Appendices

List of Figures

List of Figures (continued)

List of Tables

Introduction

The current Section 3 of the TOEFL test consists of 30 discrete vocabulary items and 30 comprehension items of various types linked to passages of 200-350 words in length (five passages with five to seven questions each). Examinees have 45 minutes to complete both parts. The two parts correlate highly, and it is assumed that both contribute to the measurement of reading comprehension. Whereas the comprehension items linked to longer contexts enjoy a high degree of face validity, the vocabulary subpart has been criticized by linguists, English as a Second Language (ESL) professionals, and others, including the TOEFL Committee of Examiners, for testing knowledge of individual vocabulary words and phrases in isolated sentences such as the one that follows.

Example 1:

> The world would look eerily different if human eyes were sensitive to infrared radiation.
>
> (A) weirdly
> (B) increasingly
> (C) surprisingly
> (D) superficially

This kind of discrete point testing based on an understanding of language proficiency as a set of linguistic abilities which could be separately measured (phonological, syntactical, lexical) was common at the time the TOEFL test was developed in 1963. Reading theory has been evolving, however, since 1967 when Goodman introduced the idea of reading as a "psycholinguistic guessing game" (Goodman, 1967). Reading is no longer viewed as a passive activity based on the bottom-up processing of separate linguistic units but rather is seen as a dynamic, interactive process involving both bottom-up and top-down processing (Carrell, 1984; Orasanu and Penney, 1986; Silberstein, 1987; Carrell, 1987; Grabe, 1988; Eskey and Grabe, 1988; Grabe, 1991). Communicative approaches to teaching and testing have evolved, based on the idea that various competencies interact in the comprehension process (Hymes, 1972; Canale and Swain, 1980; Savignon, 1983; Duran et.al., 1985; Carrell, 1987 and 1989; Bachman, 1990). Communicative tests do not focus on measuring discrete aspects of language performance since, in authentic language use, grammatical, phonological, and lexical knowledge do not manifest themselves independently. In real-world communicative situations context gives clues to meaning, and a longer context naturally gives more clues. Test items that require more context to answer are, therefore, likely to be more reflective of communicative aspects of language behavior (Swain, 1983). A number of studies have supported the importance of context in readers' ability to determine the meaning of unknown words (Drum and Konopak, 1987; Sternberg, 1987; Barnett, 1988), although the number and variety of studies in both first- and second-language research, and their sometimes conflicting results, indicate that the role context plays in the learning of word meaning is far from clear (Sternberg and

Powell, 1983; Williams and Dallas, 1983; Bensoussan and Laufer, 1984; Nagy, et al., 1985 and 1987; Schatz and Baldwin, 1986; Weiss, et al., 1986; Nagy and Gentner, 1987; Sternberg, 1987; Dollerup, et al., 1988; Laufer, 1988). While there is no evidence that communicative tests measure language proficiency better than discrete point tests, they have greater face validity and are assumed to have a better washback effect on those preparing for language tests.

In 1989 a pretest study was developed in the Test Development group of the TOEFL program to compare examinee performance on vocabulary items in the single-sentence format and performance on the same vocabulary items tested together with other reading comprehension items in extended passages (Schedl and Way, 1990). The following is an example of the extended-passage context (compare to Example 1 from the previous page). In this longer passage context there are a number of syntactical and lexical clues to the meaning of the same word tested in Example 1 from the previous page.

Example 2:

> Human vision, like that of other primates, has evolved in an arboreal environment. In the dense, complex world of a tropical forest, it is more important to see well than to develop an acute sense of smell. In the course of evolution, members of the primate line have acquired large eyes while the snout has shrunk to give the eye an unimpeded view. Of mammals, only humans and some primates enjoy color vision. The red flag is black to the bull. Horses live in a monochrome world. Light visible to human eyes, however, occupies only a very narrow band in the whole electromagnetic spectrum. Ultraviolet rays are invisible to humans, though ants and honeybees are sensitive to them. Humans have no direct perception of infrared rays, unlike the rattlesnake, which has receptors tuned in to wavelengths longer than 0.7 micron. The world would look **eerily** different if human eyes were sensitive to infrared radiation. Then, instead of the darkness of night, we would be able to move easily in a strange, shadowless world where objects glowed with varying degrees of intensity. But human eyes excel in other ways. They are, in fact, remarkably discerning in color gradation. The color sensitivity of normal human vision is rarely surpassed even by sophisticated technical devices.
>
> The word "eerily" in line 11 [line 13 in the above typeset] is closest in meaning to
>
> (A) weirdly
> (B) increasingly
> (C) surprisingly
> (D) superficially

This new vocabulary format was perceived to have advantages over the single-sentence format in that it (1) had improved face validity because of the extended-passage context; (2) allowed examinees to make use of context clues of various kinds; (3) tested reading comprehension as well as vocabulary recognition; and (4) was likely to provide

beneficial feedback to examinees, who would probably spend more time reading and less time memorizing word lists in preparation for the test. Results indicated that the items embedded in extended-passage contexts behaved well from a psychometric point of view.

In 1990 the TOEFL Committee of Examiners and the TOEFL program, through the TOEFL 2000 project, began exploring ways in which changes could be incorporated into TOEFL to make the test better reflect authentic language use. At its April 1990 meeting, the Committee of Examiners expressed a strong interest in eliminating Vocabulary as a separate part of Section 3.

At the committee's request, a 54-item, all-passage pilot test incorporating vocabulary items into six reading comprehension sets was assembled and pretested at a number of English Language Institutes (ELIs) in the United States. Examinees' scores on the pilot test were correlated with their scores on an Institutional TOEFL test[1] consisting of the usual 30 discrete vocabulary and 30 reading comprehension passage items in five passage sets. Statistical analyses of the data indicated that the new format was reliable and that items fell within the current range of difficulty for TOEFL.

However, the 54-item six-passage pilot test appeared to be significantly speeded in comparison to the TOEFL form that was taken by the same examinees; i.e., it took the institutional population longer to answer 54 items that were all associated with reading passages than it took them to answer 30 vocabulary questions in the current TOEFL single-sentence format and 30 reading comprehension questions associated with five passages in linked sets. It was thought the experimental format might not be as speeded for the TOEFL population as it was for the institutional population, since the institutional population is typically less proficient in English than the TOEFL population[2]. Another factor that may have influenced calculations of speededness based on the pilot was that participants were specifically instructed not to guess and to only answer questions for which they had time. On the TOEFL test there is no penalty for guessing, and it is assumed that examinees randomly fill in unanswered questions at the end of the test.

Purpose of the Study

Incorporating vocabulary items into reading comprehension passage sets was an attractive idea for many reasons. The new item type was not only attractive in terms of face validity, construct validity, and positive feedback to ESL students, but also performed well in terms of traditional item statistics (i.e., difficulty and item-test biserial correlations).

---

[1] Institutional TOEFL tests are previously used nondisclosed TOEFL test forms that are administered by institutions to their students using their own facilities.

[2] The mean institutional score for examinees taking the pilot test was 476.49. Usual TOEFL means are between 505 and 520.

The major issue that still needed to be addressed was speededness. Although it was anticipated that the TOEFL Program would be able to accommodate some increase in testing time, it had to be determined whether enough items could be administered to maintain acceptable standards of reliability for Section 3.

The primary goal of this study was the direct assessment of the speededness of the proposed new Section 3. This assessment was carried out by administering a prototype form of the revised Section 3 test under controlled experimental conditions, where the manipulated variables were the testing time and the length of the test. In addition to the more critical issue of speededness, however, was a concern with how the proposed revisions would affect the dimensionality of Section 3. The dimensionality question was viewed as a related research issue, and was addressed through a series of statistical analyses that were applied to sample data from the current Section 3 of the TOEFL test (rather than the data collected in this investigation). The methods and results of the preliminary dimensionality analyses are provided as Appendix A of this report. The methods and results of the principal analyses that addressed the effects of varied testing time and test length on the revised Section 3 of the TOEFL test are provided in the sections that follow.

Methodology

Experimental Design

The design of the study called for randomly assigning three different testing time limits at participating English as a Second Language (ESL) institutions, and randomly assigning experimental Section 3 test forms of three different lengths at each participating institution. This approach has been used previously by several researchers, (e.g., Wild and Durso 1979; Evans 1980), and provides a more direct exploration of the assumptions typically employed about examinee behavior in less controlled studies of speededness, such as those in Bejar (1985), Wilson (1989), and Secolsky (1989). The three experimental test forms used for the study each included regular TOEFL Sections 1 and 2, and a Section 3 that consisted of the same six reading passages and either 48, 54, or 60 items. Each of the six sets consisted of nine to 11 items. Appendix B contains the longest version of the three experimental Section 3 tests used in the study and identifies the items that were excluded in the shorter versions. The three forms shared a common core of 48 reading items, and the forms with increased numbers of items were obtained by adding one or two additional items per passage. Analyses prior to data collection had indicated that the experimental design would require the participation of a reasonably large number of institutions because it would not be possible to randomly assign different time limits within institutions. It was estimated that the participation of 45 institutions with 33 students per institution would provide satisfactory precision for the purposes of the experiment. Each institution was to be assigned one of three timing conditions for administering the experimental reading section: 50 minutes, 55 minutes, or 60 minutes. The assignment of Section 3 test length was randomized within-institution by spiralling the three test forms sent to institutions.

Participant Recruitment

Subjects were recruited by offering free Institutional TOEFL tests in exchange for participation in the study. (The Institutional TOEFL test data were utilized as covariate information in the analysis of the data.) Participating English as a second language institutions were originally contacted by phone, after which a formal letter was mailed containing a detailed description of the project and initial instructions. All experimental materials were packaged and sent out with Institutional TOEFL test materials, along with more specific instructions for administering the experimental test forms. Appendix C contains samples of the original letter sent to the participating institutions and the instructions sent with the testing materials. Institutions were instructed to administer the experimental test forms approximately one week before the Institutional TOEFL test administration. In some cases, institutions were unable to comply with the instructions. For example, one institution administered the experimental test form after the Institutional TOEFL, and a second institution administered the experimental test approximately three weeks before the Institutional TOEFL. These instances of noncompliance were not considered serious enough to invalidate the resulting data.

Of the 45 institutions originally recruited for the study, two institutions never returned materials, and a third that administered the experimental reading test under improper timing conditions was not included in the analysis of the data.

Matching the Experimental and Institutional TOEFL Data

Because the experimental and institutional TOEFL data were processed independently, it was necessary to match the records of those participants who took both tests. All nonmatching records were deleted from further analyses. Several institutions administered the experimental test using the Institutional TOEFL answer sheets rather than the answer sheets provided with the experimental test. Some of the participants at these institutions did not provide sufficient information on the answer sheets to determine which of the three experimental forms had been administered. These records were also deleted from further analyses.

Identification of Common Items

Because the three experimental reading tests were of different lengths, it was necessary to find a common metric for comparing examinee performance. It was not possible to use typical equating procedures for this purpose, because such adjustments would remove differences that might be due to the differing experimental conditions. Therefore, it was decided that performance comparisons would be based on the 48 items that were common to the three forms. An item analysis indicated that one of the common 48 items was statistically flawed, however, so this item was deleted. The resulting comparisons among forms were based on the remaining 47 items common to the three experimental forms. The statistical properties of the items removed did not appear to differ from those of the remaining items in any meaningful way.

Analysis of the Data

The analyses of the data included six specific analyses: 1) an assessment of the sampling procedures; 2) traditional speededness analyses; 3) outlier analyses; 4) score comparisons; 5) equating analyses; and 6) alternate form reliability analyses. The methods used in these analyses are described in the paragraphs below.

Assessment of sampling procedures. The performance of the examinees on the Institutional TOEFL test should not be affected by the conditions assigned to the students during the experimental exam. We investigated students' scores on the three sections of the Institutional TOEFL to check for possible corruption of the assignment of the experimental conditions and to measure the amount of chance imbalance in ability among the examinees assigned the different experimental conditions. We fit models of the form

$$Y_{ijkl} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \delta_i + \epsilon_{ijkl} \qquad (1)$$

where Y_{ijkl} represents the number correct score from a section of the Institutional

TOEFL test, μ is the mean score of students assigned the fewest items and shortest testing time on the experimental test, α_j, $j = 2,3$ and β_k, $k = 2,3$ are the additive effects of the increased number of items and test-taking time assigned for the experimental exam respectively, $(\alpha\beta)_{jk}$ are the potential interactions of these effects, δ-$N(0,\tau^2)$ are random effects used to model the clustering of examinees with similar ability within institutions, and ϵ_{ijkl}-$N(0,\sigma_2)$ are the differences in the institutional scores among examinees assigned the same experimental conditions at institution *i*. The δ_i are needed in the model to obtain standard errors that account for the random assignment of timing conditions to institutions instead of individual examinees. We computed the likelihood ratio test of the hypothesis $\alpha_j = \beta_k = (\alpha\beta)_{jk}$ for all values of j and k to measure the effectiveness of the random assignment of examinees and institutions. This null hypothesis is true if the randomization is properly executed.

Traditional speededness analyses. To provide a traditional assessment of speededness for the experimental reading test, several common speededness indices were compiled. These included the percentage completing the test, the percentage completing 75 percent of the test, and a ratio index of speededness described by Gullicksen (1987).

Outlier analyses. Plots and other data-cleaning activities revealed a small number of outliers among the experimental Section 3 scores. By examining item-level information, we determined that most of these outliers were examinees who stopped or switched to rapid guessing when they reached Section 3 of the experimental test. We regressed the number-correct score of the 47 common items of the experimental Section 3 on the number-correct score of the Institutional TOEFL Section 3, and removed examinees whose standardized residuals exceeded 2.5 in absolute value to produce a systematic approach for removing examinees with inconsistent scores from the primary reported results involving the experimental Section 3.

Score Comparisons. The primary comparisons of the performance of the examinees under the differing experimental conditions were made using the number-correct score on the common 47 items based on a model similar to (1),

$$Y_{ijkl} = \mu + \alpha_j + \beta_k + \gamma X_{ijkl} + \delta_i + \epsilon_{ijkl}\ , \qquad (2)$$

where Y_{ijkl} now represents the number-correct score on the common 47 items of the experimental Section 3, X_{ijkl} is the number-correct score of the Institutional TOEFL Section 3, γ is a regression parameter for X_{ijkl}, and the other parameters serve purposes analogous to model (1). The Institutional TOEFL Section 3 scores are included for covariate adjustment because they have high correlation with the experimental scores, thus sharply improving the precision of the comparisons. The scores on the other sections of the Institutional TOEFL test do not have much additional predictive value, so they were not included in (2). Likelihood ratio tests of the hypothesis that interactions exist between the testing time and number of items were consistent with no interaction or small interactions, so the interaction terms were not included in the reported results based on model (2).

A graphical summary of the effect of increased testing time was also produced by interpolating orthogonal polynomials between the experimentally assigned testing times. A reparameterization of the model in (2) with t_k denoting testing time, and $l(t_k)$ and $q(t_k)$ representing orthogonal linear and quadratic polynomials (3 points), yields the alternative form of (2),

$$Y_{ijkl} = \mu + \alpha_j + \lambda_1 l(t_k) + \lambda_2 q(t_k) + \gamma X_{ijkl} + \delta_i + \epsilon_{ijkl} . \quad (3)$$

A graph of the effect of increased testing time on the number-correct score on the common 47 items of the experimental Section 3 was formed by interpolating values of $l(t_k)$ and $q(t_k)$ for testing time in the range from 50 to 60 minutes with the test length set at 60 items, and the institutional test score set at the average value for the entire experimental sample.

Equating analyses. A subset of the experimental data was used to equate the scores on the experimental Section 3 to the existing TOEFL Section 3 scale. Specifically, the items in the first five passages (49 items) of the experimental Section 3 test were calibrated using LOGIST 6 (Wingersky, Patrick, and Lord, 1988). Note that all 49 items were taken only by participants who were administered the 60-item test. To calibrate the data for these items, a reformatted data matrix was created that consisted of complete data for 39 of the 49 items, data for an additional five items that were taken by approximately two-thirds of the participants, and data for five more items that were taken by only one-third of the participants. The resulting item-parameter estimates for these 49 items were then transformed to the TOEFL IRT scale using IRT item characteristic curve (ICC) scaling procedures (Stocking and Lord, 1983). It was possible to carry out the ICC scaling because 30 of the 49 items used in this portion of the study had been previously administered in operational TOEFL forms, and the previously obtained IRT statistics could be used as a basis for the transformation of the new estimates.[3] Finally, IRT true-score equating procedures were used to equate the 49-item test to the TOEFL base form. Specifically, the transformed IRT item-parameter estimates for the 49 items and the IRT item-parameter estimates obtained for the 58-item TOEFL base form were used as the basis for the true-score equating. The resulting conversions were then applied to obtain TOEFL converted scores on the experimental Section 3 for all participants who were administered the longest of the three experimental forms (as only these participants had taken all 49 of the items used in the true score equating).

Alternate forms reliability analyses. Based on the results of the equating analyses, two sets of section and total converted scores were available for the participants who were administered the longest of the three experimental Section 3 forms: one set based on the Institutional TOEFL form and a second set based on the experimental form. For Sections 1 and 2, the two sets of scores were based on parallel forms previously used as operational TOEFL tests. For Section 3, one score was based on the Institutional

---

[3] The procedures in this portion of the study were as similar to procedures carried out in operational TOEFL equatings as possible.

TOEFL form and the other score was based on the 49-item experimental TOEFL form used in the equating analyses. The means and standard deviations of the section and total converted scores were compared, and the bivariate relationships between scores were examined graphically. Pearson product moment correlations were calculated to assess the alternate-forms reliabilities of the section and total converted scores.

Results

Assessment of Sampling Procedures

Table 1 summarizes the number-correct and converted-score means and standard deviations on the TOEFL Institutional form used for the study. This form was originally administered as an operational TOEFL test form in December of 1990 and is designated as Form 3MTF12. Included in Table 1 are summary statistics for the foreign and domestic samples taking the TOEFL test in December of 1990 as well as the summary statistics for the Institutional TOEFL candidates participating in the study. The Institutional TOEFL statistics are further broken down by assignment to the three different experimental test-form conditions. Examination of Table 1 shows that examinees assigned to the longer experimental test forms were higher in ability, and examinees assigned to the longer testing times were also higher in ability. The likelihood ratio test that all of the parameters in model (1) except μ, τ^2 and σ^2 are zero for the Institutional Section 3 yields a $\chi^2_8 = 9.96$ with p-value 0.27. Similar p-values were obtained for the other sections of the Institutional TOEFL, and for the first two (common) sections of the experimental exam. Similar p-values were also obtained when (1) was restricted to contain only additive terms. Although it would have been realistic to hope for a better randomization, the imbalances among the experimental groups from repeated randomizations would often be worse than occurred in our experiment.

As expected, the participants in the Section 3 study were less proficient than candidates taking the same form in an operational administration. The mean total score on Form 3MTF12 was 479.5 for the research participants, compared with mean total scores of 514.0 (Foreign) and 508.3 (Domestic) in the operational samples. The most noticeable performance differences between the research and operational samples were on Sections 2 and 3. The TOEFL Institutional sample performed better on Section 1 (Listening Comprehension). This is consistent with the generally observed trend on the TOEFL test that domestic candidates perform better than foreign candidates on Section 1.

Traditional speededness analyses

Table 2 displays traditional speededness statistics for the experimental Section 3 by test length and test-timing conditions. Also included in Table 2 are KR-20 reliability estimates. The speededness statistics suggest a visible and expected trend related to testing time: Within each test-length condition, candidates reached more items when more testing time was provided. For example, for the test length of 60 items, 95.3 percent of the participants completed the test and 100 percent reached three-fourths of the test when 60 minutes were provided. When 55 minutes were allowed, only 81.7 percent completed the test, and 99.2 percent reached three-fourths of the test. When 50 minutes were allowed, only 77.9 percent finished the test, and only 93.6 percent finished three-fourths of the items administered. This trend is also apparent for the test length of 48 items, and to a lesser extent, for the test length of 54 items.

A second trend that might have been expected in the data -- that within a given amount of testing time, a greater percentage of candidates would complete the shorter tests than would complete the longer tests -- was not apparent in the data. For example, when 60 minutes were allowed, a smaller percentage of candidates in the 54-item condition completed 100 percent and 75 percent of the items than in either the 60-item or 48-item conditions. One possible explanation for this finding is that the effects of additional items per passage on speededness are relatively minor compared with the effects of additional passages. That is, regardless of how many items were administered, candidates still had to read all six passages in the test. Overall, the evidence from the traditional speededness statistics suggested that six passages were definitely too many when 55 or 50 minutes were allowed, and possibly too many when 60 minutes were allowed.

The KR-20 reliabilities ranged from 0.87 to 0.92 across the experimental conditions, which compare favorably with the KR-20 value of 0.90 for the Institutional TOEFL. The magnitudes of the KR-20 values did not seem to be related to test length, probably because of the effects of speededness in several of the conditions. Because estimates of reliability are inflated when tests are speeded, the KR-20 values in Table 2 should be interpreted with caution.

Outlier Analysis

Figure 1 displays a plot of the number-correct score on the common 47 items of the experimental Section 3 against the number-correct score on the Institutional Section 3, with the 31 outliers identified by the regression residual approach highlighted as asterisks (three of the outliers were hidden). The average scores for the lower-outlier examinees are equal to or slightly above the mean for all experimental subjects on each section of the Institutional TOEFL exam and the first two sections of the experimental exam. Table 3 displays the average number correct for the common items on the experimental exam by reading passage for the low-outlier examinees and the entire experimental sample. The performance of low-outlier examinees was consistent with their other test scores on the items associated with the first reading passage, but their performance dropped to below guessing on the remaining passages (two out of eight expected). The identification numbers of the examinees who appear to have lost motivation are strongly clustered, suggesting that when one examinee chose not to participate, examinees nearby were often affected. The performance of all but this small subset of examinees on the experimental Section 3 is consistent with the performance of the same examinees on the Institutional TOEFL test, where motivation due to grading and placement is often present, and also with the performance of the same examinees on the first two sections of the experimental exam. This provides good evidence that most students performed close to their maximum ability on the experimental Section 3.

The outliers were not concentrated in one or two experimental conditions. As a result, the substantive conclusions from the comparisons of the different experimental conditions did not change when the outliers were included in the analyses. Model (2) without the interaction terms fits better when the outliers are excluded.

Score Comparisons

The most important results of the experiment are summarized in Table 4. The performance of the examinees improved substantially as the amount of testing time was increased for each test length. The examinees also appear to have performed better on the common items of the longer test forms, contradicting prior expectations.

The trends in Table 4 are very similar to those of the Institutional TOEFL scores, which are highly correlated with the experimental test. Covariate adjustment accounting for the imbalances in ability that are not the result of the differing testing conditions almost entirely accounts for the differences in the performance of examinees assigned tests of different lengths. The covariate adjustment also reduces the strong trend with increased testing time, but the improvement with additional testing time remains strong even after covariate adjustment. The estimates from the covariate adjustment model in (2) were computed using the program VARCL (Longford, 1987) and are summarized in Table 5.

The large values of β_2 and β_3 compared with their standard errors in Table 5 indicate that the time trend is replicable, and the small values of α_2 and α_3 indicate that the apparent improved performance on the common items for the longer test forms is likely to be a consequence of the chance assignment of stronger examinees to these test forms. The larger standard errors for β_2 and β_3 are the result of the assignment of the timing conditions at the institutional level. The clustering of examinees of similar abilities within institutions, as measured by $\tau/\sigma = 0.25$, is substantial, and in good agreement with the estimates of this quantity computed for the sample-size calculations using Institutional TOEFL data reported to the TOEFL Program.

Figure 2 displays the predicted increase in the mean number-correct score on the 47 common items as test time increases based on the model in (3). Our prediction is that performance will improve substantially when test-taking time is increased from 50 to 55 minutes, and will level out between 55 and 60 minutes. The quadratic curve fit to the data actually decreases for the largest time values, but this decrease is small and not statistically significant. We believe that the curve should actually approach an upper asymptote.

Score Comparisons Based on Passages 1 to 5

Because the traditional speededness analyses suggested that performance on the last passage may have been affected in all timing conditions, the primary score comparisons were repeated using data from only the 39 common items in the first five passages. The results of these comparisons were virtually identical to the results obtained

using the 47 common items from six passages. Figure 3 presents a graphical summary of these results, which are very similar to the results presented in Figure 2.

Equating Analyses

In order to perform IRT true-score equating on the experimental data, it was necessary to calibrate the Section 3 data from all experimental conditions combined using LOGIST.[4] Based on the results of both the traditional speededness analyses and the score comparisons, it was clear that calibrations of the last set of items in the experimental Section 3 would be problematic because of speededness effects. For this reason, the IRT calibrations were limited to the first five sets in the experimental test forms, a total of 49 items. The parameters for running LOGIST were identical to those used in operational administrations of the TOEFL test, except the value of CRITFIXC, which was set at -2.5 because of the smaller sample sizes. (CRITFIXC affects the number of items for which LOGIST will not estimate a unique c-value. See Wingersky, Patrick, and Lord, 1988, for further information.) The LOGIST run converged in 22 stages, with 21 of the 49 items estimated with a common c-value of 0.17405.

As previously mentioned, the item-parameter estimates for the 49 experimental test items were then transformed to the TOEFL IRT scale using the ICC scaling procedures, using item parameters for 30 of the 49 items that had been previously estimated when pretested in regular TOEFL test administrations. IRT true-score equating procedures were then used to equate the 49 experimental items to the current TOEFL base form (Form 3KTF05), resulting in a Section 3 converted score for each of the 50 possible number-right scores. Figure 4 displays the resulting number-correct converted-score equating relationship for the experimental form, as well as the corresponding number-correct to converted-score line for the TOEFL base form, which contained 58 items. The differences in the equating lines reflect the equating of the shorter experimental Section 3 form to the longer TOEFL base form.

Alternate-form reliability analyses

Table 6 contains the alternate-form summary statistics for the Institutional TOEFL and experimental test results for the candidates taking the longest of the three experimental forms (Form 1C). Both number-correct and converted-score results are presented, although comparisons of alternate-form number-correct score means and standard deviations are not meaningful. The converted score summary statistics for the Institutional TOEFL and experimental TOEFL forms are quite similar. For Sections 1 and 2, the mean converted scores for the Institutional TOEFL forms are slightly higher, which could be due to practice effects that occurred between the administration of the experimental form and the Institutional TOEFL form, typically a period of less than two weeks. For Section 3, there is no evidence of practice effects, but it should be noted that

[4] Note that raw-to-converted score conversions already existed from operational TOEFL administrations for Sections 1 and 2 of the experimental forms.

any practice effects for the experimental Section 3 would have been equated away. The total-score summary statistics at the bottom of Table 6 suggest that the overall results for the Institutional and experimental TOEFL forms were very similar.

The last column of Table 6 provides the alternate-form reliability coefficients. For both number-correct and converted scores, the alternate-form reliabilities for Section 1 were noticeably higher than those for the other two sections. This may be due to several factors, including test length and test structure. For example, Section 1 has 12 more items than Section 2 and the structural differences between the Institutional and experimental Section 3 forms would have adversely affected the correlations. For all three sections, the alternate-form reliabilities are lower than those reported for typical TOEFL test administrations. This would be expected, since the alternate-form reliabilities incorporate sources of measurement error in addition to those present in the administration of a single form.

Additional information about the relationships between the Institutional TOEFL and experimental-form scores are provided in Figures 5 to 8, which display bivariate plots of the number-correct and converted scores for each section as well as the total converted scores. These figures depict strong linear relationships, although for Sections 2 and 3, a number of outliers appear to exist. To adjust the summary score statistics for outliers, the Institutional TOEFL scores for each section and the total scores were regressed on the corresponding scores from the experimental form, outliers were identified on the basis of absolute standardized residuals in excess of 2.5, and alternate form summary statistics were recomputed. Table 7 presents these results, which are similar to those in Table 6. However, it can be seen that the alternate form reliabilities for Sections 2 and 3 are much closer to those for Section 1 when the effects of outliers are removed from all three sections.

In summary, the alternate form statistics suggested that a revised Section 3 form based on five passages and 49 scored items will result in sufficiently reliable number-correct and converted scores. Although the length and composition of the experimental Section 3 differed from the TOEFL Institutional Section 3, the relationship between converted scores on the two forms was similar to the relationships seen for Sections 1 and 2. It is recognized that alternate forms of the revised Section 3 might have slightly different relationships, but it is doubtful that such relationships would be meaningfully different from the relationship seen between the experimental Section 3 and the Institutional TOEFL Section 3 in this study.

Discussion and Recommendations

The primary purpose of this study was to perform a direct assessment of the speededness of the proposed new Section 3 of the TOEFL test. The essential results of the study are summarized in Figure 2. This plot suggests that allowing 55 minutes to examinees for completing the revised Section 3 with six passages would be minimally adequate, regardless of whether 48, 54, or 60 items were included in the test. However, the statistics in Table 2 indicate that the tests with six passages border on what traditionally would be considered speeded when 55 minutes are allowed, and in the case of the 54-item test, when 60 minutes are allowed. This implies that a testing time greater than 60 minutes would be required in order to be confident that an administration of this specific test form would not be speeded in the experimental population. However, a testing time greater than 60 minutes might not be feasible from an operational perspective.

The results of the study did indicate that little additional time is needed to accommodate the addition of one or two extra items within each passage. This in part explains why performance was not related to test length, as all the test forms used in this study consisted of six passages. We believe that a five-passage test would require substantially less time than a six-passage test because of the reduced reading load. The five-passage recommendation is bolstered by the satisfactory results of the score equating and alternate-forms reliability analyses performed on the group taking the 60-item test, which were based on the 49 items in the first five passages only. Taken as a whole, these results support a revised Section 3 with five passages and an average of 10 items in each passage, for a total of 50 items; this is also feasible from a test development perspective.

Although all our experimental settings involved six reading passages, there were several results of the study from which timing recommendations for a five-passage Section 3 can be drawn. As previously mentioned, the traditional speededness statistics shown in Table 2 imply that more than 60 minutes would have to be allowed to be confident that an administration of this specific test form would not be speeded in the experimental population. Corroborating evidence is presented in Table 3, which indicates that the number-correct scores on the common items dropped off substantially on the last passage compared with the previous passages. (Note that Passage 5 had seven scored common items, whereas all other passages had eight scored common items.) Furthermore, the item analysis statistics indicated that the three poorest performing items in the test were all in the last passage. (One of these items was common to all versions of the test and, as previously noted, was deleted from scoring.) Although other interpretations are possible, we believe that these findings support the contention that an appropriate time limit for the six-passage test used in this experiment would be greater than 60 minutes. If one were to set 65 minutes as a minimally acceptable limit for this test, the proportionally equivalent limit for five passages would be about 55 minutes.

We also replicated the analyses carried out on the six-passage test using the 39 common items in the first five passages only. The results of these analyses indicated trends that were virtually identical to the trends obtained using the 47 common items

from six passages (see Figures 2 and 3). That is, the prediction based on analyses of the five-passage data is also that performance will improve with an increase in testing time from 50 to 55 minutes, and will level out between 55 and 60 minutes. Because the existence of the sixth passage could affect the performance of candidates on the first five passages, these results should be interpreted with caution.

Our analyses of the experimental data collected in this study suggest that a minimum of 55 minutes should be allowed for a five-passage, revised Section 3. This represents a 10-minute increase to the timing of the current Section 3. A followup study involving only five reading passages would be desirable. However, because the standard errors from a new study will be substantial, and because of the uncertainty created by extrapolating from a low-stakes test given to an institutional population to an operational TOEFL examination, judgments based on information external to a new experiment will still be required to set appropriate time limits.

A conservative view of the results would perhaps add even more time to this minimum. The conservative view has merit for several reasons. First, the dimensionality analyses of the current Section 3, reported in Appendix A, indicate that there is an end-of-test dimension that can be interpreted as being related to speededness. This phenomenon was also reported in a study by Oltman, Stricker, and Barrows (1988). The effect is relatively subtle, and may in fact be confounded with the trait measured by Section 3 itself. For example, the increase in the number of random responses near the end of a test section may be caused by the fact that many examinees are pressed for time, or it could be that examinees who are less proficient simply stop paying attention to the questions that are presented near the end of the test. Nevertheless, providing additional time on the revised TOEFL might lessen the salience of this effect, which would prove beneficial in applications such as item calibration and scaling. A second reason for considering more extended time limits is that operational reading sets will vary from form to form, and some of these forms will pose greater reading requirements than others. A final reason for adopting longer time limits is the inherent uncertainty in extrapolating from an institutional population in a low-stakes setting to the highly motivated TOEFL test-taking population. It is unclear whether the test would be more or less speeded in operational use.

Although the results of this study related to the equating, scaling and test reliability of the experimental Section 3 were positive, these issues should continue to be of concern as part of the implementation planning for the revised Section 3. In particular, it will be useful to implement the revised Section 3 by administering it simultaneously with a form of the current Section 3 to obtain additional information related to equating and scaling.[5] This will provide an opportunity to replicate the findings of this study with the actual TOEFL testing population. Also, some thought

[5] Such an administration would likely spiral the old and new Section 3 forms at domestic test centers, as it would be impossible to administer the two versions of Section 3 within the same test center.

should be given to the estimation of reliability in the revised Section 3 of TOEFL, as the entire section will be passage-based and IRT-based estimates of reliability will be questionable because of the effects of within-passage structures (Sireci, Thissen, and Wainer, 1991). Consequently, it will be useful to gather alternate-form reliability data based on parallel forms of the revised Section 3 of the TOEFL test soon after the revisions to the test are implemented, as these estimates of reliability would not be contaminated by within-passage effects.

References

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

Bachman, L. F. (1990). Communicative language ability. In L. F. Bachman, Fundamental Considerations in Language Testing. Oxford: Oxford University Press.

Barnett, M. A. (1988). Reading through context: How real and perceived strategy use affects L2 comprehension. Modern Language Journal, 72 (ii), 150-159.

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

Bejar, I. (1985). Test Speededness Under Number-right Scoring: An Analysis of the Test of English as a Foreign Language (ETS Research Report No. 85-11). Princeton, New Jersey: Educational Testing Service.

Benoussan, M. and Laufer, B. (1984). Lexical guessing in context in EFL reading comprehension. Journal of Reading Research, 7, 15-32.

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

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

Canale, M. and Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. Applied Linguistics, 1 (1), 1-47.

Carrell, P. L. (1984). Fostering interactive second language reading. In S. J. Savignon and M. S. Berns (Eds.), Initiatives in Communicative Language Teaching II. Reading, Massachusetts: Addison-Wesley.

Carrell, P. L. (1989). Metacognitive awareness and second language reading. Modern Language Journal, 73, 121-134.

Carrell, P. L. (1987). A view of written text as communicative interaction: Implications for reading in a second language. In J. Devine, P. Carrell, D. E. Eskey (Eds.), Research in Reading in a Second Language. Washington, D.C., TESOL.

Devine, J. (1987). General language competence and adult second language reading. In J. Devine, P. L. Carrell, and D. E. Eskey (Eds.), Research in Reading in English as a Second Language. Washington, D.C.: TESOL.

Dollerup, C., Glahn, E., and Hansen, C. R. (1988). Vocabularies in the reading process. ERIC manuscript, Copenhagen, Denmark.

Drum, P. A. and Konopak, B. C. (1987). Learning word meanings from written context. In M. G. McKeown and M. E. Curtis (Eds.) The Nature of Vocabulary Acquisition. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

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

Duran, R. P., Canale, M., Penfield, J., Stansfield, C. W. and Liskin-Gasparro, J. E. (1985). TOEFL from a Communicative Viewpoint on Language Proficiency: A Working Paper (TOEFL Research Report No. 17). Princeton, New Jersey: Educational Testing Service.

Eskey, D. E. and Grabe, W. (1988). Interactive models for second language reading: Perspectives on instruction. In P. L. Carrell, J. Devine, D. E. Eskey (Eds.), Interactive Approaches to Second Language Reading. Cambridge University Press.

Evans, F. (1980). A Study of the Relationships Among Speed and Power Aptitude Test Scores and Ethnic Identity (ETS Research Report No. 80-22). Princeton, New Jersey: Educational Testing Service.

Fraser, C. (1983). NOHARM: An IBM PC computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory. Armidale, Australia: The University of New England, Center for Behavioral Studies.

Grabe, W. (1988). Reassessing the term "interactive." In P. L. Carrell, J. Devine, D.E. Eskey (Eds.), Interactive Approaches to Second Language Reading. Cambridge University Press.

Grabe, W. (1991). Current developments in second language reading research. TESOL Quarterly, 25 (3), 375-406.

Goodman, K. S. (1967). Reading: A psycholinguistic guessing game. Journal of the Reading Specialist, 6, 126-135.

Gullicksen, H. (1987). Theory of mental tests. Hillsdale, New Jersey: Erlbaum.
Hale, G. A., Stansfield, C. W., Rock, D. A., Hicks, M. M., Butler, F. A., and Oller, J. W. (1988). Multiple-choice cloze items and the Test of English as a Foreign Language (TOEFL Research Report No. 26). Princeton, New Jersey: Educational Testing Service.

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

Hymes, D. (1972). On communicative competence. In B. J. Pride and J. Holmes (Eds.), Sociolinguistics. Harmondsworth: Penguin.

Laufer, B. (1988). Ease and difficulty in vocabulary learning: some teaching implications. Paper presented at the annual meeting of the International Association of Teachers of English as a Foreign Language, April 11-14, Edinburgh, Scotland.

Longford, N. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. Biometrika, 74, 817-827.

Lord, F. M. (1980). Applications of Item Response Theory to Practical Testing Problems. Hillsdale, New Jersey: Erlbaum.

McDonald, R. P. (1967). Nonlinear factor analysis. Psychometric Monographs, No. 15.

McDonald, R. P. (1981). The dimensionality of tests and items. British Journal of Mathematical and Statistical Psychology, 34, 100-117.

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

McKinley, R. L. (1989). Confirmatory analysis of test structure using multidimensional item response theory (ETS Research Report 89-21). Princeton, New Jersey: Educational Testing Service.

McKinley, R. L., and Way, W. D. (1992). The feasibility of modeling secondary TOEFL ability dimensions using multidimensional IRT models (TOEFL Technical Report TR-5). Princeton, New Jersey: Educational Testing Service.

Nagy, W. E., Herman, P. A., and Anderson, R. L. (1985). Learning words from context. Reading Research Quarterly, 20 (2), 233-253.

Nagy, W. E., Anderson, R. L. and Herman, P. A. (1987). Learning word meanings from context during normal reading. American Educational Research Journal, 24 (2), 237-270.

Nagy, W. E. and Gentner, D. (1987). Semantic constraints on lexical categories (Technical Report No. 413). Cambridge, Massachusetts: Boldt, Beranek and Newman, Inc.

Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality. Journal of Educational Measurement, 28, 99-117.

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

Orasanu, J. and Penney, M. (1986). Comprehension theory and how it grew. In J. Orasanu (Ed.), Reading Comprehension from Research to Practice. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Savignon, S. (1983). Communicative Competence: Theory and Classroom Practice. Reading, Massachusetts: Addison-Wesley.

Schatz, E. K. and Baldwin, R. C. (1986). Context clues are unreliable predictors of word meanings. Reading Research Quarterly, 21 (4), 439-453.

Schedl, M. and Way, W. D. (1990). TOEFL vocabulary items: The impact of context. Presented at the Educational Testing Service DIF interest group seminar, June 12. Princeton, New Jersey: Educational Testing Service.

Secolsky, C. (1989). Accounting for Random Responding at the End of the Test in Assessing Speededness on the Test of English as a Foreign Language (TOEFL Research Report No. 30, ETS Research Report No. 89-11). Princeton, New Jersey: Educational Testing Service.

Silberstein, S. (1987). Let's take another look at reading: Twenty-five years of reading instruction. English Teaching Forum, 25, 28-35.

Sireci, S., Thissen, D., and Wainer, H. (1991). On the reliability of testlet-based tests. Journal of Educational Measurement, 3, 237-247.

Sternberg, R. J. (1987). Most vocabulary is learned from context. In M. G. McKeown and M. E. Curtis (Eds.), The Nature of Vocabulary Acquisition. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Sternberg, R. J. and Powell, J. S. (1983). Comprehending verbal comprehension. American Psychologist, August, 878-893.

Stocking, M. L., and Lord, F. M. (1983). Developing a common metric in item response theory. Applied Psychological Measurement, 7, 201-210.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. Psychometrika, 52, 589-617.

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensional assessment and ability estimation. Psychometrika, 55, 293-325.

Swain, M. (1984). Large-scale communicative language testing: A case study. In S. J. Savignon and M. S. Berns (Eds.) Initiatives in Communicative Language Testing II. Reading, Massachusetts: Addison-Wesley.

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

Walker, R. C. (1981). A reader's guide to test analysis reports. Princeton, New Jersey: Educational Testing Service.

Weiss, A. S., Mangrum, C. T., Llabre, M. M. (1986). Differential effects of differing vocabulary presentations. Reading Research and Instruction, 25 (4), 265-276.

Wild, C., and Durso, R. (1979). Effect of increased test-taking time on test scores by ethnic group, age, and sex (GRE Board Professional Report No. 84-09aP, ETS Research Report No. 89-36). Princeton, New Jersey: Educational Testing Service.

Williams, R. and Dallas, D. (1984). Aspects of vocabulary in the readability of content area L2 educational textbooks: A case study. In J. C. Alderson and A. H. Urguhart (Eds.), Reading in a Foreign Language. Longman.

Wilson, K. (1989). Population Differences in Speed Versus Level of GRE Reading Comprehension: An Exploratory Study (GRE Board Professional Report No. 84-09aP, ETS Research Report No. 89-36). Princeton, New Jersey: Educational Testing Service.

Wingersky, M. S., Patrick, R., and Lord, F. M. (1988). LOGIST User's Guide (Version 6.00). Princeton, New Jersey: Educational Testing Service.

Appendix A

Preliminary Dimensionality Analyses of Section 3
of the TOEFL Test

Preliminary Dimensionality Analyses

Background

To provide background information related to the proposed changes to Section 3 of the TOEFL test, dimensionality analyses of two Section 3 tests were carried out. Although a number of investigations related to the dimensionality of the TOEFL test have been undertaken (c.f., McKinley and Way, 1992; Hale, Rock and Jirele, 1989; Oltman, Stricker, and Barrows, 1988; Hale, Stansfield, Rock, Butler and Oller, 1988; Dunbar, 1982), all of these studies have examined the TOEFL test as a whole rather than Section 3 in isolation. In practice, however, each section of the TOEFL test maintains a distinct reported score scale.

The reasons for looking at the dimensionality of the current Section 3 as part of a feasibility study of a revised Section 3 are implicit in the assumptions of IRT. If the assumptions of the IRT model are met for the current Section 3, then the deletion of the vocabulary subpart from the test will have no psychometric consequences. To the extent that the vocabulary and reading comprehension subparts measure distinct ability dimensions, however, the unidimensional IRT model will actually measure a composite of these two dimensions, and eliminating the vocabulary subpart may have undesirable effects on the IRT scale for Section 3.

Data

The data used for the dimensionality assessments were spaced samples of 2,903 Section 3 item-response vectors from the September 1991 administration and 2,761 Section 3 item-response vectors from the October 1991 administration. To maintain consistency with the sampling procedures used in operational equatings of the TOEFL test, only candidate records from domestic test centers were sampled for the study. The test forms selected for study included 29 scored vocabulary questions and 29 scored reading comprehension questions. There were six reading passages in the September TOEFL, and five reading passages in the October TOEFL. The inclusion of six reading passages in the September form was unusual. In fact, TOEFL test development specifications have been adopted since this form was assembled that formally limit the number of reading passages in Section 3 to five.

Methods

Three techniques for assessing the dimensionality of dichotomously-scored test items were employed: 1) Stout's procedure for assessing essential unidimensionality (Stout, 1987; Nandakumar, 1991); 2) applications of multidimensional IRT models as implemented by the computer programs CONFIRM (McKinley, 1989) and NOHARM (Fraser, 1983); and 3) comparisons of full test versus content-based unidimensional calibrations of the data (Bejar, 1980).

Stout's Procedure. The nonparametric approach for assessing latent trait unidimensionality developed by Stout (1987) was applied to both the September 1991 and October 1991 data. This theory-based procedure tests the hypothesis H_o: $d_E = 1$ versus the alternative hypothesis H_1: $d_E > 1$. A statistical index, T, is computed and the hypothesis H_o is rejected if $T \geq Z_\alpha$, where Z_α is the upper $100(1 - \alpha)$ percentile of the standard normal distribution, with α as the desired level of significance.

Stout's procedure was carried out using a series of computer programs written by Stout and his colleagues. A brief description of the steps involved in the procedure is provided below. For additional theoretical and procedural details, see Stout (1987) and Nandakumar (1991).

In the first step of Stout's procedure, the data are divided in half. The first half of the data is used to perform a principal factor analysis of the interitem tetrachoric correlations. Two factors are extracted, and the unrotated loadings on the second factor are used to identify a homogeneous subset of M items called "assessment subtest 1", or AT1. A second assessment subtest (AT2) of M items is also selected to be parallel in difficulty to AT1. The remaining items on the test make up a partitioning subtest (PT) which is used to divide examinees into subgroups. Subgroups are defined in terms of the number of correct scores on the set of partitioning items. Only PT scores with frequencies greater than or equal to some specified minimum (usually 20) are included in the calculation of the test statistic.

A test statistic, T_L, is obtained by calculating two estimates of the variance of the AT1 proportion-correct scores for the examinees at each PT number-correct score level. One of these estimates is an examinee-wise estimate and consists of the variance of the examinee proportion-correct scores. The second estimate is an item-wise estimate and consists of the variance of the proportion-correct scores across the items in AT1 for that particular PT subgroup. The difference between these two variance estimates is normalized within each subgroup and combined across subgroups to produce T_L. Because T_L tends to exhibit a positive bias for tests that are not long and is also susceptible to bias if AT1 consists of items that are overly homogeneous with respect to item difficulty, the above procedures are repeated on the items in AT2 to produce a second statistic, T_B. The overall test statistic is then defined as $T = (T_L - T_B) / \sqrt{2}$.

In applying Stout's procedure to the two data sets used in this study, all item responses were scored right/wrong, that is, items omitted and not reached were scored wrong. For each data set, the procedure was applied to all 58 items, to the vocabulary and reading comprehension subparts separately, and to shortened (47-item) tests with the items in the last two passages excluded.

Multidimensional calibrations. One-, two-, and three-dimensional calibrations of each data set were performed using the computer programs CONFIRM (McKinley, 1989) and NOHARM (Fraser, 1983). Although confirmatory analyses are possible using both of these procedures, only exploratory analyses were undertaken for this study. For

the CONFIRM program, exploratory calibrations result in parameter estimates of a multidimensional three-parameter logistic model, or M3PL, given by:

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

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

Evaluation of model-data fit for the CONFIRM program in the present study was done by comparing the one-, two-, and three-dimensional solutions in terms of three model-data fit statistics: 1) a likelihood ratio chi-square statistic (Bock, Gibbons, and Muraki, 1985); 2) Akaike's (1987) AIC statistic; and 3) a variation on the AIC statistic, called the consistent AIC statistic, or CAIC (Bozdogan, 1987). These statistics are described in detail in a previous TOEFL research report (McKinley and Way, 1992).

The program NOHARM is based on the work of McDonald in nonlinear factor analysis (1967; 1982), and uses harmonic analysis to fit a multidimensional normal ogive model, which can be stated as:

$$P_i(\underline{\theta}_j) = c_i + (1 - c_i)\, N[f_{Oi} + \underline{f}_{1i}'\underline{\theta}_j],(2)$$

where $P_i(\underline{\theta}_j)$ is the probability of a correct response to item i by examinee j, N represents the normal distribution function, f_{Oi} is a threshold value for item i, $\underline{f}_{1i}$ is a k by 1 vector of coefficients for item i, $\underline{\theta}_j$ is a k by 1 vector of factor scores for examinee j, and there are k dimensions.

NOHARM computes and prints out the residual item covariances after fitting the model, and provides the root mean square of the residual item covariances. McDonald (1981) recommends inspecting the residual matrix for clusters of large values that may indicate groups of items that do not fit the model. Fraser (1983) suggests that, "If the root mean-square residual is in the order of the typical standard error of the residuals (4 times the reciprocal of the square root of the sample size) we have a rough indication that a refined test of significance would not reject the hypothesized model" (p. 1). Model-data fit based on NOHARM was evaluated by examining the residual matrices and the root mean-square residual for each solution.

As with Stout's procedure, both the CONFIRM and NOHARM programs were applied to data scored right/wrong. In addition, the guessing parameter was fixed at 0.2 for all items using both programs.

Full test versus content-based calibrations. Because the proposed content change to Section 3 of the TOEFL test will involve the elimination of the vocabulary subpart, it is important to compare how this change might affect unidimensional calibrations of the

reading items. One procedure that can be used to investigate this question was proposed by Bejar (1980). This procedure compares item-parameter estimates based on calibrating all the items in a test (or full-test calibrations) with item-parameter estimates obtained from separate calibrations of different content-based parts of the test. In the present study, full-test and content-based calibrations were compared for the reading comprehension subpart only. This comparison was considered to be analogous to the situation that will occur after the proposed change to Section 3 is made. That is, calibrations of reading comprehension items will be linked back to an IRT scale that was originally based on both reading comprehension and vocabulary items.

The full-test and content-based calibrations were carried out using the LOGIST computer program (Wingersky, Patrick, and Lord, 1988). For these calibrations, the data were scored to allow for items omitted and not reached. As with the multidimensional calibrations, the guessing parameter was fixed at 0.2. For both data sets, the a- and b-parameter estimates of the reading comprehension items resulting from the content-based calibrations were compared with those resulting from the full-test calibrations. To place the estimates from the two different calibrations on the same ability scale, the content-based estimates were transformed to the scale defined by the full-test estimates. This was accomplished by a straightforward mean and sigma transformation of the examinee ability estimates.

To aggregate the differences in the content-based and full-test item-parameter estimates, true-score equatings were carried out between the two sets of reading comprehension item parameters, and the extent to which the true-score equating relationship deviated from identity was examined.

Results

Summary Statistics. The Section 3 number-correct score means, standard deviations, minimums, and maximums presented in Table A.1 indicate that the September 1991 test form was slightly more difficult and the scores were more variable than the October form. The KR-20 estimated reliabilities were moderately high, .92 and .91 for the September and October forms, respectively. It can also be seen from Table A.1 that the odd/even correlations were considerably higher than the vocabulary/reading comprehension correlations for both forms, .87 versus .76 for the September form, and .85 versus .74 for the October form. These results suggest the existence of some structural differences between the two subparts of the test.

The extent to which sufficient time is allowed for completing the test is assessed by computing the percentages of examinees completing successively smaller numbers of items, until the number of items completed by virtually all (99.5 percent) the examinees has been computed. According to the general ETS guidelines as detailed by Walker (1981), a test is usually regarded as essentially unspeeded if at least 80 percent of the examinees reach the last question and if virtually everyone reaches at least 75 percent of the items. It can be seen from the bottom of Table A.1 that timing may have been somewhat more of a problem for the September form, where 99.5 percent of the

examinees completed 48 (83 percent) of the 58 items, versus 52 items (90 percent) for the October form. The percent of examinees completing both forms was quite high, however: 96.1 percent for the September form and 97.6 percent for the October form. This suggests that speededness in the traditional sense is not a problem for either.

These results are consistent with typical TOEFL administrations. It should be noted, however, that the criteria mentioned above have limitations for rights-only tests, as random or patterned responding at the end of a test is encouraged for examinees who do not have time to attend to every question.

Tests of Essential Unidimensionality. Table A.2 provides a summary of the results of the tests of essential unidimensionality as obtained from Stout's procedure. These analyses were performed separately for the entire section (58 items), the vocabulary subpart (29 items), the reading comprehension subpart (29 items), and the shorter section with the items from the last two passages omitted (47 items). For each form, Table A.2 provides the first and second eigenvalues and the ratio of the first to the second from the principal factor analysis of the interitem tetrachoric correlations. Also provided are the assessment subtest (AT1 and AT2) T statistics and Stout's T statistic.

The T statistics in the last row of Table A.2 indicate that the test of essential unidimensionality was rejected at a 0.05 level of significance for the entire sections and for the reading comprehension subparts of both forms. For the complete sections, the overall T statistics were 8.48 and 2.16 for the September and October forms, respectively; while for reading comprehension subparts alone, these values were 7.05 and 3.14. Inspection of the results from these analyses revealed that for both forms, the subtest AT1 was comprised of the items from the last passage. This finding was of significance because, as previously mentioned, in Stout's procedure the selection of items for AT1 is based on a principal factor analysis of the interitem tetrachoric correlations with two factors extracted. Those items with the highest absolute loadings on the unrotated second factor are typically selected for AT1. When Stout's procedure was performed after eliminating the items in the last two reading passages, the resulting T statistics were not statistically significant. Furthermore, the items selected for AT1 included both reading and vocabulary items. This suggested that the departures from essential dimensionality for the two sections were primarily due to end-of-test effects that might be related to speededness.

Multidimensional Calibrations. Table A.3 summarizes the exploratory multidimensional CONFIRM and NOHARM analyses for both forms. The likelihood ratio chi-square, Akaike's AIC, consistent AIC (CAIC) model-data fit statistics, and the NOHARM root mean-square residuals for the one-dimensional, two-dimensional, and three-dimensional solutions are shown. On the basis of the chi-square and AIC statistics, the three-dimensional solutions obtained from CONFIRM would initially appear to be optimal. The decreases in magnitude in these statistics from the two-dimensional to the three-dimensional solutions were relatively small, however, suggesting the improvements in fit provided by the three-dimensional solutions were not of practical consequence.

In contrast to the chi-square and AIC statistics, the CAIC statistics in Table A.3 indicate the two-dimensional solution would be optimal for both forms. The selection of the simpler model based on the CAIC statistic is consistent with the fact that this statistic includes a strong penalty for overparameterization. McKinley and Way (1992) point out that selecting the CAIC statistic over the AIC statistic is equivalent to selecting a larger critical value for testing whether a particular model fits best, thus reducing the Type I error rate.

The NOHARM root mean-square residuals at the bottom of Table A.3 also are lowest for the three-dimensional solution. However, with large sample sizes, adding a factor would normally cause this statistic to become smaller. Additionally, as with the likelihood chi-square and AIC statistics from the CONFIRM program, the relative decreases in the magnitudes of the NOHARM root mean-square residuals in going from the one- to the two-factor solutions are greater than the decreases seen when going from the two- to the three-factor models. These findings were consistent with the CONFIRM results, and suggested the two-factor solutions provided adequate fit to the data. Therefore, further investigations of the CONFIRM and NOHARM results were undertaken only for the two-dimensional solutions.

Figures A.1a to A.2b display bivariate plots of the two-dimensional item discriminations from the CONFIRM and NOHARM solutions. Note that as in the case of factor analysis, these solutions are subject to rotational indeterminacy. No attempts at rotating the solutions were undertaken as part of this study. Therefore, the magnitudes of the discrimination-parameter estimates on each ability dimension should be interpreted with caution, although the patterns of the estimates do provide meaningful information.

Bivariate plots of the two-dimensional item-discrimination estimates from the CONFIRM analyses are presented in Figures A.1a and A.1b. In each plot, the vocabulary (VOC) subpart is represented by squares, reading passages 1-4 (September form) and 1-3 (October form) by pluses, and reading passages 5 and 6 (September) and 4 and 5 (October) are represented by diamonds. In Figure A.1a for the September form, the items in passages 5 and 6 appear to form a distinct cluster, suggesting the presence of a second ability dimension. The plot for the October form (Figure A.1b) indicates a less pronounced but similar separation of the items in the last two passages from the remainder of the test. In this plot, the items in the last two passages cluster below the rest of the section, and the vocabulary items appear to form another, although less distinct, cluster.

Figures A.2b and A.2b display the bivariate plots of the two-dimensional discrimination estimates from the NOHARM analyses. The pattern in Figure A.2a is similar to that observed in Figure A.1a, with the items in the last two passages of the September form clustering distinctly from the remainder of the section. In Figure A.2b, a more distinct cluster of the items in the last two passages in the October form is observed than was seen in Figure A.1b. The vocabulary items also tend to cluster distinctly from the reading items as a whole, although this clustering is not clear-cut.

It would appear from inspection of Figures A.1a-A.2b that the data suggest the presence of a dimension related to end-of-test effects, especially for the September form. The patterns of Figures A.1a and A.2b are very similar, with the last two passages loading heavily on the second dimension; whereas Figures A.1b and 2B present a less obvious separation of Section 3 into two dimensions. Although both forms were comprised of the same number of items, the September form contained six reading passages, instead of five as is typical of TOEFL forms. This may have accounted for the more pronounced patterns in the plots based on the September data.

Full-Test versus Content-Based Calibrations. Bivariate plots of the content-based (RC) discrimination estimates against the full test-based estimates are presented in Figures A.3a and A.3b. In these plots, passages 1-4 (September form) and 1-3 (October form) are represented by squares, and passages 5 and 6 and 4 and 5 are represented by pluses. In Figure A.3a, the items in the last two passages of the September form fall slightly above the remaining passages, which generally fall close to the diagonal. This pattern is less distinct for the October form, where all the items fall close to the diagonal, although the items from passages 4 and 5 are all above the diagonal. Based on these results, it is reasonable to assume that the effects of calibrating reading items in the absence of vocabulary items on the resulting discrimination-parameter estimates will not be practically significant, particularly if the timing of the revised Section 3 is adequately addressed.

The full-test versus RC-based true-score equating differences are presented in Figure A.4. The differences are plotted as a function of the number-right true scores. It can be seen that the differences between the two forms follow similar patterns for most of the true-score scale, where these differences are less than 0.10. The largest differences, from 0.10 to 0.20, occur for the September form at the upper end of the scale, between the true scores of about 21-29. For the October form, the differences at this end of the scale are trivial. These results suggest that any slight departures from dimensionality that are due to end-of-test effects have minimal effect on the equating results for Section 3.

Discussion

The purpose of the preliminary dimensionality analyses of Section 3 of TOEFL was to assess the possible effects on the IRT scale of a revised Section 3. If the assumptions of the IRT model are met for the current Section 3, the deletion of the vocabulary subpart will have no psychometric consequences. The data for this part of the study were comprised of scored item responses for 29 vocabulary and 29 reading comprehension items from two operational forms of TOEFL. The September 1991 form included six reading passages, whereas the October 1991 form included the more typical five passages.

The results of the preliminary analyses indicated some departures from unidimensionality for Section 3 and for the reading subpart, especially in the September form. In particular, the departures from unidimensionality appeared to be primarily due

to end-of-test effects that are probably related to the timing of the test. It is worth noting that similar results were reported by Oltman, Stricker, and Barrows (1988) on the basis of a three-way multidimensional scaling of the TOEFL test using samples of examinees that systematically varied in native language and level of English proficiency. These authors concluded, "The end-of-test dimension is an artifact of the timing of the test, for it would disappear if the time limit for the TOEFL section involved (Vocabulary and Reading Comprehension) were increased to permit everyone to finish" (p. 29).

Despite the salience of the end-of-test phenomenon seen in the preliminary dimensionality analyses, it is unlikely that revising Section 3 to eliminate discrete vocabulary items will have noticeable effects on the IRT-based Section 3 score scale. On the other hand, the results of the preliminary dimensionality analyses clearly underscored the need for systematic research to determine appropriate test length and time limits for the revised Section 3.

Table 1: Summary Statistics for Form 3MTF12 Administered in the Operational Samples and in the TOEFL Institutional Samples Participating in the Study

Number Correct Scores

		Section 1		Section 2		Section 3	
Sample	N	Mean	Std.	Mean	Std.	Mean	Std.
Op. Foreign	13735	30.6	11.1	26.4	7.6	40.7	11.6
Op. Domestic	8877	33.2	9.6	24.8	7.3	38.2	10.9
Inst. TOEFL	1323	29.5	9.4	22.7	7.2	34.5	10.4
Exp. 48 Items	434	28.9	9.2	22.6	7.2	33.7	10.1
Exp. 54 Items	449	30.0	9.5	23.0	7.0	35.1	10.7
Exp. 60 Items	440	29.6	9.6	22.6	7.3	34.6	10.5
Exp. 50 Min.	430	28.0	9.8	21.8	7.3	33.2	10.7
Exp. 55 Min.	366	31.7	9.4	23.3	7.3	34.6	10.3
Exp. 60 Min.	527	29.2	8.9	23.1	6.9	35.4	10.3

Scaled Scores

		Section 1		Section 2		Section 3		Total Scores	
Sample	N	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.
Op. Foreign	13735	50.3	8.2	51.7	8.4	52.2	8.5	514.0	77.5
Op. Domestic	8877	52.1	7.0	50.0	7.8	50.4	7.8	508.3	68.9
Inst. TOEFL	1323	49.4	6.8	46.8	7.5	47.6	7.5	479.5	65.2
Exp. 48 Items	434	48.9	6.5	46.7	7.6	47.1	7.3	475.8	64.0
Exp. 54 Items	449	49.8	6.9	47.0	7.3	48.0	7.6	482.8	65.5
Exp. 60 Items	440	49.5	6.9	46.7	7.5	47.7	7.4	479.7	65.9
Exp. 50 Min.	430	48.3	7.2	45.9	7.8	46.7	7.8	469.6	69.2
Exp. 55 Min.	366	51.0	6.6	47.3	7.6	47.8	7.3	487.0	63.8
Exp. 60 Min.	527	49.1	6.3	47.2	7.1	48.3	7.2	482.3	61.8

Note: The Institutional TOEFL conditions were identical for all groups.

Table 2: Summary of Traditional Speededness Analyses for the Experimental Section 3 by Test Length and Timing Conditions

Test Length	Time	N	100%	75%	σ^2_{NR}/σ^2_R	KR-20
48 Items	50 Min.	135	88.9	96.3	0.21	0.87
	55 Min.	120	90.8	96.7	0.16	0.87
	60 Min.	170	97.6	100.0	0.02	0.90
54 Items	50 Min.	145	86.2	93.1	0.36	0.89
	55 Min.	120	87.5	97.5	0.12	0.89
	60 Min.	175	86.9	98.3	0.29	0.91
60 Items	50 Min.	140	77.9	93.6	0.27	0.89
	55 Min.	120	81.7	99.2	0.12	0.92
	60 Min.	170	95.3	100.0	0.04	0.87
Inst. TOEFL	45 Min.	1320	93.9	99.2	0.04	0.90

Note: N = sample size, 100% = percent completing the test, 75% = percentage of examinees completing three-fourths of the test, σ^2_{NR}/σ^2_R = variance of not-reached items divided by variance of rights, KR-20 = Kuder-Richardson Formula 20 reliability. Analysis program deleted cases at random to reduce sample sizes to a multiple of five. Analyses included one examinee who answered only the first item on the test. Speededness statistics for Section 3 of the Institutional TOEFL form across all samples is provided in the last row of the table.

Table 3: Low Outlier Examinees Compared with Entire Experimental Section 3: Average Number Correct on Common Items per Passage

	Passage Number					
Group	1	2	3	4	5	6
Low Outlier	5.3	1.3	1.5	1.3	1.4	1.1
Entire Sample	5.5	4.9	5.3	4.7	3.3	3.3

Table 4: Average Number Correct Score on the Common 47 Items of the Experimental Section 3

Test Length	Time	N	Average Score
48 Items	50 Min.	134	24.7
	55 Min.	119	26.9
	60 Min.	168	28.1
54 Items	50 Min.	149	26.7
	55 Min.	118	28.2
	60 Min.	169	28.7
60 Items	50 Min.	142	24.6
	55 Min.	119	28.6
	60 Min.	171	28.9

Table 5: Summary of Covariate Adjustment Model Fit

	μ	α_2	α_3	β_2	β_3	γ	τ^2	σ^2
Estimate	3.04	0.30	0.07	1.51	1.38	0.67	1.3	20.0
SE		0.3	0.3	0.6	0.5	0.01	0.04	

Table 6: Alternate-Form Summary Statistics for the Institutional TOEFL and Experimental Test Results -- All Valid Candidates Taking Form 1C (60 Item Form)

	Test Form Summary Statistics					
Form / Score	N	Mean	SD	Min	Max	$r_{xx'}$
Inst. S1 Number Correct	439	29.7	9.6	8	49	0.87
Exp. S1 Number Correct	439	28.5	9.3	9	49	0.87
Inst. S1 Converted	439	49.5	6.9	31	66	0.85
Exp. S1 Converted	439	48.9	6.1	33	66	0.85
Inst. S2 Number Correct	439	22.6	7.3	2	38	0.83
Exp. S2 Number Correct	439	22.8	7.2	6	38	0.83
Inst. S2 Converted	439	46.7	7.5	20	68	0.79
Exp. S2 Converted	439	46.4	7.5	24	68	0.79
Inst. S3 Number Correct	439	34.6	10.5	11	56	0.82
Exp. S3 Number Correct	439	29.9	9.2	9	49	0.82
Inst. S3 Converted	439	47.8	7.4	28	64	0.81
Exp. S3 Converted	439	48.0	7.7	28	67	0.81
Inst. Total Score	439	480.0	65.6	303	653	0.91
Exp. Total Score	439	477.5	62.3	303	670	0.91

Note: Alternate-form reliabilities ($r_{xx'}$) based on Pearson product-moment correlations.

Table 7: Alternate-Form Summary Statistics for the Institutional TOEFL and Experimental Test Results -- Candidates Taking Form 1C (60 Item Form) with Outliers Removed

	Test Form Summary Statistics					
Form / Score	N	Mean	SD	Min	Max	$r_{xx'}$
Inst. S1 Number Correct	432	29.7	9.5	9	49	0.88
Exp. S1 Number Correct	432	28.5	9.3	9	49	0.88
Inst. S1 Converted	432	49.5	6.8	32	66	0.86
Exp. S1 Converted	432	48.9	6.1	33	66	0.86
Inst. S2 Number Correct	432	22.7	7.2	6	38	0.85
Exp. S2 Number Correct	432	22.9	7.2	6	38	0.85
Inst. S2 Converted	432	46.9	7.3	24	68	0.81
Exp. S2 Converted	432	46.4	7.5	24	68	0.81
Inst. S3 Number Correct	433	34.6	10.4	11	56	0.86
Exp. S3 Number Correct	433	29.9	9.2	9	49	0.86
Inst. S3 Converted	433	47.7	7.4	28	64	0.84
Exp. S3 Converted	433	48.1	7.7	28	67	0.84
Inst. Total Score	430	481.8	63.9	313	653	0.92
Exp. Total Score	430	479.3	62.1	303	670	0.92

Note: Alternate-form reliabilities ($r_{xx'}$) based on Pearson product-moment correlations.

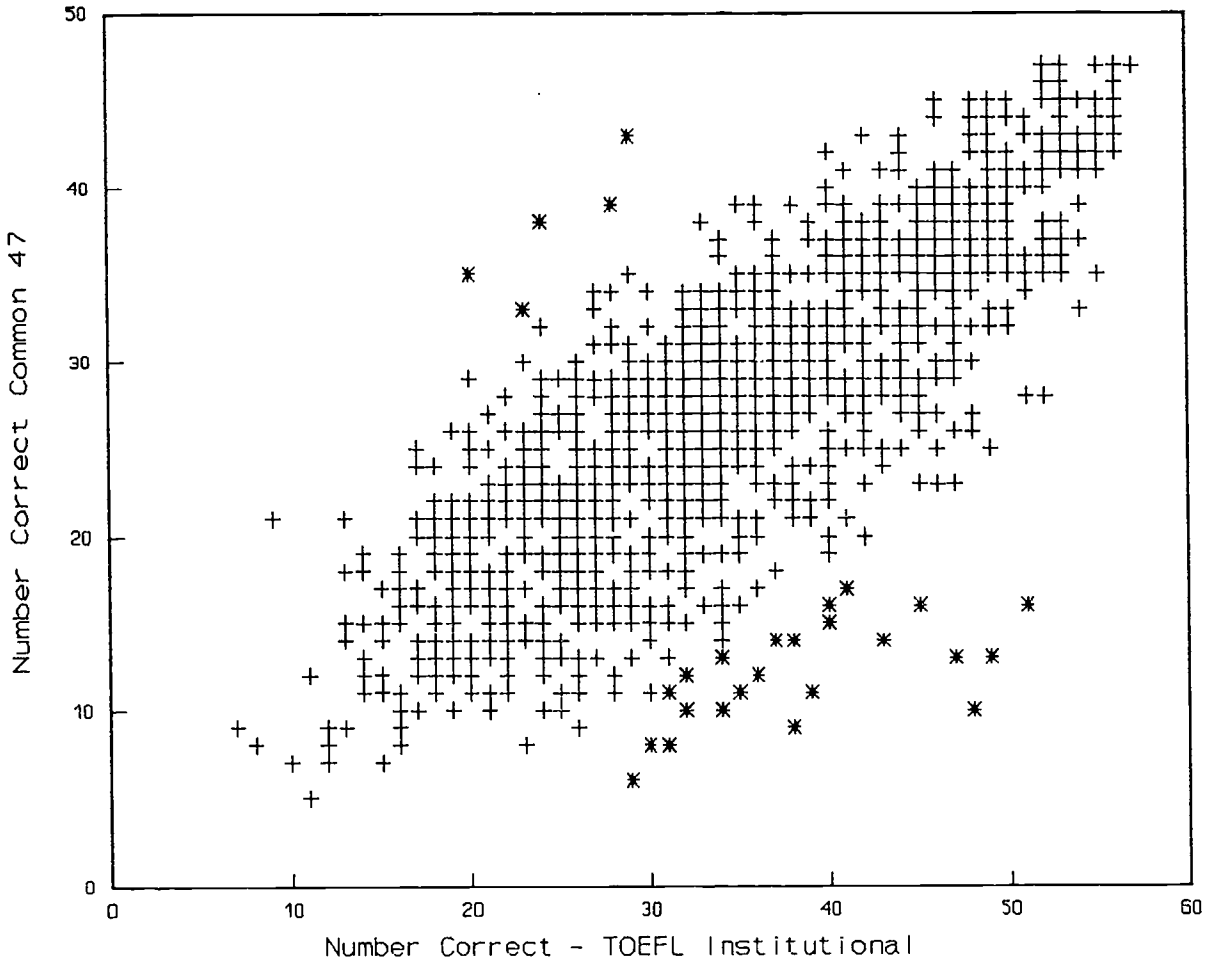


Figure 1: Number Correct Score on the Common 47 Items Plotted Against Number Correct Score on the Institutional TOEFL Section 3 with Outliers Highlighted

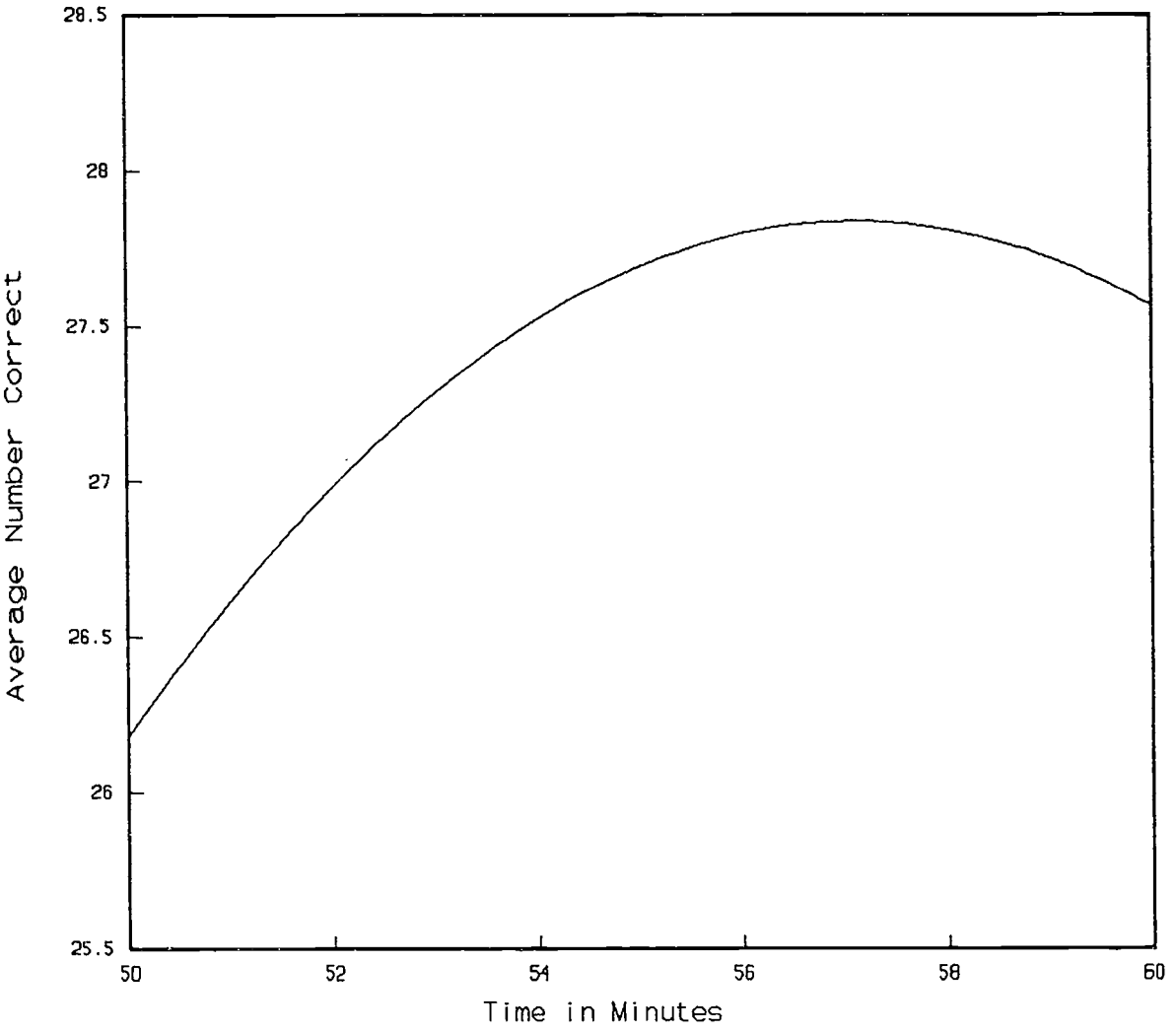


Figure 2: Predicted Increase in Mean Number Correct on the Common 47 Items Plotted as a Function of Increases in Testing Time

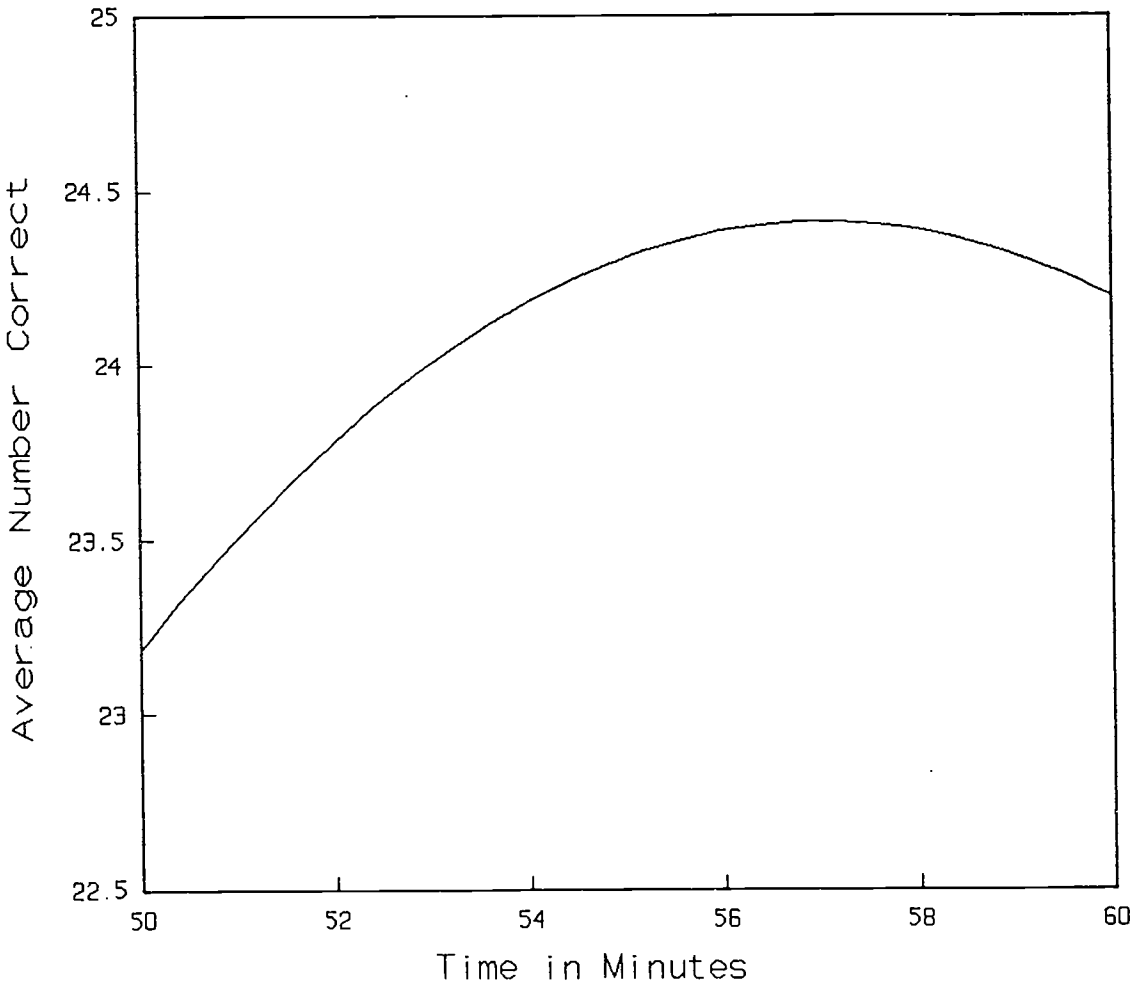


Figure 3: Predicted Increase in Mean Number Correct on the Common 39 Items in the First Five Passages Plotted as a Function of Increases in Testing Time

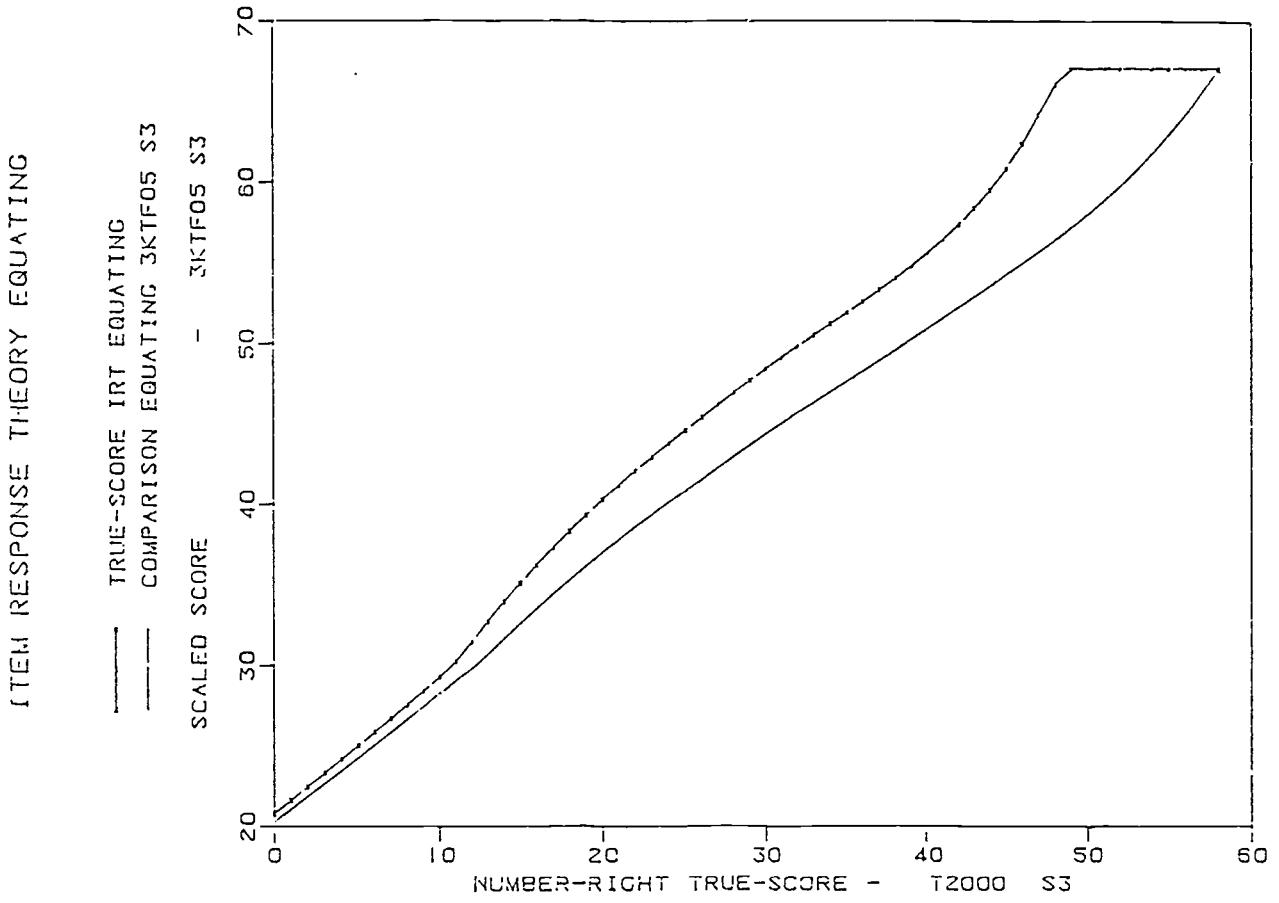
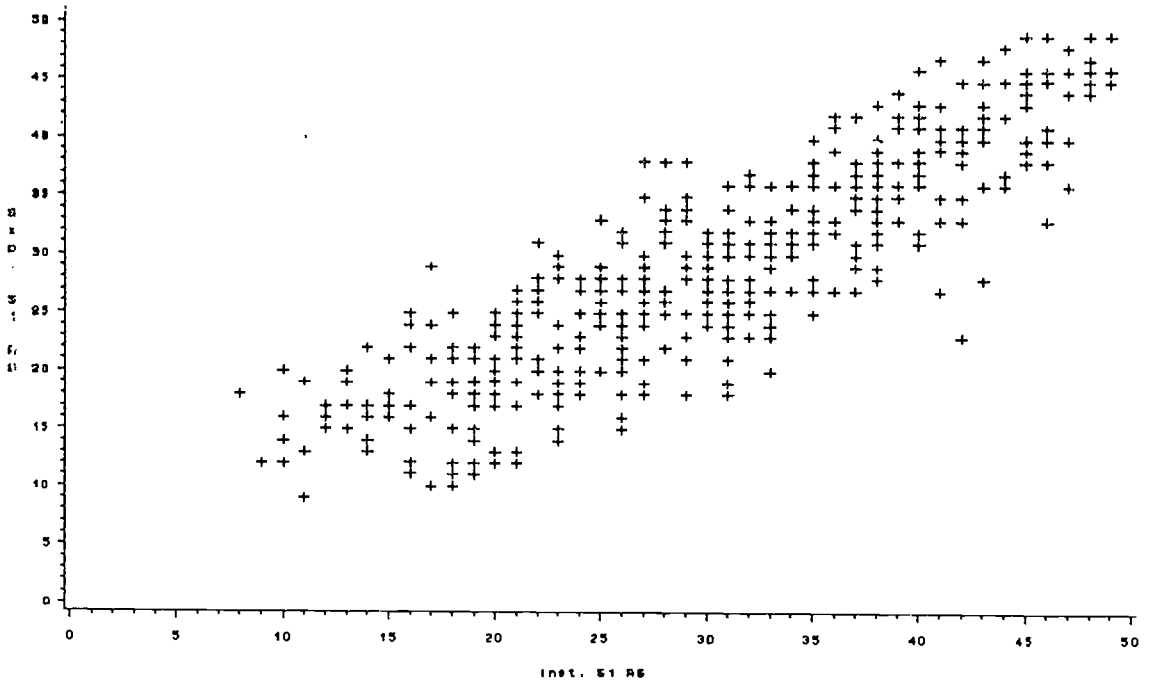
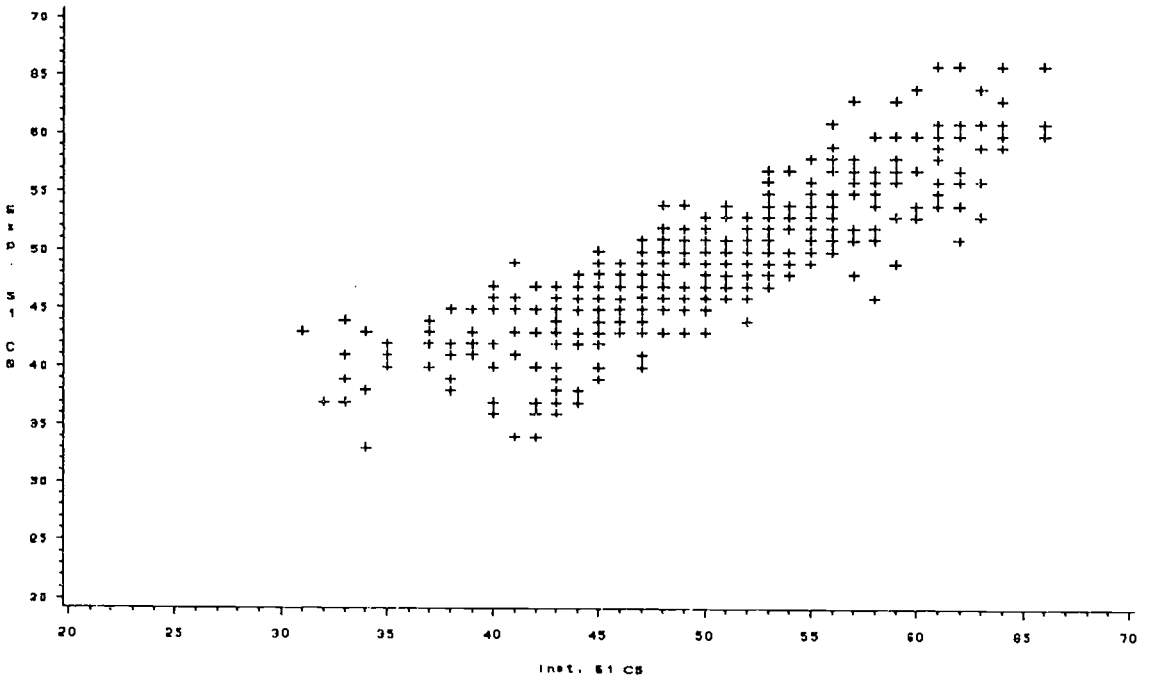


Figure 4: Number Correct to Converted Score Equating Relationship for the Experimental Section 3 Form

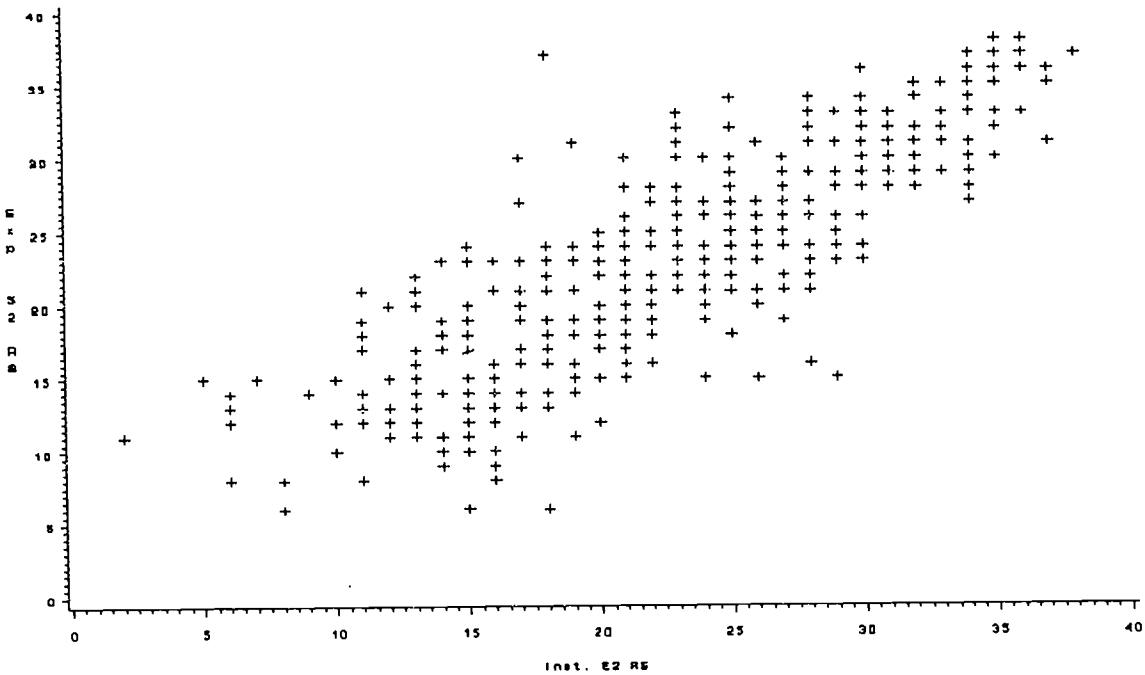


A) Section 1 Number Correct Scores

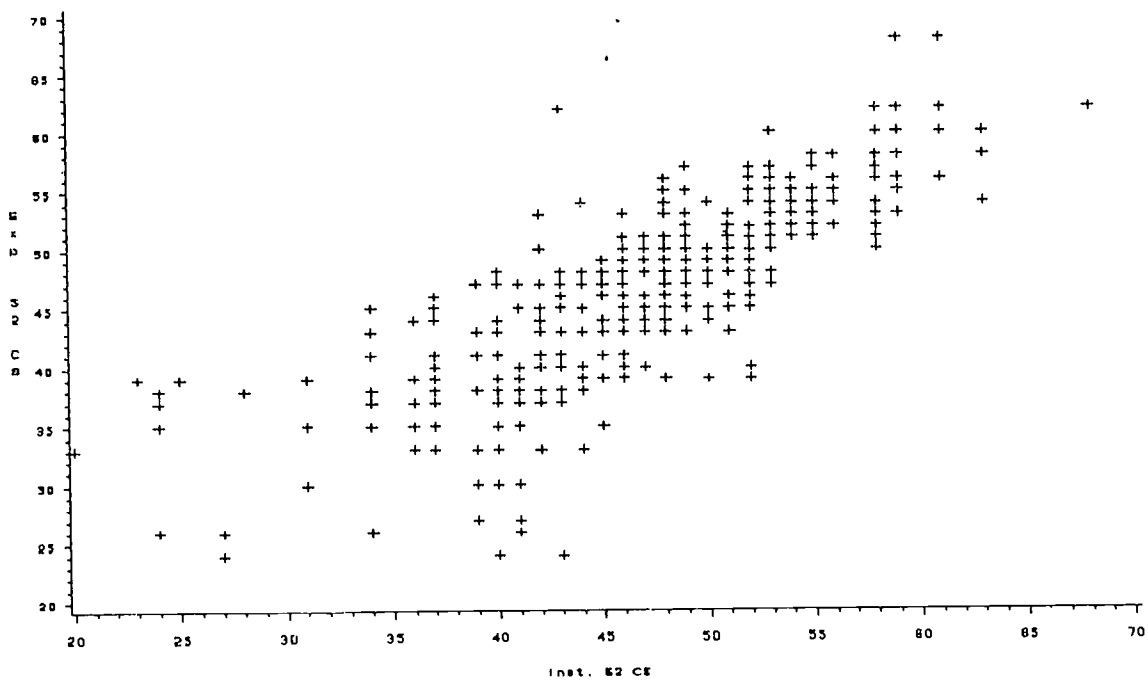


B) Section 1 Converted Scores

Figure 5: Experimental Section 1 Number Correct and Converted Scores Plotted Against Institutional TOEFL Section 1 Number Correct and Converted Scores for Candidates in the 60-Item Conditions

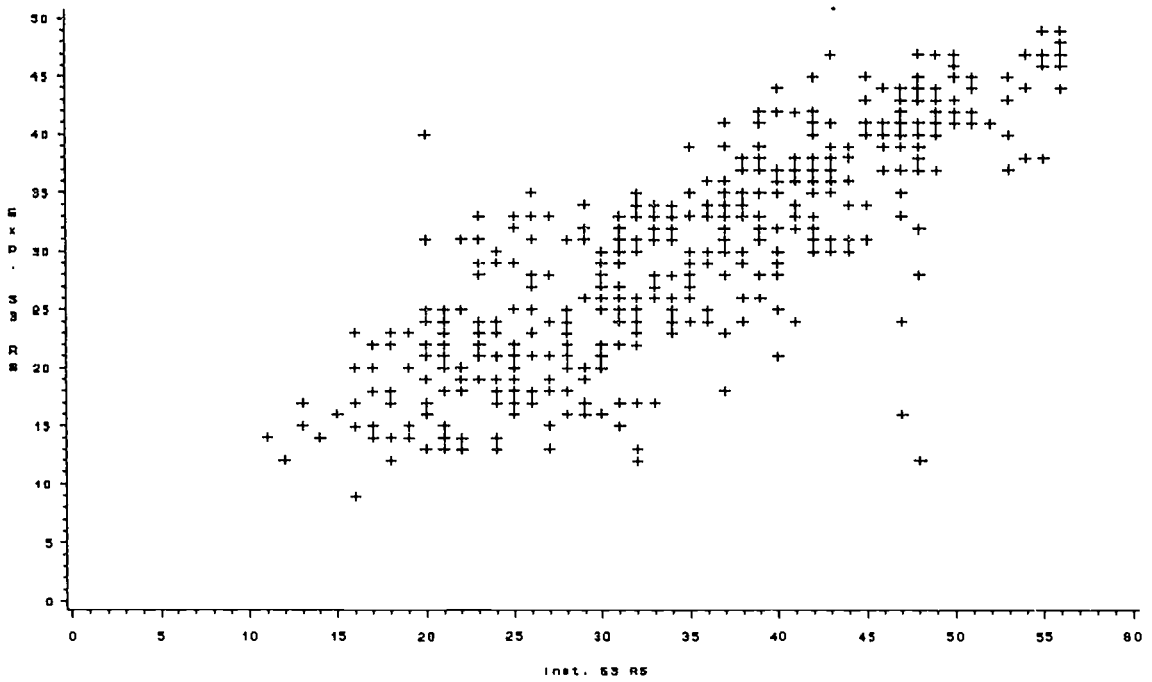


A) Section 2 Number Correct Scores

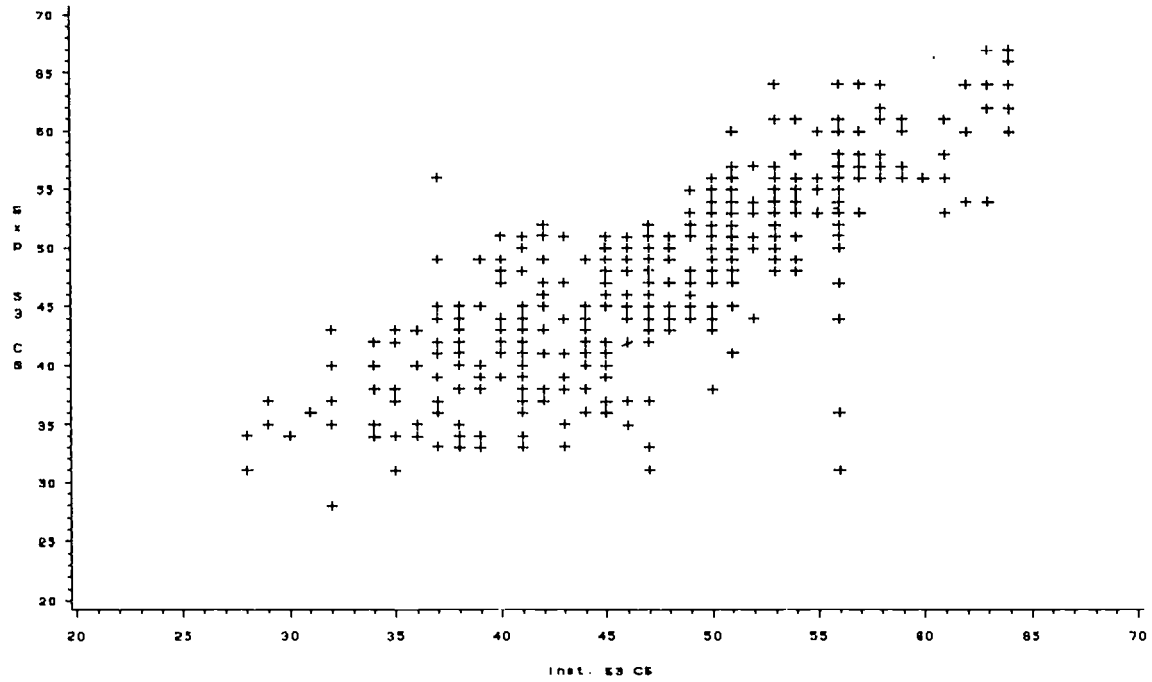


B) Section 2 Converted Scores

Figure 6: Experimental Section 2 Number Correct and Converted Scores Plotted Against Institutional TOEFL Section 2 Number Correct and Converted Scores for Candidates in the 60-Item Conditions



A) Section 3 Number Correct Scores



B) Section 3 Converted Scores

Figure 7: Experimental Section 3 Number Correct and Converted Scores Plotted Against Institutional TOEFL Section 3 Number Correct and Converted Scores for Candidates in the 60-Item Conditions

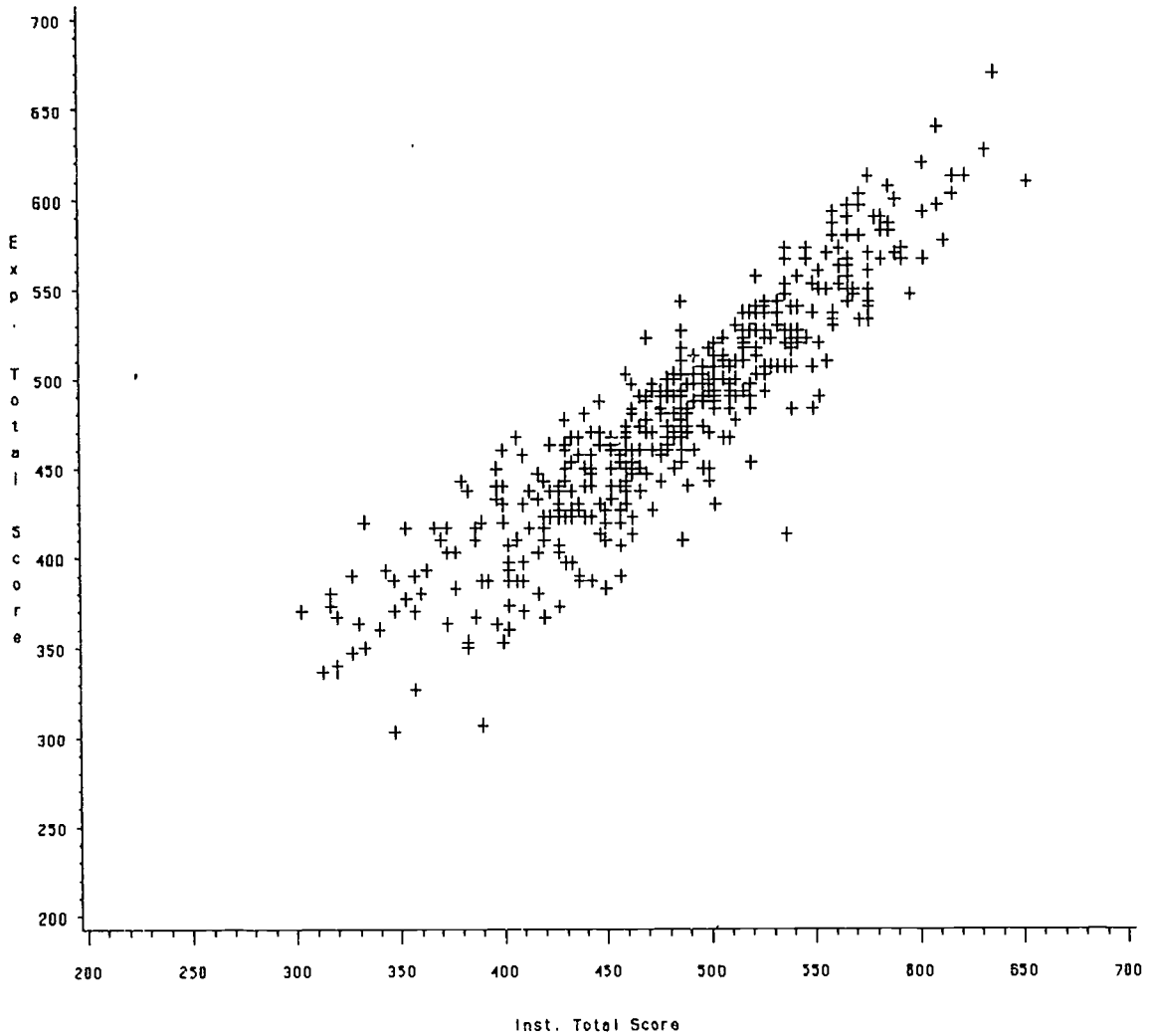


Figure 8: Experimental Total Converted Scores Plotted Against Institutional TOEFL Total Converted Scores for Candidates in the 60-Item Conditions

Table A.1: Raw Score Summary Statistics

	September 1991	October 1991
N	2903	2761
Mean	38.27	38.67
SD	11.35	10.55
Minimum	0	0
Maximum	58	58
r_{vr}	0.76	0.74
r_{oe}	0.87	0.85
KR-20	0.92	0.91

# Items	Percent Completing	
58	96.1	97.6
57	96.7	98.3
56	97.0	98.6
55	97.4	98.8
54	97.7	98.8
53	98.3	98.9
52	98.5	99.5
51	98.8	
50	99.0	
49	99.4	
48	99.5	

Table A.2: Tests of Essential Dimensionality

	Section 3 58 Items		Section 3 47 Items		Vocabulary		Reading Comprehension	
	9/91	10/91	9/91	10/91	9/91	10/91	9/91	10/91
λ_1	17.97	15.99	14.78	13.89	9.80	9.75	9.48	7.40
λ_2	2.09	1.74	1.11	1.61	0.75	1.07	1.38	1.03
λ_1/λ_2	8.60	9.19	13.31	8.60	13.08	9.14	6.88	7.21
AT1 T	12.43	4.37	1.90	3.44	4.90	0.89	10.99	6.38
AT2 T	0.43	1.31	0.40	3.59	2.71	2.88	1.02	1.95
Overall T	8.48	2.16	1.06	-0.10	1.55	-1.41	7.05	3.14
Prob > T	0.00	0.02	0.14	0.48	0.06	0.48	0.00	0.00

Table A.3: Multidimensional Analyses

	9/91			10/91		
CONFIRM	1D	2D	3D	1D	2D	3D
χ^2	131052.7	129272.3	128958.6	121022.9	120036.6	119629.5
AIC	177386.9	175722.4	175524.8	164814.7	163944.4	163653.3
CAIC	178195.8	176935.8	177142.7	165617.8	165149.1	165259.6
NOHARM						
RMS resid.	0.0061	0.0041	0.0038	0.0054	0.0045	0.0039

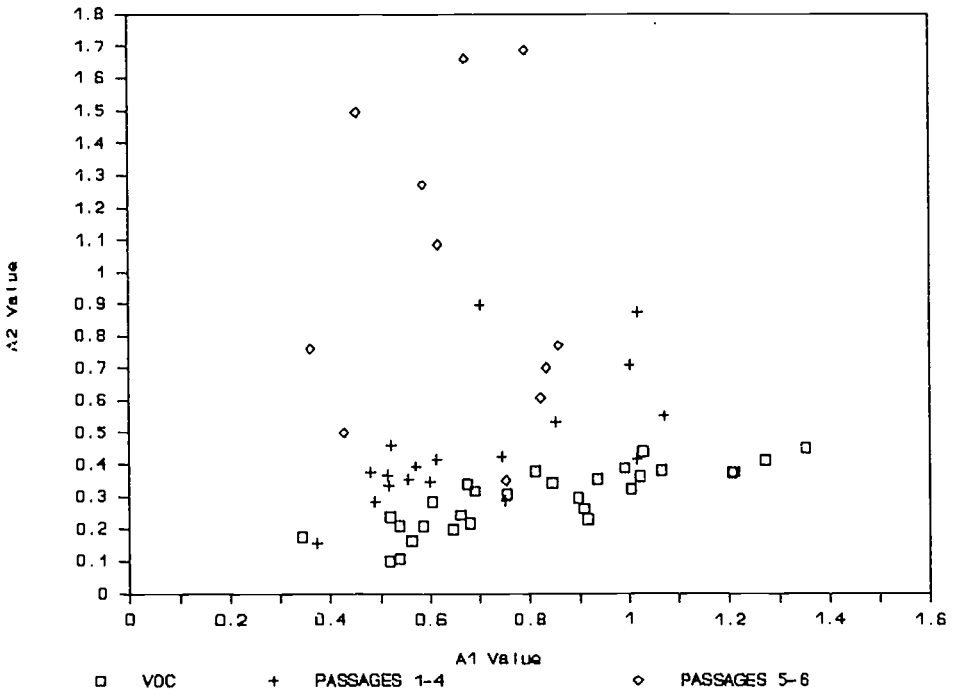


Figure A.1a: A_2 Estimates Plotted Against A_1 Estimates for the CONFIRM Two-Dimensional Exploratory Run - September 1991 Administration

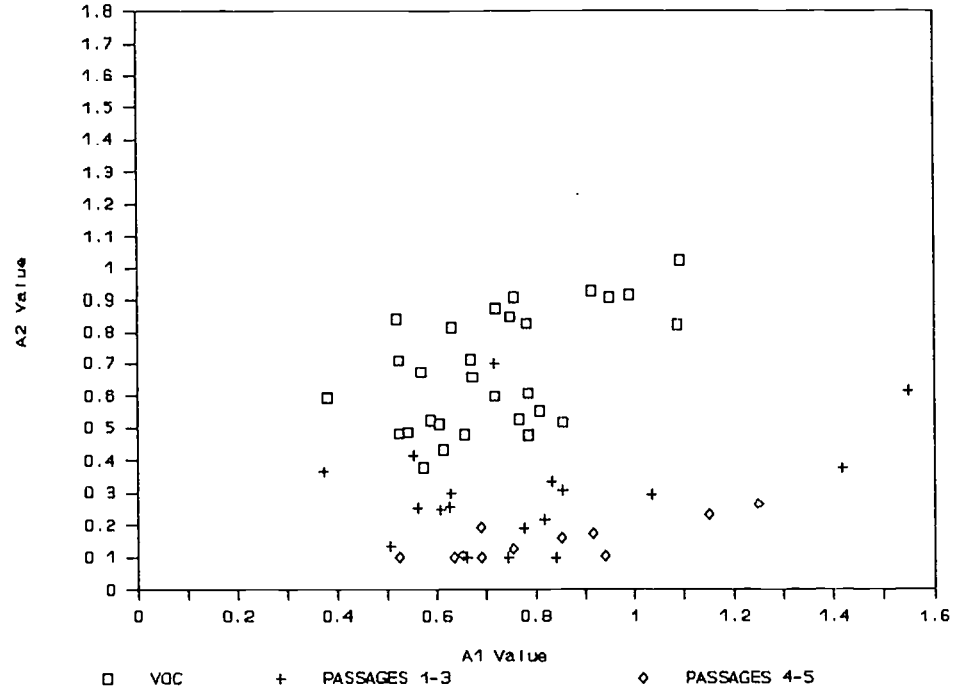


Figure A.1b: A_2 Estimates Plotted Against A_1 Estimates for the CONFIRM Two-Dimensional Exploratory Run - October 1991 Administration

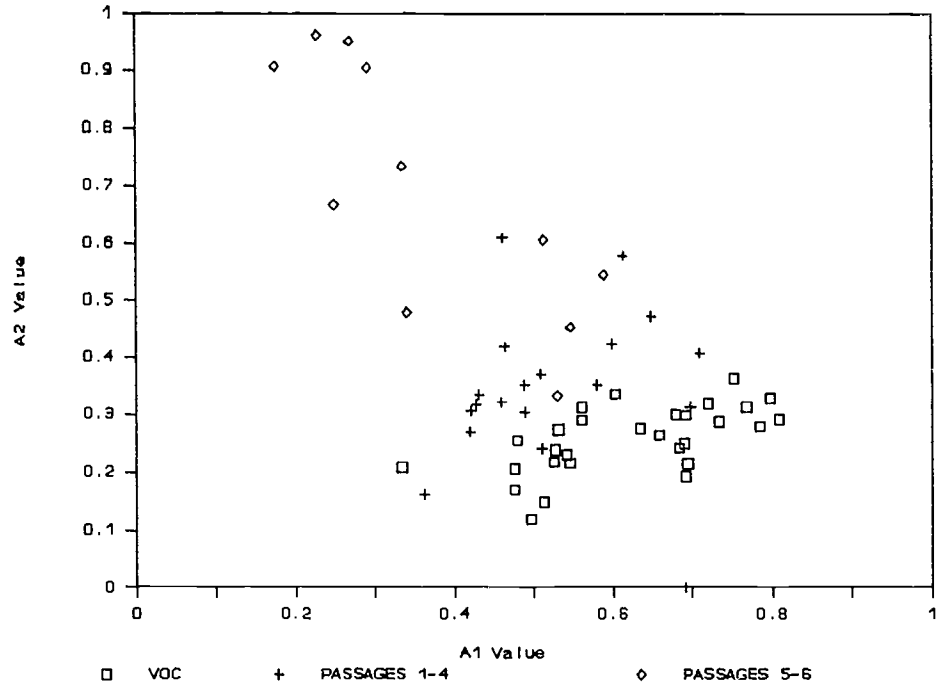


Figure A.2a: A_2 Estimates Plotted Against A_1 Estimates for the NOHARM Two-Dimensional Exploratory Run - September 1991 Administration

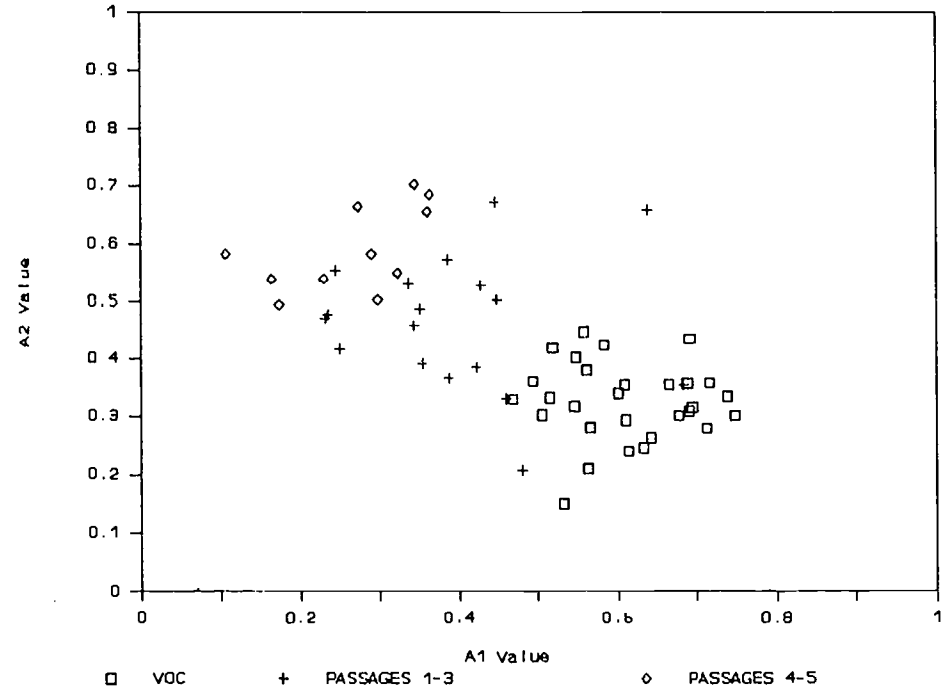


Figure A.2b: A_2 Estimates Plotted Against A_1 Estimates for the NOHARM Two-Dimensional Exploratory Run - October 1991 Administration

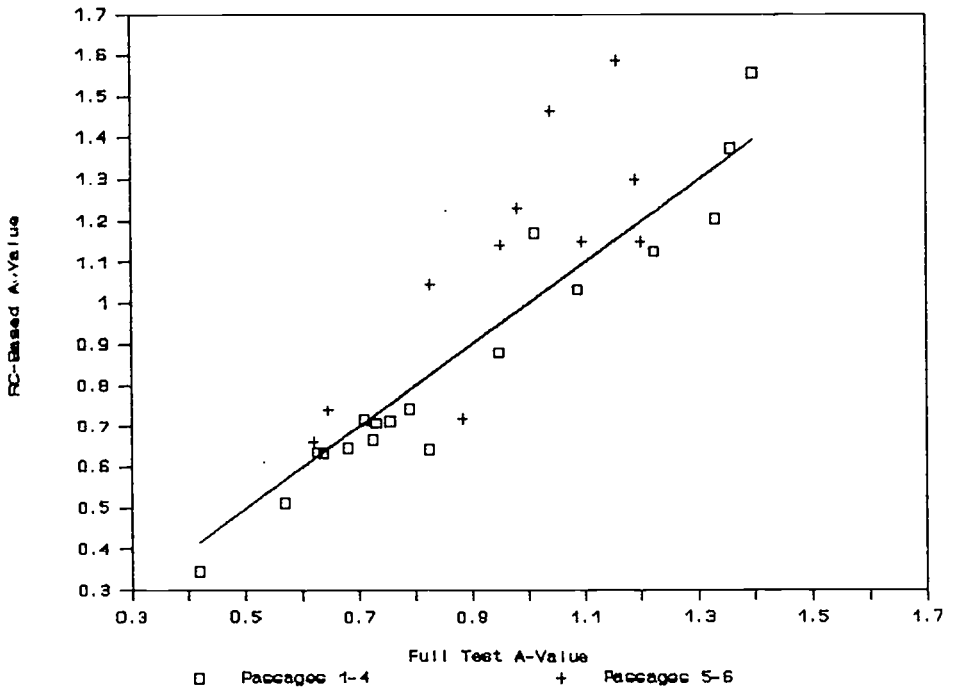


Figure A.3a: RC-Based A-Estimates Plotted Against Total Test-Based A-Estimates for the September 1991 Administration

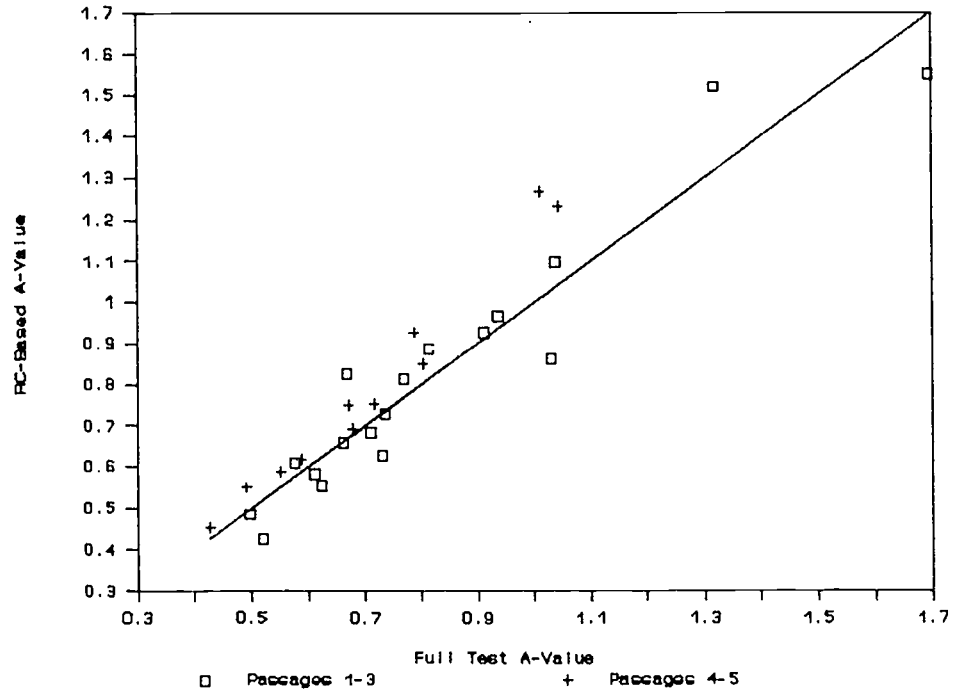


Figure A.3b: RC-Based A-Estimates Plotted Against Total Test-Based A-Estimates for the October 1991 Administration

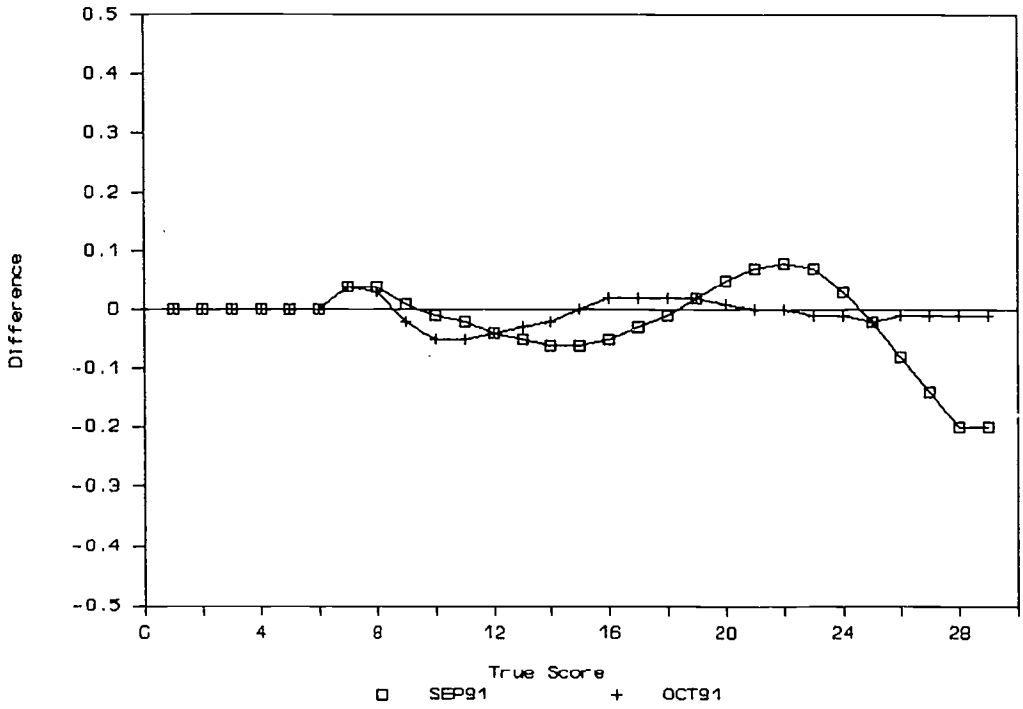


Figure A.4: True Score Equating Differences - Total Test-Based vs. RC-Based

Appendix B

Experimental TOEFL Section 3 with 60 Items

The experimental TOEFL Section 3 had three versions. The 60 items in the longest version are provided on the following pages. The shorter versions of the experimental TOEFL Section 3 contained the same six reading passages with one or two fewer questions associated with each passage. The question numbers that were not included in the shorter versions are listed below:

48-item test: 9, 10, 19, 20, 26, 30, 36, 39, 45, 49, 59, 60.

54-item test: 10, 20, 26, 39, 49, 59.

57

NO TEST MATERIAL ON THIS PAGE

3 • 3 • 3 • 3 • 3 • 3 • 3

Directions: In this section you will read several passages. Each one is followed by several questions about it. You are to choose the one best answer, (A), (B), (C), or (D), to each question. Then, on your answer sheet, find the number of the question and fill in the space that corresponds to the letter of the answer you have chosen.

Answer all questions following a passage on the basis of what is stated or implied in that passage.

Read the following passage:

The railroad was not the first institution to impose regularity on society, or to draw attention to the importance of precise timekeeping. For as long as merchants have set out their wares at daybreak and communal festivities have been celebrated, people have been in rough agreement with their neighbors as to the time of day. The value of this tradition is today more apparent than ever. Were it not for public acceptance of a single yardstick of time, social life would be unbearably chaotic: the massive daily transfers of goods, services, and information would proceed in fits and starts; the very fabric of modern society would begin to unravel.

Line (5)

Example I **Sample Answer**

What is the main idea of the passage? Ⓐ Ⓑ ● Ⓓ

(A) In modern society we must make more time for our neighbors.
(B) The traditions of society are timeless.
(C) An accepted way of measuring time is essential for the smooth functioning of society.
(D) Society judges people by the times at which they conduct certain activities.

The main idea of the passage is that societies need to agree about how time is to be measured in order to function smoothly. Therefore, you should choose answer (C).

Example II **Sample Answer**

In line 5, the phrase "this tradition" refers to Ⓐ Ⓑ Ⓒ ●

(A) the practice of starting the business day at dawn
(B) friendly relations between neighbors
(C) the railroad's reliance on time schedules
(D) people's agreement on the measurement of time

The phrase "this tradition" refers to the preceding clause, "people have been in rough agreement with their neighbors as to the time of day." Therefore, you should choose answer (D).

Now begin work on the questions.

GO ON TO THE NEXT PAGE

Questions 1-10

Alice Walker makes her living by writing, and her poems, short stories, and novels have won many awards and fellowships for her. She was born in Eatonton, Georgia. She went to public schools there, and then to Spelman College in Atlanta before coming to New York to attend Sarah Lawrence College, from which she graduated in 1966. For a time she lived in Jackson, Mississippi, with her lawyer husband and small daughter. About *Langston Hughes, American Poet*, her first book for children, she says, "After my first meeting with Langston Hughes I vowed I would write a book about him for children someday. Why? Because I, at twenty-two, knew next to nothing of his work, and he didn't scold me; he just gave me a stack of his books. And he was kind to me; I will always be grateful that in his absolute warmth and generosity he fulfilled my deepest dream (and need) of what a poet should be.

"To me he is not dead at all. Hardly a day goes by that I don't think of him or speak of him. Once, just before he died, when he was sick with the flu, I took him a sack full of oranges. The joy I felt in giving that simple gift is undiminished by time. He said he liked oranges, too."

1. What is the main topic of the passage?
   (A) Alice Walker's reflections on Langston Hughes
   (B) The influence of Alice Walker on the writing of Langston Hughes
   (C) Langston Hughes' book about Alice Walker
   (D) A comparison of the childhoods of Alice Walker and Langston Hughes

2. In the passage, Alice Walker is described as
   (A) a research fellow at Spelman College
   (B) a professor at Sarah Lawrence College
   (C) a prize-winning writer of prose and poetry
   (D) an author of plays for children

3. Before attending college, Alice Walker went to school in
   (A) Atlanta, Georgia
   (B) Eatonton, Georgia
   (C) Jackson, Mississippi
   (D) Lawrence, Massachusetts

4. The word "vowed" in line 7 is closest in meaning to which of the following?
   (A) Confided to him
   (B) Believed
   (C) Denied
   (D) Promised myself

5. It can be inferred from the passage that Alice Walker was twenty-two years old when
   (A) she moved to Jackson, Mississippi
   (B) she moved to New York
   (C) she first met Langston Hughes
   (D) Langston Hughes died

6. It can be inferred from lines 9-10 that Alice Walker's first impressions of Langston Hughes were derived mostly from
   (A) talking with his friends
   (B) reading his autobiography
   (C) studying his poetry
   (D) meeting him

7. The word "dream" in line 10 is closest in meaning to
   (A) nightmare
   (B) expectation
   (C) sleep
   (D) misconception

GO ON TO THE NEXT PAGE

3 • 3 • 3 • 3 • 3 • 3 • 3

8. What does Alice Walker imply when she says Langston Hughes "is not dead at all" (line 11)?
   (A) Langston Hughes believed in eternal life.
   (B) She had not been informed of Langston Hughes' death.
   (C) For her, Langston Hughes had never really existed.
   (D) Langston Hughes is still present in her thoughts.

9. The word "undiminished" in line 13 is closest in meaning to which of the following?
   (A) Not exaggerated
   (B) Not lessened
   (C) Disappointed
   (D) Unequaled

10. According to the passage, what did Alice Walker give Langston Hughes before he died?
   (A) A job
   (B) An award
   (C) Some oranges
   (D) A stack of books

GO ON TO THE NEXT PAGE

Questions 11-20

Human vision, like that of other primates, has evolved in an arboreal environment. In the dense, complex world of a tropical forest, it is more important to see well than to develop an acute sense of smell. In the course of evolution, members of the primate line have acquired large eyes while the snout has shrunk to give the eye an unimpeded view. Of mammals, only humans and some primates enjoy color vision. The red flag is black to the bull. Horses live in a monochrome world. Light visible to human eyes, however, occupies only a very narrow band in the whole electromagnetic spectrum. Ultraviolet rays are invisible to humans, though ants and honeybees are sensitive to them. Humans have no direct perception of infrared rays, unlike the rattlesnake, which has receptors tuned in to wavelengths longer than 0.7 micron. The world would look eerily different if human eyes were sensitive to infrared radiation. Then, instead of the darkness of night, we would be able to move easily in a strange, shadowless world where objects glowed with varying degrees of intensity. But human eyes excel in other ways. They are, in fact, remarkably discerning in color gradation. The color sensitivity of normal human vision is rarely surpassed even by sophisticated technical devices.

11 What does the passage mainly discuss?

(A) Ultraviolet rays
(B) Human vision
(C) Sight and smell
(D) The environment of primates

12 Why does the author mention the "tropical forest" in line 2?

(A) To explain why primates have developed keen vision
(B) To suggest that primates need to see only the color green
(C) To give an example of environmental change
(D) To indicate where large-eyed primates can be found

13 What does the author mean by stating that "The red flag is black to the bull" (line 5)?

(A) Bulls are attracted to red objects
(B) Bulls do not notice flags
(C) Bulls attack all flags
(D) Bulls do not see the color red

14 The word "monochrome" in line 6 is closest in meaning to which of the following?

(A) Monotonous
(B) Ultraviolet
(C) One dimension
(D) One color

15 In line 8, "them" refers to which of the following?

(A) Human eyes
(B) Ultraviolet rays
(C) Humans
(D) Wavelengths

16 The word "eerily" in line 10 is closest in meaning to which of the following?

(A) Strangely
(B) Increasingly
(C) Slightly
(D) Superficially

17 It can be inferred from the passage that humans could move more easily at night if they

(A) had a narrower field of vision
(B) were color-blind
(C) had infrared vision
(D) lived in an arboreal environment

GO ON TO THE NEXT PAGE

18. The word "surpassed" in line 14 is closest in meaning to which of the following?

(A) Recorded
(B) Exceeded
(C) Found
(D) Provided

19. According to the passage, the ability of humans to distinguish color differences is

(A) average
(B) weak
(C) excellent
(D) variable

20. Where in the passage does the author mention the development over time of certain physical changes among primates?

(A) Lines 3-4
(B) Lines 5-6
(C) Lines 7-9
(D) Lines 12-14

GO ON TO THE NEXT PAGE

Questions 21-30

Ancient people made clay pottery because they needed it for their survival. They used the pots they made for cooking, storing food, and carrying things from place to place. Pottery was so important to early cultures that scientists now study it to learn more about ancient civilizations. The more advanced the pottery in terms of decoration, materials, glazes, and manufacture, the more advanced the culture itself.

The artisan who makes pottery in North America today utilizes his or her skill and imagination to create items that are beautiful as well as functional, transforming something ordinary into something special and unique.

The potter uses one of the earth's most basic materials, clay. Clay can be found almost anywhere. Good pottery clay must be free from all small stones and other hard materials that would make the potting process difficult. Most North American artisan-potters now purchase commercially processed clay, but some find the clay they need right in the earth, close to where they work.

The most important tools potters use are their own hands; however, they also use wire loop tools, wooden modeling tools, plain wire, and sponges. Plain wire is used to cut away the finished pot from its base on the potter's wheel.

After a finished pot is dried of all its moisture in the open air, it is placed in a kiln and fired. The first firing hardens the pottery, and it is then ready to be glazed and fired again.

For areas where they do not want any glaze, such as the bottom of the pot, artisans paint on melted wax that will later burn off in the kiln. They then pour on the liquid glaze and let it run over the clay surface, making any kind of decorative pattern that they want.

21. What does the passage mainly discuss?

(A) Different kinds of clay
(B) The training of an artisan
(C) The making of pottery
(D) Crafts of ancient civilizations

22. Which of the following is NOT mentioned in the passage as a way that ancient people used pottery?

(A) To hold food
(B) To wash clothes
(C) To cook
(D) To transport objects

23. The word "it" in line 3 refers to

(A) clay
(B) culture
(C) survival
(D) pottery

24. According to the passage, which of the following can be learned about an ancient civilization by examining its pottery?

(A) Its food preferences
(B) Its developmental stage
(C) Its geographic location
(D) Its population

25. The word "functional" in line 7 is closest in meaning to which of the following?

(A) Useful
(B) Strong
(C) Inexpensive
(D) Original

26. The word "basic" in line 9 is closest in meaning to which of the following?

(A) Familiar
(B) Fundamental
(C) Versatile
(D) Dirty

27. According to the passage, how do most North American potters today get the clay they need?

(A) They buy it
(B) They make it
(C) They dig it from the earth
(D) They barter for it

GO ON TO THE NEXT PAGE

3 • 3 • 3 • 3 • 3 • 3 • 3

28. It can be inferred from the passage that clay is processed commercially in order to

(A) make it dry more evenly
(B) remove hard substances
(C) prevent the glaze from sticking
(D) make it easier to color

29. According to the author, what do potters use to remove the pot from the wheel?

(A) Melted wax
(B) A wire loop
(C) A sponge
(D) Plain wire

30. The word "pattern" in line 21 is closest in meaning to which of the following?

(A) Model
(B) Color
(C) Puzzle
(D) Design

GO ON TO THE NEXT PAGE

Questions 31-40

The status of women in colonial North America has been well studied and described and can be briefly summarized. Throughout the colonial period there was a marked shortage of women, which varied with the regions and was always greatest in the frontier areas. This
Line favorable ratio enhanced women's status and position and allowed them to pursue different
(5) careers. The Puritans, the religious sect that dominated the early British colonies in North America, regarded idleness as a sin, and believed that life in an underdeveloped country made it absolutely necessary that each member of the community perform an economic function. Thus work for women, married or single, was not only approved, it was regarded as a civic duty. Puritan town councils expected widows and unattached women to be self-supporting and
(10) for a long time provided needy spinsters with parcels of land. There was no social sanction against married women working; on the contrary, wives were expected to help their husbands in their trade and won social approval for doing extra work in or out of the home. Needy children, girls as well as boys, were indentured or apprenticed and were expected to work for their keep

(15) The vast majority of women worked within their homes, where their labor produced most articles needed for the family The entire colonial production of cloth and clothing and partially that of shoes was in the hands of women. In addition to these occupations, women were found in many different kinds of employment. They were butchers, silversmiths, gunsmiths, upholsterers They ran mills, plantations, tanyards, shipyards and every kind of
(20) shop, tavern, and boardinghouse. They were gatekeepers, jail keepers, sextons, journalists, printers, apothecaries, midwives, nurses, and teachers

31 What does the passage mainly discuss?

(A) Colonial marriages
(B) The Puritan religion
(C) Colonial women's employment
(D) Education in the colonies

32 The word "marked" in line 2 is closest in meaning to

(A) underlined
(B) graded
(C) prolonged
(D) distinct

33 According to the passage where in colonial North America were there the fewest women?

(A) Puritan communities
(B) Seaports
(C) Frontier settlements
(D) Capital cities

34 The word "enhanced" in line 4 is closest in meaning to which of the following?

(A) Supplemented
(B) Confirmed
(C) Improved
(D) Determined

35 It can be inferred from the passage that the Puritans were

(A) uneducated
(B) hardworking
(C) generous
(D) wealthy

36 According to the passage, Puritans believed that an unmarried adult woman should be

(A) financially responsible for herself
(B) returned to England
(C) supported by her family
(D) trained to be a nurse

37 The phrase "unattached women" in line 9 is closest in meaning to which of the following?

(A) Women without high social status
(B) Women without property
(C) Unmarried women
(D) Unemployed women

GO ON TO THE NEXT PAGE

38. According to the passage, what did the Puritans expect from married women?

(A) They should adopt needy children.
(B) They should assist in their husbands' trade or business.
(C) They should work only within their own homes.
(D) They should be apprenticed.

39. According to the passage, which products were made entirely by women?

(A) Gunpowder and bullets
(B) Cups and plates
(C) Paper and books
(D) Cloth and clothing

40. The lists in lines 18-21 are intended to show which of the following?

(A) The influence of the Puritans in the colonies
(B) The limits of job opportunities in the colonies
(C) The main industries of the colonial economy
(D) The variety of work done by colonial women

GO ON TO THE NEXT PAGE

Questions 41-49

Beneath the deep oceans that cover two-thirds of the Earth are concealed some of the most tantalizing secrets of our planet. There the crust of the Earth is thinner and the unknown mantle — the layer beneath the crust — lies closest, tempting scientists to drill into it. The first
Line (5) such attempt, the ambitious Project Mohole, got under way during the 1960's and proved the value of deep-sea drilling by making several test holes in the mantle beneath the crust before spiraling costs led to its cancellation.

Soon afterward, however, work began on the more modest Deep Sea Drilling Project, which is not aimed at reaching the mantle but at exploring the crust itself. This venture uses a special ship, the Glomar Challenger, which can be held precisely in position in the
(10) sea -- without any anchor -- by sound-wave guiding systems and computer-controlled propellers. From this stable platform, scientists lowered drilling pipes into waters four miles deep to scoop up cores of ocean sediment and bedrock. Analysis of the fossil contents has indicated that the ocean floors spread, moving continents around the Earth.

41. The passage mainly discusses
    (A) analysis of fossils in the ocean
    (B) exploration beneath the ocean bottom
    (C) the composition of the Earth's crust
    (D) the construction of the Glomar Challenger

42. According to the passage, one of the objectives of Project Mohole was to
    (A) increase public support for underwater experimentation
    (B) test the ocean bottom for unusual ocean sediment
    (C) estimate the age of the Earth's crust
    (D) study the Earth's mantle

43. The word "spiraling" in line 6 is closest in meaning to which of the following?
    (A) Rising
    (B) Necessary
    (C) Unpredictable
    (D) Circular

44. It can be inferred from the passage that Project Mohole originally was intended to
    (A) involve deeper drilling than the Deep Sea Drilling Project
    (B) cost less than the Deep Sea Drilling Project
    (C) employ fewer scientists than the Deep Sea Drilling Project
    (D) yield more fossil discoveries than the Deep Sea Drilling Project

45. The expression "more modest" in line 7 is closest in meaning to
    (A) more sophisticated
    (B) more timid
    (C) less ambitious
    (D) less controversial

46. The word "precisely" in line 9 is closest in meaning to which of the following?
    (A) Exactly
    (B) Clearly
    (C) Economically
    (D) Practically

GO ON TO THE NEXT PAGE

3 • 3 • 3 • 3 • 3 • 3 • 3

47. According to the passage, computers are used on the Glomar Challenger in order to
    (A) measure the spread of the ocean floors
    (B) lower its drilling pipes into the water
    (C) keep it in one place
    (D) detect the location of the Earth's mantle

48. The phrase "stable platform" in line 11 refers to
    (A) the Glomar Challenger
    (B) a ship's anchor
    (C) sound-wave guiding systems
    (D) computer-controlled propellers

49. For which of the following terms does the author supply a definition?
    (A) "mantle" (line 3)
    (B) "anchor" (line 10)
    (C) "sound-wave guiding systems" (line 10)
    (D) "bedrock" (line 12)

GO ON TO THE NEXT PAGE

Questions 50-60

Until fairly recently, the scientist was apt to be viewed as a high-minded seeker after truth, like Sinclair Lewis' Martin Arrowsmith, or as a madman, like Mary Shelley's Dr. Frankenstein — eccentric, maybe even dangerous, but at least pure. Some of the same
Line (5) stereotypes are evident in the 1965 publication *The Scientist*, which pictures the typical scientist as "an august scholar, a remote ascetic, a bright-eyed visionary, or a sweat-soaked mechanic." One thing the scientist was not was a social animal, stitched into the fabric of the workaday world and determined to earn recognition.

An indication that something was wrong with this picture came in 1968, when James D. Watson published *The Double Helix*, an opinionated account of the discovery of the structure (10) of DNA, which won a Nobel Prize for him, Francis Crick, and Maurice Wilkins. When not holding forth on nucleotides or hydrogen bonds, Watson let on that he thought Crick talked too much and laughed too loudly; that another key figure in the story, the x-ray crystallographer Rosalind Franklin, was secretive and stubborn; that he, Watson, wanted very much to be famous; and that he and Crick had always been acutely aware of the chemist (15) Linus Pauling's progress toward the same goal they were pursuing -- that it had been a real race animated by a real prize. *The Double Helix* tweaked the public's imagination. The pursuit of scientific knowledge, it suggested, was a dramatic enterprise, peopled with colorful characters and shaped by their relationships with one another.

50 What is the writer's main point?

(A) Scientists are scholarly visionaries who seek the truth.
(B) The public's image of scientists has changed recently
(C) Most scientists want to be famous
(D) Many scientists pursue similar goals

51 Why is Dr. Frankenstein mentioned in lines 2-3?

(A) Because he was a sweat-soaked mechanic
(B) To show how scientists differ from nonscientists
(C) As an example of a mad scientist
(D) To demonstrate that science is a dramatic enterprise

52 The word "evident" in line 4 is closest in meaning to which of the following?

(A) Factual
(B) Apparent
(C) Presented
(D) Repeated

53 The phrase "stitched into the fabric of" in line 6 is closest in meaning to which of the following?

(A) An integral part of
(B) Trying to repair
(C) Feeling isolated from
(D) Escaping from

54 The author implies that James D. Watson was all of the following EXCEPT

(A) opinionated
(B) secretive
(C) knowledgeable
(D) ambitious

55 *The Scientist* and *The Double Helix* are compared in the passage because they

(A) present contradictory images of scientists
(B) deal with different scientific achievements
(C) describe different famous scientists
(D) show various ways scientists earn recognition

63

GO ON TO THE NEXT PAGE

30

56. The word "key" in line 12 is closest in meaning to which of the following?

(A) Respected
(B) Major
(C) Difficult
(D) Famous

57. Which of the following can be inferred from the passage about Rosalind Franklin?

(A) She helped Watson write *The Double Helix*.
(B) She worked closely with Watson, Crick, and Wilkins.
(C) She was not interested in a prize.
(D) She was Watson's supervisor.

58. The phrase "the same goal" in line 15 refers to

(A) writing a book
(B) working together with Rosalind Franklin
(C) discovering the structure of DNA
(D) developing relationships with other scientists

59. The word "enterprise" in line 17 is closest in meaning to which of the following?

(A) Story
(B) Undertaking
(C) Business organization
(D) Reward

60. Where in the passage does the author refer to fictional characters?

(A) Lines 1-3
(B) Lines 6-7
(C) Lines 8-10
(D) Lines 16-18

THIS IS THE END OF SECTION 3

IF YOU FINISH BEFORE TIME IS CALLED, CHECK YOUR WORK ON SECTION 3 ONLY.
DO NOT READ OR WORK ON ANY OTHER SECTION OF THE TEST.

31

Appendix C

Confirmation Letter Sent to Participating Institutions and Instructions for Supervisors Administering the Experimental Section 3

EDUCATIONAL TESTING SERVICE PRINCETON, N.J. 08541

609 921 9000
CABLE EDUCTESTSVC
FAX 609 734 5410

February 4, 1992

Dear Colleague:

Thank you for agreeing to help us with our research study. As was mentioned in our telephone conversation, the TOEFL program is considering making revisions to the TOEFL test. To ensure the reliability of ETS tests, new material is field-tested with students whose backgrounds are similar to those expected to take the tests. Please let us know how many students will be tested by the end of February.

This pilot test version of the TOEFL includes typical Sections 1 and 2 and a new version of Section 3 with no vocabulary subpart. Vocabulary items are incorporated into the reading passage sets. The administration of the test will take about two hours. The test may be given in a language laboratory or in a conventional classroom if a tape or cassette recorder with a good speaker is available.

The research design involves the students taking the pilot test version approximately one week before the Institutional TOEFL. It will be necessary for the students to be assigned identification numbers which they will put on both the pilot test answer sheet and the Institutional TOEFL answer sheet so we can link their performances on these two components. Details on student identification will be shipped with the test materials. All pilot test and Institutional TOEFL materials should be returned to ETS as soon as possible after administration. We will not issue scores on the pilot test questions; however, we will provide you with a scoring stencil if you wish to grade your students on a percent correct basis.

An order form will be included with your test shipment stating that the Institutional TOEFL will be free of charge for the number of students participating in the research study, up to a limit of 100. Pilot testing materials will be shipped together with the Institutional TOEFL materials approximately two weeks before the testing date.

I would appreciate your sharing this letter with any colleagues at other institutions who are teaching a course in English as a Second Language and might be interested in this opportunity. If you have any questions about the research study, feel free to call me collect. Thank you in advance for your assistance. We are looking forward to working with you on this project.

Sincerely yours,

Barbara K. Suomi
Examiner
School and Higher
Education Programs

BKS/ly

Instructions for Supervisors of the English Proficiency Pretest (EPP) Pilot Study

Thank you again for agreeing to participate in the English Proficiency Pretest (EPP) pilot study. As explained in the confirmation letter sent to you previously, this pilot version of the TOEFL test includes typical Sections 1 and 2 and a new version of Section 3 with no vocabulary subpart. The EPP will take about two hours to administer. It should be administered approximately one week before the Institutional TOEFL. In exchange for participation in the EPP pilot study, up to 100 Institutional TOEFL tests may be administered free of charge. All materials necessary for administering the EPP are enclosed, as are the materials necessary for administering the Institutional TOEFL. The instructions for administering the EPP are generally the same as those stated in the TOEFL Institutional Testing Program's Manual for Supervisors. Additional special instructions for administering the EPP are provided below.

1) Identification Numbers. The research design for this study depends upon having the same students take both the EPP and the Institutional TOEFL. To facilitate this, we have attached a list providing six digit identification numbers. **Please assign one of these numbers to each student in your institution who participates in the EPP pilot study. Write their names in the corresponding spaces on the attached list, and please make sure that each student marks in the same sequence number on both the EPP answer sheet and the Institutional TOEFL answer sheet.** The attached list should be completed and returned to ETS with the other testing materials. In addition, please encourage your students to mark in their names on the EPP answer sheet. This will provide a second source of information for matching the EPP and Institutional records. Participant names will not be used for any purpose except matching test records.

2) Sections 1 and 2. Sections 1 and 2 of the EPP pilot should be administered with the time limits and general instructions provided in the TOEFL Institutional Program's Manual for Supervisors.

3) Timing Section 3. A crucial factor in this study is the amount of time provided to students for completing Section 3 of the EPP. **IMPORTANT: Please note that for your institution, students should be given 55 minutes for completing Section 3 of the EPP.** This time limit should replace the time limit of 45 minutes provided for Section 3 in the Manual for Supervisors.

4) Section 3 Versions. There are three versions of Section 3 in the EPP pilot study: Form T2000A contains 60 questions, Form T2000B contains 54 questions, and Form T2000C contains 48 questions. By design, different students at your institution will take Section 3 tests of different lengths. If the students bring this to your attention, please inform them that this is intentional, and that they are to

attempt to answer all questions in their particular booklet as if they were taking an actual TOEFL test.

5) Scoring the EPP. ETS will not be scoring the EPP. However, three scoring stencils are included with the EPP materials in case you wish to provide your students with feedback on their performance. The three scoring stencils correspond to the three versions of Section 3 on the EPP. If you use the stencils, please be sure to match the answer sheets with the appropriate stencils. Score the EPP tests as soon as possible so that there will be no delay in returning materials to ETS.

6) Checking EPP Materials. At the end of the EPP test, collect the test books and answer sheets from each student. **Please check all answer sheets to make sure the students have gridded the appropriate identification numbers on their answer sheets.** Students should grid this same number on their answer sheets when they take the Institutional TOEFL approximately one week later. The EPP materials should held until the Institutional TOEFL administration is completed, at which point both the EPP and Institutional TOEFL materials should be returned to ETS.

EPP Pilot Study Student Identification List

ID #	Student Name	ID #	Student Name
214001		214026	
214002		214027	
214003		214028	
214004		214029	
214005		214030	
214006		214031	
214007		214032	
214008		214033	
214009		214034	
214010		214035	
214011		214036	
214012		214037	
214013		214038	
214014		214039	
214015		214040	
214016		214041	
214017		214042	
214018		214043	
214019		214044	
214020		214045	
214021		214046	
214022		214047	
214023		214048	
214024		214049	
214025		214050	

Printed on Recycled Paper

57909 37569 • Y25M.5 • 275591 • Printed In U.S.A.