DOCUMENT RESUME

ED 385 568 TM 024 006

AUTHOR	DeMauro, Gerald E.
TITLE	An Investigation of the Appropriateness of the TOEFL Test as a Matching Variable To Equate TWE Topics.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-92-26; TOEFL-RR-37
PUB DATE	May 92
NOTE	44p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*English (Second Language); *Equated Scores; *Evaluation Methods; Test Format; *Test Use
IDENTIFIERS	*Anchor Tests; Appropriateness Measurement; Equipercentile Equating; Essay Topics; Linear Equating Method; *Test of English as a Foreign Language; Test of Written English; Writing Prompts

ABSTRACT

The feasibility of using linear and equipercentile equating methods (W. H. Angoff, 1984) to equate forms of the Test of Written English (TWE) by using the Test of English as a Foreign Language (TOEFL) as an anchor was explored. These two equating methods assume that either the TOEFL test and TWE test measure the same skills or that the examinee groups across TWE administrations are equivalent in skills. The differences between equated and observed scores (equating residuals) and differences among the mean equated scores for examinee groups were further examined in terms of characteristics of the TWE topics. An evaluation of the assumptions underlying the equating methods suggests that the TOEFL and TWE tests do not measure the same skills and that examinee groups are often dissimilar in skills. Therefore, use of the TOEFL test as an anchor to equate the TWE tests does not appear appropriate. An alternative equating model based on expert judgment during pretest evaluation of potential essay prompts is recommended for future investigation. Eleven tables present study data, and an appendix provides information about the sample. (Contains 14 references.) (Author/SLD)

TEST OF ENGLISH AS A FOREIGN LANGUAGE

Research Reports

REPORT 37
MAY 1992

An Investigation of the Appropriateness of the TOEFL Test as a Matching Variable to Equate TWE Topics

Gerald E. DeMauro

EDUCATIONAL TESTING SERVICE

An Investigation of the Appropriateness of the TOEFL Test as a Matching Variable to Equate TWE Topics

Gerald E. DeMauro

Educational Testing Service
Princeton, New Jersey

RR-92-26

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

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

✣ ✣ ✣

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

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

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

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

Abstract

The feasibility of using linear and equipercentile equating methods (Angoff, 1984) to equate forms of the Test of Written English (TWE) by using the TOEFL test as an anchor was explored. These two equating methods assume that either the TOEFL test and TWE test measure the same skills or that the examinee groups across TWE administrations are equivalent in skills. The differences between equated and observed scores (equating residuals) and differences among the mean equated scores for examinee groups were further examined in terms of characteristics of the TWE topics.

An evaluation of the assumptions underlying the equating methods suggests that the TOEFL and TWE tests do not measure the same skills and that examinee groups are often dissimilar in skills. Therefore, use of the TOEFL test as an anchor to equate TWE tests does not appear appropriate. An alternative equating model based on expert judgment during pretest evaluation of potential essay prompts is recommended for future investigation.

Table of Contents

List of Tables

Introduction

Overview of the Task

The Test of Written English (TWE) is a one-topic essay test of English writing skills that is administered to foreign populations four times a year with the multiple-choice Test of English as a Foreign Language (TOEFL test). The TWE test uses different topic types. For example, the compare/contrast type presents a point of view and an opposing point of view, sometimes implied. Examinees are required to develop arguments in favor of one and/or in opposition to the alternative. The chart/graph type presents data in the form of a chart or graph. Examinees are required to draw inferences from the data and develop cogent arguments based on these inferences. Other topic types are used as well, but these two are the only types used in the TWE forms examined in this study.

The TWE test is holistically scored, and scores are reported on a scale ranging from 1.0 to 6.0 with half-point intervals. All TWE topics are pretested on populations of limited-English proficient undergraduate and graduate students in the United States and Canada who have English skills that are similar to those of the examinees who normally take the TWE and the TOEFL test.

The TOEFL test, described in greater detail later, provides three section scores and an overall score. Section 2 is of particular interest to this study because it measures structure and written expression. The TWE exam was designed as a direct test of standard written English to complement the indirect measure provided by Section 2. In particular, the TWE essay test was conceived as a measure of productive skills, as distinguished from the recognition skills evaluated by multiple-choice questions (Angelis, 1982).

Rationales for an Equating Study of the TWE Test

Corporate guidelines of Educational Testing Service (ETS) require that programs assure "adequate comparability of score on different forms" of essay tests (Breland, Conlan, Fowles, and Livingston, 1987; Guideline 11ii, Writing Test Specifications). Investigations of equating methods and applications are also consistent with priorities endorsed by the TOEFL Research Committee for the TWE (Stansfield & Ross, 1988), and with the ETS Standards for Quality and Fairness (Educational Testing Service, 1987).

Golub-Smith, Reese, and Steinhaus (1991) recently investigated the impact of topic variations in TWE scores, but there is little available research on how examinees with different levels of writing skill are affected by topic variations. Equating methodologies may help assess the differential impact of topic variations.

Focus of the Study

This study proposes to equate the TWE scores derived from the November 1986, May 1987, July 1987, November 1987, May 1988, September 1988, and October 1988 administrations of the TWE test to the scale of the July 1986 TWE administration. For five of these seven test administrations, the world was divided into three TOEFL-TWE administrative regions (A, B, and C), and each region received a different essay topic. There were worldwide administrations of a single chart/graph type topic in July 1987 and of a single compare/contrast type topic in September 1988. Because different topics were used for most of the administrations in each of the three regions of the world, analyses and equating procedures are done within region.

Statement of Research Problem

This study addresses three major questions:

1. Based on the assumptions of the equating models, is one section of the TOEFL test preferred over other sections as a matching variable for TWE equating?

2. Is the relationship between the TWE and TOEFL test scores stable enough across test administrations and TWE topic types to support the use of one of the equating methods examined in this study?

3. If the assumptions of the equating methods are met, are certain topic types more difficult than others, as indicated by differences in equated scores, or are certain topics more difficult than the July 1986 topic, as indicated by differences between observed and equated scores?

If the equating procedures under investigation are appropriate, it may be possible to assess the relative difficulty of different topics by evaluating the differences between scores that have all been converted to a common scale. It may also be possible to evaluate how different each essay topic is in difficulty from the July 1986 topic by examining the difference between observed and equated scores (equating residuals). That is, since readers award the same point values for the same level of demonstrated skill, regardless of the essay topic addressed, differences between the observed score and the equated score could either indicate that the assumptions that support the equating procedures are not met or that the topic on which the observed score is awarded differs in difficulty from the July 1986 topic, on which the equated score is based. If the equating procedures seem appropriate, there is greater confidence that the equating residuals are an index of differences in topic difficulty compared to the July 1986 topic.

Methods of Equating

If randomly assigned examinees take two forms of a test that measure the same skill, then the examinees should perform equally well on the forms provided the forms are equally difficult measures of that skill. Any differences in observed scores for these equivalent examinee groups would then reflect variation in the difficulty of the test forms, which could be adjusted through equating procedures. As discussed later, it cannot be assumed that the examinees who take different forms of the TWE have the same level of ability, nor can it be assumed that the groups that take different forms of the TWE are equally able, on average, and exhibit the same distribution of ability, because they have not been randomly assigned to test forms. Therefore, some common measure of their ability, called an anchor test, must be used to gauge their skills or match them in ability. This is the suggested use of the TOEFL test in this study.

Both linear and equipercentile equating methods are explored in this study.[1] These methods are derived from Angoff's Design IV (1984). This design does not require that the groups that take different forms of the TWE are randomly assigned to the forms given in those administrations, but it does assume that the groups are not widely different in skills. There is some evidence that the groups that take different forms of the test actually differ in skills, however, calling into question the use of these methods. The national composition of the examinee samples varies from administration to administration and, consequently, from topic to topic (C. Taylor, personal communication, 1991), and nationality may well be related to differences in English writing skills. Observed differences in skill will be examined in the study as part of the investigation of the appropriateness of the equating models.

Why the July 1986 Scale Was Chosen as a Standard

As mentioned earlier, the scores from all forms considered in this study are transformed to the TWE scale based on the July 1986 form. This form was chosen as a standard for two reasons.

1. It is the first operational form of the TWE.

2. When different forms of the TWE are read and scored, reading managers first review responses to this form that exemplify different scores. The managers then select the papers from the forms to be scored that also exemplify the qualities associated with different TWE scores and use these papers to train readers.

[1] IRT methods (Masters, 1982; Phillips, 1989) are not currently possible because they depend on having either some examinees in common to both test forms or some essay topic in common to all examinees (Wright & Stone, 1979).

Methods

Study Questions

Question 1: Based on the assumptions of the equating models, is one section of the TOEFL test preferred over other sections as a matching variable for equating forms of the TWE test? Both the linear and equipercentile equating methods under investigation, from Angoff's Design IV, require that if the forms of the TWE test to be equated are not parallel in function to the anchor test, then the groups that take each form are equivalent in skills (Angoff, 1984). The linear procedures are sometimes called "Tucker" equating. The equipercentile analog of Tucker linear equating is based on distributions derived from frequency estimation (Angoff, 1984).

The specific linear equating method examined in this study was derived from the work of L. R. Tucker (Angoff, 1984). The procedures assume that within administrative region (A, B, or C), the regression slopes and intercepts of the TWE test on the TOEFL test and the variance error of estimates of the TWE test from the TOEFL test were the same for the examinees who took each form of the TWE test (these values can be observed) as for a composite or synthetic group of all of the examinees who took either form of the test (these values cannot be observed). The equipercentile method assumed that the joint distributions of TOEFL and TWE test scores for the groups that took each form of the TWE test were the same as the joint distributions for this synthetic group.

Frequency estimation combined the frequencies of each administration's sample with that of the July 1986 sample for each TWE score at each level of the matching variable, TOEFL total score. The frequency attaining each TOEFL score in each test form was adjusted to be the same as this combined frequency. These adjusted frequencies were then summed over TOEFL scores to produce two TWE distributions for the synthetic group, one on the July 1986 form and the other on the test form that was to be equated to the July 1986 form. Equipercentile equating was then applied to these frequencies. In the current application, whenever there were no examinees or only small numbers of examinees who achieved a certain TOEFL score, the frequencies at that score were combined with that of the next score closest to the middle of the TOEFL score distribution, following a process detailed by Angoff (1984).

As a first step in evaluating the assumptions underlying the equating models, correlational and regression analyses were made of the relationship of TOEFL scores to TWE scores to assess whether the test forms were each parallel to the anchor test and to identify the TOEFL score that would be the most appropriate matching variable for equating purposes. The means and standard deviations on the TOEFL test sections that had the highest correlations with the TWE were also examined to assess the comparability of skills in different examinee samples. Unfortunately, if the groups that took each form of the test had had identical TOEFL means and standard deviations, it would not have guaranteed that the groups were equivalently skilled on the TWE test, particularly if the TOEFL test is not parallel in function to the TWE test. Because there was no other common measure of the groups, however, our analysis of the equivalence of examinee group skills focused on performance on the related skills measured by the TOEFL test.

This analysis examined both the magnitude of mean differences in the scores used in equating and differences in dispersion of these scores for each examinee sample compared to the July 1986 sample. These evaluations employed criteria used in other equating studies (Modu & Stern, 1975) to evaluate the appropriateness of the Tucker procedures. Because the same assumptions of parallel anchor test or groups' skills equivalence apply to the curvilinear methods, the criteria were used to assess the appropriateness of both the linear and equipercentile models employed in this study.

These criteria are that groups that take different forms of the TWE have standardized TOEFL mean differences (mean differences divided by the combined standard deviation) of no greater than .25 and that the ratios of TOEFL variances of the groups fall in the .80 to 1.25 range. If these criteria are not met, then the preferred linear procedure is Levine's (1956) true score equating, as opposed to the Tucker method. As a second step in the assessment of the assumptions underlying equating, then, the Modu and Stern criteria were applied to the TWE distributions under examination to determine the appropriateness of the Tucker method. Note that Levine's procedures require estimates of both TWE and TOEFL reliabilities within administrative regions. Since these estimates were not available, Levine equating was not a reasonable alternative for Tucker equating.

Analyses were made to complement the correlation analyses described above in discerning which, if any, components could provide the most accurate means of gauging examinee writing skills, either as possible anchors for linear equating, or as a means of matching examinees for frequency estimation equating. To assess this, biserial correlations of TWE score and TOEFL item scores were computed for the two compare/contrast (July 1986 and May 1987) and the two chart/graph topics (July 1987 and May 1988) discussed earlier. The biserial coefficients were then classified according to whether they were below .30, from .30 to .39, or above .40. Chi-square analyses evaluated whether the distributions of biserial values in these categories varied by section of the TOEFL test for any of the three administrative regions.

Question 2: Is the relationship between the TWE and TOEFL test scores stable enough across test administrations and TWE topic types to support the use of one of the equating methods examined in this study? General linear model analyses of variance were made of the regression of the TWE scores on each TOEFL score in relation to administrative region, for each of four administrations of the TWE: July 1986, May 1987, July 1987, and May 1988. In these analyses, the TOEFL score was the dependent variable and the TWE score (continuous) and administrative region (discrete) were independent variables. The interaction of the independent variables estimated the homogeneity of the regression of TWE test scores on TOEFL test scores, thereby enabling an evaluation of the stability of the relationship across the topic variations of the administrative regions. Because there were no regional topic variations in July 1986 or in July 1987, if we assume that the samples were as different in skills from region to region in these administrations as they were in the May administrations, then the differences in the magnitude of the interaction effects between the two July administrations and the two May administrations should be related to regional variation in topics.

Question 3: If the assumptions of the equating methods are met, are certain topic types more difficult than others, as indicated by a difference between observed and equated scores? Means and standard deviations of the converted scores were prepared for each administration and each administrative region. These enabled comparison of performance on these topics on the common July 1986 scale. As discussed earlier, equating residuals were also computed, in which the observed score is subtracted from the equated score, to assess differences in difficulty of each topic with the July 1986 topic.

Sample

Joint distributions of TWE and TOEFL scores were prepared separately for each of the three worldwide TWE administrative regions for the July 1986, November 1986, May 1987, July 1987, May 1988, September 1988 and October 1988 TWE administrations. Within each region, as part of the equipercentile procedure for each TOEFL score, a distribution of the TWE scores for the July 1986 administration was matched to the distribution of the TWE scores of each particular administration under investigation. Table 1 shows the administration dates and examinee sample sizes by administrative region in which different TWE topics were administered in November 1986, May 1987, November 1987, May 1988, and October 1988. In this way, the scores derived from the topics used in each region for each administration were separately equated to the scale of that region for the July 1986 form.

The Examinations

TWE test. The TWE test is a direct measure of writing designed to complement Section 2 of the TOEFL test, which is an indirect measure of English writing (Educational Testing Service, 1990). The TWE is a 30-minute essay test that is scored holistically by two readers on a scale of 1 to 6. If the readers disagree by more than one point, an experienced third reader is used to adjudicate the score. Because averaging introduces half points, the reported scale score has 11 possible values.

TOEFL test. The TOEFL test is a three-section multiple choice test that reports four scaled scores, including one for each section and a total score. Section 1, Listening Comprehension, assesses listening comprehension of statements, of conversations, and of oral presentations. Section 2, Structure and Written Expression, assesses understanding of basic grammar and knowledge of the grammar of written English. Because Section 2 is an indirect test of writing, it is of particular interest to this study. Section 3, Vocabulary and Reading Comprehension assesses word phrase knowledge and understanding of written passages (Educational Testing Service, 1990). The total score is the sum of the section scale scores multiplied by 10/3. Throughout the study, the TOEFL scores referred to are scaled scores.

Results

Question 1: Adequacy of TOEFL Test as a Matching Criterion

Table 2 presents the correlation coefficients of TWE and TOEFL scores, showing some degree of linear prediction, especially for the TOEFL total score. The ranges of the correlations were wide, from .549 to .731 for Section 2 and from .613 to .752 for TOEFL total scores. Moreover, the pattern of relationships varied from administration to administration. For example, although the total score had the highest correlation with TWE scores in all administrations as expected, Section 3 scores had the next highest correlations with TWE scores for the May 1987 and October 1988 administrations, and the lowest correlations for the July 1986, July 1987, and September 1988 administrations.

Because reliability was not estimated either for the TOEFL test or for the TWE within administrative region, it was not possible to disattenuate these correlations. Even if the TWE and TOEFL test reliabilities within administrative regions were each as low as .800 (this is not likely), a disattenuated correlation of .900 between TWE and TOEFL scores would require TWE correlations with TOEFL scores to be .720 or higher. The magnitude of the observed correlations between TOEFL scores and TWE scores did not permit us to assume that the requirements of the equating models that these two tests are parallel was satisfied.

Section scores. Table 2 shows that the TOEFL total score provided about as much information about the writing skills measured by the TWE test the optimally weighted combination one could derive from multivariate regression of the TOEFL section scores on TWE scores. In fact, in each of the administrations, the correlation of the TOEFL total scale score with TWE scores was within .10 of the multiple correlation of the three TOEFL section scores. The higher correlation of the TOEFL total score with the TWE scores supported using the total score as the best choice among possible TOEFL scores as a matching variable or anchor test for equating.

Perhaps most perplexing was the variation in Region A from administration to administration in which one of the section scores was the best predictor of TWE performance. This suggests that one section score should not be chosen as a matching variable without first examining the data.

Comparability of examinee groups (see Tables 3 and 4). Table 3 shows the TOEFL and TWE score means and standard deviations of the examinee samples. There is a tendency to attract similarly skilled examinees, at least in the skills measured by the TOEFL total score, at similar times of the year. For example, there were very small total score mean differences between the July 1986 (505.92) and July 1987 (505.43) administrations; the November 1986 (525.85), November 1987 (533.33), and October 1988 (526.90) administrations; and the May 1987 (518.11) and May 1988 (517.68) administrations. These groups, which were so similar in TOEFL total scores, were very different in TWE scores. Perhaps the TWE test is more sensitive to a certain skill than the TOEFL test, or the differences in difficulty of the TWE forms are so large even similarly-skilled examinee groups achieve different mean scores.

Ratios of TOEFL score variance of examinees who took the July 1986 form of the TWE to the variances of examinees who took each of the other forms of the TWE fell within the .80 to 1.25 interval, showing groups taking different forms were similarly distributed on the TOEFL test. However, the second criterion, that the groups have standardized mean TOEFL differences (July 1986 to other administrations) no greater than .25 was violated. In Region B, standardized mean differences on Section 2 exceeded the second criterion for the November 1986 (.63), May 1987 (.44), November 1987 (.71), May 1988 (.56), and October 1988 (.57) administrations. In Region C, standardized mean differences on Section 2 violated the second criterion for the May 1987 (.33) and September 1988 (.26) administrations.

The .25 criterion of mean differences on the TOEFL total score was violated in Region A for the November 1987 (.34) and October 1988 (.34) administrations. In Region B, the difference criterion was violated in the November 1986 (.71), May 1987 (.51), November 1987 (.85), May 1988 (.54), and October 1988 (.62) administrations (Table 4). These two criteria of group equivalence for Tucker equating, then, were only partially satisfied.

Biserial correlations (Table 5). Overall, Section 2 had higher median biserial values (.36) than Section 3 (.33) or Section 1 (.32). The median biserial coefficients for Region B (.38) were also higher than those for either Region A (.32) or Region C (.31). The item biserial correlations with TWE scores were related to item biserial correlations to TOEFL total scores (Spearman r (n=1752) = .814), showing the items most discriminating for TOEFL test performance were also most discriminating for TWE performance.

For all three administrative regions, a larger proportion of biserial correlations of .40 or higher were found in Section 2 of the TOEFL test. For this reason, Section 2 and the TOEFL total score were both used as anchor tests for equating purposes. These analyses do not, however, permit the assumption that either the total TOEFL test or TOEFL Section 2 is parallel in function to the TWE test.

Summary of TOEFL test adequacy for TWE equating. The correlational analyses failed to demonstrate that the TOEFL and TWE tests were parallel. Moreover, analyses of the comparability of the skills of groups that took each TWE form suggested that the groups that took some forms in some regions were not very similar to the group that took the form administered in July 1986. The equating procedures are presented below with the caution that they may not be appropriate, and a discussion follows examining the effects of violations of the assumptions of the procedures.

Question 2: Stability of the Relationship between the TWE and TOEFL Test Scores

Design IV (Angoff, 1984) assumes that the TWE-to-TOEFL score regression slopes, intercepts, and error variances for each sample are equal to the estimated slopes, intercepts, and error variances for that sample and the July 1986 sample combined. This assumption also applies to the slopes, intercepts, and error variances for July 1986. Table 6 gives the values of the regression slopes and intercepts for the July 1986 and May 1987 (compare/contrast) and July 1987 and May 1988 (chart/graph) forms used in Regions A, B, and C. The slopes appear to be reasonably consistent. The intercepts, however, varied considerably. Note that these statistics could not directly evaluate the assumptions underlying the equating model, because the regression slopes and intercepts for the combined group could not be observed. They do show, however, that the relationship between scores on the TWE and TOEFL tests varied in different administrations. This presents the dilemma that one TOEFL score may predict different TWE scores in different regions or administrations

For the July 1986 and May 1987 compare/contrast topics and the July 1987 and May 1988 chart/graph topics, analyses were made of the homogeneity of regression over the three regions in which the TWE test was administered. In these analyses, the TOEFL total or section scores were the dependent variables, while region and the interaction of region and TWE score were independent variables. The results are shown in Table 7. Note that the regressions of the TOEFL Section 2 and the TOEFL total scores on TWE scores were significantly different over administrative regions. The groups also appear to be of different skill levels for all TOEFL sections in each administration except for Section 3 on the May 1988 administration. Once again, this analysis did not directly test the assumptions underlying the equating models, but it did suggest that the same score on the TWE test would predict different TOEFL scores in different administrative regions. If the skill measured by the TOEFL test were the same as that measured by the TWE, the same TWE score, which is based on the same scoring criteria in all regions, should predict the same TOEFL score.

It is especially interesting that the differences in regression over regions were smaller in July 1986 and July 1987, when the topics did not vary from region to region. It is possible that these smaller differences in the relationships for the two July administrations reflected similarities in the skills of the July 1987 and July 1986 examinee groups. It is also possible that the larger F-ratios associated with the May 1987 and May 1988 administrations, when different topics were used in each region, mean that the relationship between the TOEFL and the TWE scores was more sensitive to topic variations (as there were in the May administrations) than to population variations (as there were in all four administrations). The existence of even these relatively small regional differences in regressions within the two July administrations when there were no regional differences in topics suggests that there are important population differences from region to region.

Question 3: Differential Difficulty of Topics That Met the Equating Criteria

Tables 8 and 9 show the equivalent July 1986 scores that would be attained for each TWE score in each region for each TWE administration. The differences between observed and equated scores were smaller for the frequency estimation conversions than for the linear equating conversions.

Table 10 shows the average converted scores and differences between equated and observed scores for each of the seven TWE administrations in each region across examinees. Perhaps closest to the observed scores were the equated scores for the July 1987 administration. Because the means and variances on the TOEFL test were so close for the two July administrations, there is greater confidence in the appropriateness of the equating models. The lack of regional topic variants in the two July administrations did not in itself seem to affect the differences between observed and equated scores. For example, the observed and equated scores for the September 1988 administration were not nearly as close as they were for the July 1987 administration, even though the topics were the same for each region in both of these administrations.

Equating residuals. Scores associated with the November 1987 and October 1988 administrations had lower equated and observed score differences (Table 10). Table 10 also shows that the July 1987 topic was more difficult than other topics in terms of the mean equated scores. Interestingly, the November 1987 and October 1988 topics were the compare/contrast type. The July 1987 topic also had small linear equating residuals.

In fact, the largest equating residuals were found for the September 1988 administration. The average differences between equated and observed scores were largest in Region A, and were larger for conversions when Section 2 was used as an anchor test or matching variable than when the TOEFL total score was used as an anchor test or matching variable. Also note that the frequency estimation conversions showed smaller residuals than did the linear conversions.

The chart/graph type essays seem more sensitive to the precision of the matching procedures. For example, the equating residuals for May 1987 (chart/graph) in Region A ranged from .36 (TOEFL total score matching, frequency estimation) to .45 (Section 2 matching, linear equating). The correlations of TWE scores with the TOEFL scores range from .58 to .66 (Table 2). In Region B, where the TOEFL/TWE score correlations ranged from .70 to .72, the differences between observed and equated scores range from .07 to .13 for the May 1987 administration. A similar relationship was found for the July 1987 administration which also involved a chart/graph type essay.

Because the same TWE protocol was used by TWE readers at all test administrations, the equating residual can be interpreted as a measure of how different examinee skills were from those described in the scoring protocol. That is, when the residuals are large, the same examinee would be described as having one level of skills at one administration, but with the same paper would be described as having another level of skills at another. These differences, however, may not be related to some inconsistency in how the scoring criteria were applied but, rather, may reflect limitations in the use of the TOEFL test to properly and consistently gauge the examinee abilities measured by the TWE test.

Discussion

Summary and Conclusions

Study questions 1 and 2 are primarily concerned with assessing tne extent to which use of the TOEFL test as an anchor test or matching variable could satisfy the assumptions of the linear and equipercentile equating for TWE topics under Angoff's (1984) Design IV. If this use of the TOEFL test was supported, the differential difficulty of the topics could be assessed by comparing the equated scores achieved by the different examinee groups. The equating models we used require that if the target test (TWE) and anchor test (TOEFL) are not parallel, the populations that take each form of the target test must be equivalently skilled. It appears that it could not be assumed the TWE and TOEFL test were parallel. Moreover, the TOEFL scores of the examinee groups did not provide conclusive evidence about how equivalent the groups were in the skills measured by the TWE test.

Two aspects of the July 1986 administration are noteworthy in explaining observed skill differences between several of the examinee groups and the July 1986 sample. Primarily, the July 1986 examinee group achieved the lowest mean scores on Section 2 of all groups in all regions except Region C, where it was tied for last with the November 1987 group mean. The July 1986 examinee group also achieved the lowest mean TOEFL total score of all examinee groups in all regions except in Region C, where only the July 1987 examinee group achieved a lower mean score. Also, whereas examinees in Region B achieved the lowest TOEFL score means in July 1986, examinees in Region B achieved the highest total score mean on four of the other seven administrations and also the highest Section 2 score mean in four of the other seven administrations. This combination of observations explains some of the observed group differences, particularly those differences in Region B that were large enough, by the standardized mean difference criterion, to preclude the equating assumption that the groups were equivalent in skills to the July 1986 groups.

Among the administrations in which the criteria for equating to July 1986 were met, the chart/graph topic used at the July 1987 administration seemed more difficult in terms of equated scores than the topics used at other administrations (Table 10), and that the chart/graph topics produced lower equated means than did the compare/contrast topics (Table 11). It is not known whether this was a general phenomenon involving this type of topic or was specific to the difficulty of the July 1987 topic, which, along with the Region A and Region C topics for May 1987, comprised all the chart/graph topics examined in the study (the May 1987 Region B topic failed to meet the group standardized mean difference criterion). Note from Table 10 that the mean equated scores for the May 1987 topics that met the group equivalence criteria were quite similar to the means for the compare/contrast topics that met the equating criteria, while the means for the July 1987 topic were considerably lower.

Although the TOEFL test provides an accurate measure of the required English skills of examinees, it may not be an appropriate anchor test or matching variable for equating TWE scores because it is concerned with only part of the unique blend of skills measured by the TWE test. TWE score variation might still be attributed, at least in part, to factors other than English proficiency, some of which are unique to expository writing tasks and not measured by multiple-choice tests (e.g., style and presentation), as well as variation associated with interreader and intrareader reliability.

Well-written TWE essay topics are carefully screened even before they are pretested to assure that they can elicit responses from all examinees. However, variations in the lexical and rhetorical components of the essay topics may contribute considerably to the differences in the correlations between TWE and TOEFL scores.

Identifying the Demands of the Topics

The TWE essay topic administered in July 1987 was given to two experts in topic development for large essay programs at Educational Testing Service. Each was also given the July 1986 topic and the distribution of examinees in terms of the percentage of July 1986 examinees who achieved each of the 11 TWE scaled scores from 1.0 to 6.0. The purpose of this exercise was to help interpret the results of the study by identifying possible components of the essay topics that would affect score distributions. The experts were asked to judge how the July 1986 population of examinees would fare on the July 1987 topics, in either of two ways:

(1) estimate the percentage of July 1986 examinees who would achieve each of the scaled scores;

(2) estimate a cumulative percentage for certain scores for examinees who would achieve that score or lower (e.g. 90% would score 5.0 or lower) and, by subtraction, estimate the percent that would achieve each discrete score.

From these estimates, both equipercentile and linear methods are useful for equating. In fact, neither of the experts completed this assignment (see Appendix), and both cited uncertainty about the effects of the lexical components of the topic as the reason. Their comments are instructive.

Note from Table 2 that the correlations of Section 2 and TOEFL total scores with TWE scores for the July 1987 examinees were relatively small in all regions but Region B, which had only 7% of the examinees. One of the expert reviewers found the July 1987 topic to be "multidimensional," involving skills in drawing empirical inferences. It may well be that sensitivity of the July 1987 TWE topic to other skills suppressed the correlation with TOEFL scores. Note, too, from Table 10 that the July 1987 topic appears to be the most difficult of all. It is perhaps these very skills in drawing empirical inferences that increased the difficulty of the July 1987 topic. Neither expert was able to project distributions on the July 1987 topic, as both said that more would have to be known about the nature of the rhetorical task.

Other Equating Models

Allen and Holland (1989) are developing a model to estimate scores on an optional essay topic for examinees who choose other options. This model first examines how sensitive the results of equating are to different untestable assumptions about the form of the score distributions. Such an approach could apply to the current need to estimate scores for a synthetic group of examinees based on the relationship between scores on the TOEFL and TWE tests.

It might also be possible to augment TOEFL test data on the skills of the examinees with expert judgments about what is being measured by each TWE topic to estimate the relative difficulty of topics. Such judgments could be informed by the system that the TWE program is developing to classify essay topics based on their rhetorical demands.

If information is available on how various rhetorical tasks affect the abilities of examinees with different TOEFL scores to write essays, perhaps expert judgment might be used to project TWE score distributions. Once essays are classified according to their rhetorical tasks, the proportion of examinees who achieved each TWE score (1.0 to 6.0) for different ranges of TOEFL scores could be computed. Then, expected score distributions could be composed for each new TWE essay topic, based on the observed TOEFL scores of the examinees and the topic's rhetorical task classification. These expected distributions would be amenable to equipercentile equating. Test developers or the TWE Core Reader Group could classify topics after pretest data are collected. The Core Reader Group, writing experts composed from the academic community, is responsible for developing and approving all TWE topics.

Recommendations

Perhaps judgmental or empirical equating could be done at pretest time using some combination of TOEFL total score and topic rhetorical analysis. Because the objective is to administer topics that are equivalent in difficulty, we would want to retain all topics for which the equated score (rounded to the nearest TWE scale half point) and the observed score were the same. All other topics should be revised and pretested again. Ultimately, topics that are close in difficulty could be administered, and the scores would be reported on the same, easily interpreted 1.0 to 6.0 scale.

It is expected that with matching criteria for equating that accurately define examinee writing skills the residuals between equated scores and observed scores would be smaller. Although there is no set standard of appropriateness of equating models, it is clear that assumptions underlying the equating models examined in this study are not easily met when TOEFL scores are used to gauge the skills of TWE examinees. If such models are used, they should be augmented with information about topic demands. Research into the application of judgmental models should be pursued, as these may well prove to be the best and most accessible methods.

Table 1

Test of Written English
Administration Dates and Examinee
Sample Sizes

Admin. Date (Topic type)		Region			Different Regional Topics?[1]
	A	B	C	Total	
July 1986 (compare/contrast)	4,620	724	5,069	10,413	No
November 1986 (compare/contrast)	19,966	7,196	15,920	43,082	Yes
May 1987 (chart/graph)	27,776	10,568	18,586	56,930	Yes
July 1987 (chart/graph)	4,984	704	5,292	10,980	No
November 1987 (compare/contrast)	22,459	8,350	18,772	49,587	Yes
May 1988 (compare/contrast)	31,166	10,443	19,917	61,526	Yes
September 1988 (compare/contrast)	8,667	1,726	4,843	15,236	No
October 1988 (compare/contrast)	36,972	7,951	7,731	62,654	Yes

[1] Either a single topic ("no") or three different topics ("yes") were administered worldwide.

Table 2

Pearson Product Moment
Correlations of the Test of
Written English with the TOEFL Scaled Scores
and Multiple Correlation of the Test of
Written English with All Three Sections
of the TOEFL Test, by Region

TWE Admin Date	Region	Section 1	Section 2	Section 3	Total Score	Multiple R
July 1986	A	.611	.544	.549	.636	.655
	B	.618	.706	.685	.726	.729
	C	.575	.620	.603	.662	.663
	All	.597	.576	.568	.653	.662
November 1986	A	.639	.679	.652	.717	.721
	B	.651	.683	.558	.690	.699
	C	.628	.680	.651	.714	.715
	All	.589	.689	.659	.711	.714
May 1987	A	.598	.580	.607	.658	.660
	B	.587	.697	.684	.717	.725
	C	.564	.619	.597	.657	.658
	All	.568	.619	.621	.671	.671
July 1987	A	.634	.589	.537	.647	.669
	B	.610	.722	.664	.716	.728
	C	.515	.588	.561	.613	.615
	All	.579	.589	.545	.634	.642
November 1987	A	.652	.665	.643	.716	.719
	B	.551	.645	.630	.665	.670
	C	.620	.670	.643	.708	.708
	All	.626	.664	.638	.709	.709

Table 2 (cont.)

TWE						
Admin		Section	Section	Section	Total	Multiple
Date	Region	1	2	3	Score	R
May 1988	A	.625	.599	.592	.668	.675
	B	.565	.652	.623	.674	.677
	C	.578	.625	.605	.665	.667
	All	.581	.623	.605	.669	.671
September 1988	A	.640	.574	.558	.652	.672
	B	.637	.731	.725	.752	.758
	C	.571	.616	.610	.664	.664
	All	.627	.600	.581	.667	.677
October 1988	A	.647	.552	.641	.671	.687
	B	.544	.646	.631	.655	.666
	C	.599	.600	.610	.663	.664
	All	.620	.565	.625	.663	.670

Table 3

Sample Means and Standard Deviations on the TOEFL Total Score and Section 2 Score, and on the TWE Score, by Administration and Administrative Region

		TOEFL Total		TOEFL Sec. 2		TWE	
Administration	Region	Mean	S.d.	Mean	S.d.	Mean	S.d.
July 1986	A	506.00	61.97	51.30	7.42	3.19	1.03
	B	495.24	75.43	49.41	8.43	3.33	1.11
	C	507.37	64.62	49.68	7.29	3.40	0.98
	All	505.92	64.34	50.38	7.47	n/a	n/a
November 1986	A	522.12	68.51	52.48	8.07	3.61	1.08
	B	548.01	72.15	54.84	8.44	3.83	1.05
	C	520.50	69.07	50.04	8.07	3.30	0.98
	All	525.85	70.05	52.21	8.26	n/a	n/a
May 1987	A	515.74	63.75	52.15	7.38	3.71	0.97
	B	532.68	72.61	53.03	8.17	3.79	1.07
	C	513.37	65.84	52.15	7.38	3.61	0.87
	All	518.11	66.53	51.66	7.70	n/a	n/a
July 1987	A	507.36	62.98	51.59	7.52	3.45	0.96
	B	499.77	75.54	50.07	8.83	3.51	1.08
	C	504.37	63.66	50.04	7.53	3.58	0.91
	All	505.43	64.21	50.74	7.65	n/a	n/a
November 1987	A	529.15	67.71	52.59	7.96	3.61	1.08
	B	556.50	69.65	55.30	8.15	4.00	1.05
	C	527.99	67.41	49.68	8.02	3.67	0.98
	All	533.33	68.73	52.49	8.14	n/a	n/a
May 1988	A	514.73	64.83	59.26	7.27	3.71	0.95
	B	534.58	72.65	54.09	8.31	3.91	0.90
	C	513.43	66.27	50.87	7.60	3.53	0.91
	All	517.68	67.12	52.48	7.65	n/a	n/a
September 1988	A	510.25	67.04	51.62	7.43	3.53	0.91
	B	507.31	83.39	50.94	9.41	3.77	1.03
	C	515.14	66.76	51.63	7.49	3.78	0.86
	All	511.47	69.05	51.36	7.71	n/a	n/a
Oc ober 1988	A	529.31	67.72	54.03	8.01	3.65	0.99
	B	543.45	76.39	54.23	6.29	3.85	0.92
	C	514.47	66.96	50.51	7.71	3.68	0.92
	All	526.96	69.26	53.08	8.12	n/a	n/a

Table 4

Ratio of Variances and Standardized Mean Differences on TOEFL Scores for the July 1986 Examinee Groups and Other Examinee Groups, by Administration Date and Region
(Underlined Values Violate Equating Criteria)

TOEFL	Admin.		Administrative Region		
Score	Date		A	B	C
Sec. 2	11/86	Var. Ratio	0.84	1.00	0.82
		Mean Diff.	0.15	0.63	0.05
	5/87	Var. Ratio	1.01	1.06	0.98
		Mean Diff.	0.12	0.44	0.33
	7/87	Var. Ratio	0.97	0.91	0.94
		Mean Diff.	0.04	0.08	0.05
	11/87	Var. Ratio	0.87	1.07	0.83
		Mean Diff.	0.16	0.71	0.00
	5/88	Var. Ratio	1.04	1.03	0.92
		Mean Diff.	0.23	0.55	0.16
	9/88	Var. Ratio	1.00	0.80	0.95
		Mean Diff.	0.04	0.17	0.26
	10/88	Var. Ratio	0.86	1.03	0.90
		Mean Diff.	0.34	0.57	0.11
Total	11/86	Var. Ratio	0.82	1.09	0.88
		Mean Diff.	0.24	0.71	0.19
	5/87	Var. Ratio	0.95	1.08	0.96
		Mean Diff.	0.15	0.51	0.09
	7/87	Var. Ratio	0.97	1.00	1.03
		Mean Diff.	0.02	0.06	0.05
	11/87	Var. Ratio	0.84	1.17	0.92
		Mean Diff.	0.34	0.85	0.31
	5/88	Var. Ratio	0.91	1.08	0.95
		Mean Diff.	0.14	0.54	0.09
	9/88	Var. Ratio	0.85	0.82	0.94
		Mean Diff.	0.07	0.15	0.12
	10/88	Var. Ratio	0.84	0.98	0.93
		Mean Diff.	0.34	0.62	0.11

Table 5

Biserial Correlations of TOEFL Items and TWE Scores, by TOEFL Test Section for Each Administrative Region, for the Combined July 1986, May 1987, July 1987, and May 1988 Administrations

Region	TOEFL Section		Lower than 0.30	0.30 to 0.39	0.40 or Higher	Total	Chi-Square[1]
A	1	N	76	105	19	200	14.86
		%	38%	53%	10%	100%	
	2	N	51	71	30	152	
		%	34	47	20	100	
	3	N	109	95	28	232	
		%	47	41	12	100	
	All Sections	N	236	271	77	584	
		%	40	46	13	100	
B	1	N	60	88	52	200	63.15
		%	57	41	20	100	
	2	N	15	38	99	152	
		%	10	25	65	100	
	3	N	31	87	114	232	
		%	13	38	49	100	
	All Sections	N	106	213	265	584	
		%	18	36	45	100	
C	1	N	107	85	8	200	32.42
		%	54	43	4	100	
	2	N	53	64	35	152	
		%	35	42	23	100	
	3	N	100	103	29	232	
		%	43	44	13	100	
	All Sections	N	260	252	72	584	
		%	45	43	12	100	

[1] All chi-square values exceed the .01 level of significance.

Table 6
Regression of TWE Scores on TOEFL Scores
by Administrative Region,
May 1987-May 1988

TWE Admin		TOEFL Section 2		TOEFL Total Score	
Date	Region	Slope[1]	Intercept	Slope[1]	Intercept
July 1986	A	8.45	-1.15	1.05	-2.13
	B	9.27	-1.25	1.07	-1.95
	C	8.31	-0.73	1.00	-1.68
November 1986	A	9.08	-1.16	1.13	-2.29
	B	8.52	-0.84	1.01	-1.68
	C	8.29	-0.85	1.02	-1.99
May 1987	A	7.60	-0.25	1.00	-1.43
	B	9.14	-1.06	1.06	-1.84
	C	7.32	-0.20	0.87	-0.86
July 1987	A	7.54	-0.45	0.99	-1.58
	B	8.82	-0.91	1.02	-1.61
	C	7.12	0.02	0.88	-0.85
November 1987	A	9.05	-1.15	1.15	-2.45
	B	8.27	-0.57	1.00	-1.55
	C	8.17	-0.39	1.03	-1.75
May 1988	A	7.79	-0.41	0.97	-1.30
	B	7.05	0.10	0.83	-0.54
	C	7.48	-0.27	0.91	-1.15
September 1988	A	7.02	-0.09	0.88	-0.98
	B	7.98	-0.29	0.93	-0.93
	C	7.10	0.11	0.86	-0.65
October 1988	A	8.02	-0.68	0.98	-1.56
	B	7.16	-0.03	0.79	-0.44
	C	7.16	0.06	0.91	-1.00

[1] In hundredths.

Table 7

Analysis of Regression Slopes, TOEFL Scores on TWE Scores, in Relation to Administrative Region
May 1987-May 1988
(Means Sharing the Same Underline Are Not Significantly Different)

TWE Admin. Date	TOEFL Score	F-ratio, Region by TWE Regression	Adjusted Means for Region A	 B	 C	F-ratio Region
7/86	Section 1	1.86	49.02	50.21	51.80	26.49**
	Section 2	29.89**	51.73	49.27	49.22	88.66**
	Section 3	30.07**	52.33	48.71	49.90	115.91**
	Total	27.14**	510.27	493.99	503.09	50.89**
5/87	Section 1	3.09*	50.95	53.81	54.30	121.47**
	Section 2	164.15**	52.07	52.54	50.59	248.62**
	Section 3	119.37**	51.47	52.05	50.35	165.60**
	Total	90.60**	514.94	528.00	517.45	56.24**
7/87	Section 1	20.31**	50.05	52.00	51.84	57.78**
	Section 2	15.95**	51.90	50.12	49.72	39.61**
	Section 3	23.40**	51.13	47.94	48.91	72.74**
	Total	8.71**	510.24	500.20	501.58	15.69**
5/88	Section 1	18.18**	49.78	52.53	53.19	255.62**
	Section 2	199.52**	52.85	52.74	51.69	228.16**
	Section 3	135.86**	51.45	51.44	51.44	120.47**
	Total	100.02**	513.58	522.34	521.05	45.79**

* Exceeds the p < .05 level of significance.
** Exceeds the p < .001 level of significance.

Table 8

Linear Equating of TWE Topics
to July 1986 TWE Scale, by
Region and Administration Date
(Rounded Converted Scores That Do Not
Equal Observed Score Are Underlined)

Anchor Test: TOEFL Total

Observed Score	Region	11/86	5/87	7/87	11/87	5/88	9/88	10/88
1.0	A	0.77	0.38	0.58	0.87	0.28	0.28	0.59
	B	0.97	0.90	0.81	0.92	0.23	0.30	0.38
	C	1.18	0.51	0.62	0.89	0.71	0.29	0.58
1.5	A	1.27	0.92	1.12	1.36	0.84	0.87	1.13
	B	1.49	1.40	1.32	1.43	0.83	0.87	0.99
	C	1.69	1.07	1.16	1.40	1.25	0.87	1.12
2.0	A	1.76	1.45	1.66	1.85	1.39	1.45	1.67
	B	2.00	1.91	1.83	1.94	1.43	1.44	1.59
	C	2.20	1.64	1.69	1.91	1.80	1.44	1.66
2.5	A	2.26	1.99	2.19	2.34	1.94	2.03	2.20
	B	2.52	2.42	2.34	2.45	2.04	2.01	2.20
	C	2.71	2.20	2.22	2.42	2.34	2.01	2.20
3.0	A	2.76	2.53	2.73	2.84	2.49	2.62	2.74
	B	3.03	2.92	2.86	2.96	2.64	2.57	2.81
	C	3.23	2.77	2.78	2.92	2.68	2.59	2.74
3.5	A	3.25	3.06	3.26	3.33	3.05	3.20	3.27
	B	3.54	3.43	3.32	3.46	3.24	3.14	3.41
	C	3.74	3.33	3.29	3.43	3.42	3.16	3.28
4.0	A	3.75	3.60	3.80	3.82	3.60	3.78	3.81
	B	4.06	3.94	3.88	3.97	3.85	3.71	4.02
	C	4.25	3.89	3.82	3.94	3.87	3.73	3.82

Table 8 (cont.)

Anchor Test: TOEFL Total

Observed Score	Region	11/86	5/87	7/87	11/87	5/88	9/88	10/88
4.5	A	4.25	4.14	4.33	4.31	4.15	4.36	4.34
	B	4.57	4.44	4.39	4.48	4.45	4.28	4.63
	C	4.76	4.46	4.35	4.45	4.51	4.30	4.36
5.0	A	4.74	4.67	4.87	4.80	4.70	4.95	4.88
	B	5.09	4.95	4.91	4.99	5.06	4.85	5.24
	C	5.27	5.02	4.88	4.96	5.05	4.88	4.89
5.5	A	5.24	5.21	5.41	5.29	5.25	5.53	5.42
	B	5.60	5.45	5.42	5.50	5.66	5.42	5.84
	C	5.78	5.59	5.62	5.47	5.60	5.45	5.43
6.0	A	5.73	5.75	5.94	5.78	5.81	6.11	5.95
	B	6.11	5.96	5.93	6.00	6.26	5.99	6.45
	C	6.30	6.15	6.14	5.95	6.14	6.02	5.98

Table 8 (cont.)

Anchor Test: TOEFL Section 2

Observed Score	Region	11/86	5/87	7/87	11/87	5/88	9/88	10/88
1.0	A	0.73	0.39	0.60	0.76	0.41	0.36	0.60
	B	0.85	0.82	0.76	0.74	0.19	0.30	0.37
	C	1.05	0.66	0.63	0.63	0.73	0.39	0.56
1.5	A	1.22	0.92	1.13	1.25	0.95	0.92	1.13
	B	1.37	1.33	1.28	1.26	0.81	0.87	0.96
	C	1.57	1.23	1.17	1.15	1.28	0.96	1.10
2.0	A	1.71	1.45	1.67	1.73	1.49	1.49	1.67
	B	1.90	1.84	1.81	1.78	1.42	1.44	1.56
	C	2.08	1.79	1.72	1.67	1.83	1.53	1.65
2.5	A	2.20	1.98	2.20	2.22	2.02	2.05	2.20
	B	2.42	2.35	2.33	2.30	2.03	2.02	2.16
	C	2.60	2.35	2.26	2.18	2.37	2.10	2.19
3.0	A	2.69	2.51	2.74	2.71	2.56	2.62	2.73
	B	2.95	2.86	2.86	2.82	2.64	2.59	2.76
	C	3.12	2.91	2.80	2.70	2.92	2.67	2.73
3.5	A	3.18	3.03	3.27	3.19	3.10	3.18	3.26
	B	3.48	3.36	3.38	3.34	3.25	3.16	3.36
	C	3.63	3.48	3.34	3.22	3.46	3.24	3.28
4.0	A	3.67	3.56	3.81	3.67	3.64	3.75	3.79
	B	4.00	3.87	3.91	3.87	3.86	3.73	3.95
	C	4.15	4.04	3.88	3.74	4.01	3.81	3.82

Table 8 (cont.)

Anchor Test: TOEFL Section 2

Observed Score	Region	11/86	5/87	7/87	11/87	5/88	9/88	10/88
4.5	A	4.17	4.09	4.34	4.16	4.18	4.31	4.32
	B	4.53	4.38	4.43	4.39	4.47	4.30	4.55
	C	4.67	4.60	4.43	4.26	4.55	4.38	4.36
5.0	A	4.66	4.62	4.88	4.65	4.71	4.88	4.85
	B	5.06	4.89	4.96	4.91	5.09	4.87	5.15
	C	5.19	5.16	4.97	4.78	5.10	4.96	4.90
5.5	A	5.15	5.15	5.41	5.13	5.25	5.44	5.39
	B	5.58	5.40	5.48	5.43	5.70	5.44	5.75
	C	5.71	5.73	5.51	5.30	5.65	5.53	5.45
6.0	A	5.64	5.68	5.95	5.62	5.79	6.00	5.92
	B	6.11	5.91	6.01	5.95	6.30	6.01	6.35
	C	6.23	6.29	6.05	5.82	6.19	6.10	5.99

Table 9

Frequency Estimation
Conversions of TWE Scores
onto July 1986 Scale, by
Region and Administration Date
(Rounded Converted Scores That Do Not
Equal Observed Scores Are Underlined)

Anchor Test: TOEFL Section 2

Observed Score	Region	11/86	5/87	7/87	11/87	5/88	9/88	10/88
1.0	A	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	B	1.00	1.17	1.07	1.25	1.00	1.00	1.00
	C	1.24	1.00	1.00	1.00	1.00	1.00	1.00
1.5	A	1.00	1.00	1.00	1.08	1.00	1.00	1.00
	B	1.20	1.53	1.48	1.50	1.13	1.00	1.04
	C	1.62	1.00	1.00	1.28	1.03	1.00	1.00
2.0	A	1.62	1.48	1.69	1.66	1.43	1.56	1.61
	B	1.86	1.84	1.82	1.79	1.59	1.58	1.64
	C	2.16	1.65	1.78	1.78	1.82	1.56	1.66
2.5	A	2.04	1.78	2.08	2.04	1.81	1.87	1.99
	B	2.35	2.24	2.24	2.18	1.91	1.89	2.01
	C	2.59	2.09	2.19	2.19	2.30	1.90	2.06
3.0	A	2.81	2.59	2.78	2.77	2.65	2.73	2.76
	B	2.96	2.78	2.78	2.80	2.65	2.68	2.78
	C	3.19	2.78	2.83	2.85	2.96	2.73	2.78
3.5	A	3.17	2.93	3.13	3.10	2.98	3.06	3.15
	B	3.34	3.14	3.19	3.13	3.05	3.00	3.17
	C	3.69	3.24	3.22	3.28	3.45	3.12	3.18

Table 9 (cont.)

Anchor Test: TOEFL Section 2

Observed Score	Region	11/86	5/87	7/87	11/87	5/88	9/88	10/88
4.0	A	3.69	3.68	3.86	3.74	3.75	3.92	3.90
	B	4.07	3.88	3.95	3.87	4.01	3.87	4.19
	C	4.25	3.97	3.96	3.90	4.04	3.92	3.92
4.5	A	3.98	4.07	4.36	3.98	4.09	4.38	4.35
	B	4.65	4.50	4.64	4.46	4.68	4.44	4.75
	C	4.74	4.47	4.45	4.34	4.50	4.36	4.36
5.0	A	4.45	4.69	4.93	4.48	4.63	4.98	4.90
	B	5.18	5.15	5.13	5.10	5.38	5.03	5.42
	C	5.33	5.19	4.96	4.94	5.11	4.98	4.98
5.5	A	4.93	5.17	5.58	5.43	5.02	5.40	5.36
	B	5.57	5.64	5.55	5.57	5.76	5.54	5.73
	C	5.71	5.66	5.49	5.46	5.58	5.55	5.45
6.0	A	6.00	6.00	6.00	6.00	6.00	6.00	6.00
	B	6.00	6.00	6.00	6.00	6.00	6.00	6.00
	C	6.00	6.00	6.00	6.00	6.00	6.00	6.00

Table 9 (cont.)

Anchor Test: TOEFL Total

Observed Score	Region	11/86	5/87	7/87	11/87	5/88	9/88	10/88
1.0	A	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	B	1.03	1.25	1.13	1.34	1.00	1.00	1.00
	C	1.32	1.00	1.00	1.04	1.00	1.00	1.00
1.5	A	1.00	1.00	1.00	1.19	1.00	1.00	1.00
	B	1.29	1.57	1.52	1.53	1.14	1.00	1.03
	C	1.66	1.00	1.00	1.42	1.02	1.00	1.00
2.0	A	1.67	1.51	1.68	1.75	1.32	1.53	1.62
	B	1.94	1.91	1.85	1.86	1.60	1.58	1.65
	C	2.22	1.68	1.75	1.87	1.80	1.53	1.68
2.5	A	2.14	1.81	2.07	2.20	1.77	1.84	2.04
	B	2.48	2.33	2.25	2.29	1.92	1.90	2.02
	C	2.63	2.13	2.15	2.30	2.27	1.85	2.08
3.0	A	2.88	2.61	2.77	2.89	2.63	2.73	2.80
	B	3.00	2.83	2.77	2.88	2.64	2.66	2.82
	C	3.26	2.80	2.80	2.93	2.93	2.69	2.79
3.5	A	3.31	2.96	3.15	3.33	2.98	3.09	3.24
	B	3.37	3.19	3.17	3.22	3.02	2.97	3.22
	C	3.76	3.27	3.16	3.40	3.41	3.06	3.20
4.0	A	3.84	3.74	3.90	3.95	3.79	3.98	3.97
	B	4.09	3.92	3.91	3.90	3.96	3.83	4.25
	C	4.35	3.99	3.92	3.99	4.03	3.89	3.94

Table 9 (cont.)

Anchor Test: TOEFL Total

Observed Score	Region	11/86	5/87	7/87	11/87	5/88	9/88	10/88
4.5	A	4.23	4.20	4.43	4.36	4.20	4.58	4.47
	B	4.63	4.50	4.58	4.53	4.59	4.37	4.78
	C	4.87	4.52	4.40	4.47	4.52	4.35	4.39
5.0	A	4.74	4.82	5.00	4.88	4.79	5.15	4.98
	B	5.20	5.14	5.13	5.19	5.39	4.98	5.48
	C	5.45	5.25	4.94	5.14	5.20	4.96	5.02
5.5	A	5.25	5.33	5.67	5.37	5.27	5.52	5.47
	B	5.63	5.64	5.62	5.64	5.78	5.50	5.78
	C	5.78	5.69	5.49	5.60	5.66	5.54	5.48
6.0	A	6.00	6.00	6.00	6.00	6.00	6.00	6.00
	B	6.00	6.00	6.00	6.00	6.00	6.00	6.00
	C	6.00	6.00	6.00	6.00	6.00	6.00	6.00

Table 10

Means and Standard Deviations of Equated Scores and Equating Residuals, by Region, Administration Date, Anchor Test, and Method of Equating

						Residuals			
Anchor=Sec. 2		Linear		Equipercent.		Linear		Equipercent.	
Region	Admin.[1]	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
A	11/86	3.29	1.06	3.27	1.06	0.32	0.02	0.34	0.15
	5/87	3.26	1.03	3.29	1.05	0.45	0.05	0.42	0.14
	7/87	3.21	1.03	3.22	1.04	0.23	0.07	0.23	0.13
	11/87	3.30	1.05	3.30	1.10	0.31	0.03	0.31	0.14
	5/88	3.33	1.02	3.33	1.02	0.38	0.07	0.39	0.15
	9/88	3.22	1.03	3.30	1.06	0.31	0.12	0.23	0.20
	10/88	3.42	1.06	3.43	1.09	0.23	0.06	0.22	0.14
B	11/86	3.83	1.11	3.84	1.14	0.01	0.06	0.01	0.12
	5/87	3.66	1.09	3.68	1.16	0.13	0.02	0.11	0.16
	7/87	3.39	1.13	3.40	1.17	0.12	0.05	0.11	0.16
	11/87	3.87	1.09	3.90	1.14	0.13	0.05	0.11	0.15
	5/88	3.76	1.10	3.85	1.15	0.16	0.20	0.07	0.29
	9/88	3.47	1.17	3.53	1.18	0.30	0.14	0.24	0.20
	10/88	3.77	1.10	3.86	1.16	0.08	0.18	0.01	0.28
C	11/86	3.43	1.02	3.48	1.00	0.13	0.03	0.19	0.09
	5/87	3.60	0.98	3.47	1.02	0.01	0.11	0.14	0.16
	7/87	3.43	0.99	3.43	0.99	0.15	0.08	0.14	0.11
	11/87	3.40	1.02	3.52	1.02	0.2[illegible]	0.04	0.15	0.07
	5/88	3.50	0.99	3.50	0.98	0.03	0.08	0.03	0.09
	9/88	3.56	0.99	3.56	1.00	0.22	0.12	0.22	0.17
	10/88	3.47	1.00	3.49	1.02	0.21	0.08	0.19	0.13

[1] May 1987 and July 1987 were chart/graph type topics.

Table 10 (cont.)

						Residuals			
Anchor=Total		Linear		Equipercent.		Linear		Equipercent.	
Region	Admin.[1]	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
A	11/86	3.36	1.07	3.41	1.11	0.25	0.01	0.20	0.10
	5/87	3.29	1.04	3.35	1.08	0.42	0.07	0.36	0.16
	7/87	3.20	1.03	3.24	1.07	0.24	0.07	0.21	0.15
	11/87	3.44	1.06	3.48	1.12	0.17	0.02	0.13	0.08
	5/88	3.28	1.04	3.36	1.08	0.43	0.10	0.36	0.11
	9/88	3.24	1.06	3.32	1.08	0.30	0.15	0.21	0.22
	10/88	3.44	1.06	3.49	1.11	0.22	0.07	0.16	0.15
B	11/86	3.89	1.08	3.88	1.12	0.05	0.03	0.04	0.10
	5/87	3.73	1.08	3.72	1.14	0.07	0.14	0.07	0.14
	7/87	3.37	1.11	3.38	1.15	0.14	0.03	0.12	0.15
	11/87	3.98	1.06	3.96	1.13	0.03	0.02	0.04	0.15
	5/88	3.74	1.09	3.82	1.15	0.17	0.19	0.10	0.29
	9/88	3.43	1.17	3.51	1.17	0.32	0.14	0.26	0.19
	10/88	3.84	1.12	3.91	1.17	0.01	0.20	0.06	0.29
C	11/86	3.53	1.01	3.57	1.03	0.23	0.02	0.27	0.08
	5/87	3.46	0.98	3.50	1.02	0.16	0.11	0.11	0.17
	7/87	3.38	0.97	3.40	0.99	0.20	0.07	0.18	0.11
	11/87	3.60	0.99	3.63	1.04	0.07	0.02	0.04	0.08
	5/88	3.39	1.01	3.49	1.01	0.15	0.13	0.05	0.11
	9/88	3.48	0.99	3.54	1.01	0.30	0.13	0.24	0.17
	10/88	3.47	0.99	3.51	1.02	0.21	0.07	0.17	0.13

[1] May 1987 and July 1987 were chart/graph type topics.

Table 11

Means and Standard Deviations for Equated Scores by Topic Type (Chart/Graph or Compare/Contrast), Type of Equating, Matching Variable, and Administrative Region (Samples Not Meeting the Equating Criteria Eliminated)

Type of	Matching		Chart/Graph		Compare/Contrast	
Equating	Variable	Region	Mean	S.D.	Mean	S.D.
Linear	Sec. 2	A	3.25	1.03	3.30	1.04
		B	3.39	1.13	3.47	1.17
		C	3.43	0.99	3.45	1.01
		All	3.28	1.03	3.37	1.03
	Total	A	3.28	1.04	3.30	1.06
		B	3.37	1.07	3.45	1.17
		C	3.44	0.98	3.46	1.00
		All	3.35	1.02	3.38	1.04
Equiperc.	Sec. 2	A	3.28	1.05	3.30	1.07
		B	3.40	1.16	3.53	1.18
		C	3.43	0.99	3.50	1.01
		All	3.31	1.05	3.39	1.05
	Total	A	3.33	1.08	3.37	1.09
		B	3.38	1.15	3.51	1.17
		C	3.48	1.02	3.52	1.02
		All	3.39	1.06	3.44	1.06

Appendix

On pages 33-34 of the Test of Written English Guide, is a series of essay topics. All were administered to international examinee populations for whom English was the second language. The first was administered in July 1986. It was scored using the scale shown on page 29.

On the topic labelled "Sample 1" on page 33, the following distributions were achieved:

Scale Score	% Examinees	Cumulative % Examinees
1.0	2	2
1.5	2	4
2.0	8	12
2.5	10	22
3.0	23	45
3.5	16	61
4.0	17	78
4.5	9	87
5.0	6	93
5.5	4	97
6.0	3	100

Please look at Sample 5 on page 34. Considering the relative difficulty of the topic relative to the July 1986 topic, and the July 1986 distribution given above, please estimate the percentage of the July 1986 examinees that would achieve each of the scale scores from 1.0 to 6.0 This can be accomplished in one of two ways:

1.) estimate the percentage that would achieve each score, to total 100 over all scores;

2.) estimate a cumulative percentage for certain scores of examinees who would achieve th .t score or lower, and, by subtraction, estimate the percent that would achieve each separate score.

Score	% Examinees	Cum % Examinees
1.0		
1.5		
2.0		
2.5		
3.0		
3.5		
4.0		
4.5		
5.0		
5.5		
6.0		

References

Allen, N. L., & Holland, P. W. (1989). Approaches to nonignorable nonresponse with application to selection bias. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Angelis, P. J. (1982). Academic needs and priorities for testing. *American Language Journal*, *1*, 41-56.

Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Princeton, NJ: Educational Testing Service.

Breland, H., Conlan, G., Fowles, M., & Livingston, S. (1987). *Guidelines for developing and scoring free-response tests*. Princeton, NJ: Educational Testing Service.

Educational Testing Service (1987). *Standards for quality and fairness*. Princeton, NJ: Author

Educational Testing Service (1990). *TOEFL test and score manual*. Princeton, NJ: Author

Keats, J. A. & Lord, F. M. (1962). A theoretical distribution for mental test scores. *Psychometrika,* 1962, *27*, 59-72.

Golub-Smith, M., Reese, C., & Steinhaus, K. (1991). Topic and topic type comparability on the Test of Written English. Manuscript submitted for publication.

Levine, R. (1955). Equating the score scales of alternate forms administered to samples of different ability. *Research Bulletin* RB-55-23. Princeton, NJ: Educational Testing Service.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.

Modu, C.C. & Stern, J. (1975). The stability of the SAT score scale. (Research and Development Report 74-75, No. 3). Princeton, NJ: Educational Testing Service.

Phillips, G. W. (1989). Statistical issues in equating writing assessment. Paper presented for symposium at the annual conference of the American Educational Research Association, San Francisco.

Stansfield, C. W., & Ross, J. (1988). A long term research agenda for the Test of Written English. Princeton, NJ: Educational Testing Service.

Wright, B.D., & Stone, M. H. (1979). *Best test design: Rasch measurement*. Chicago: MESA Press.

Printed on Recycled Paper

57906-01199 • Y52M.5 • 275583 • Printed in U.S.A.