TOEFL TEST OF WRITTEN ENGLISH GUIDE

Overview of the TWE Test

The Test of Written English (TWE) is the essay component of the Test of English as a Foreign Language (TOEFL), the multiple-choice test used by more than 2,400 institutions to evaluate the English proficiency of applicants whose native language is not English. As a direct, productive skills test, the TWE® test is intended to complement TOEFL Section 2 (Structure and Written Expression). The TWE test is holistically scored, using a criterion-referenced scale to provide information about an examinee's ability to generate and organize ideas on paper, to support those ideas with evidence or examples, and to use the conventions of standard written English.

Introduced in July 1986, the TWE test is currently (1996) offered as a required component of the TOEFL test at five administrations a year — in February, May, August, October, and December. There is no additional fee for the TWE test.

The TOEFL Test

First administered in 1963-64, the TOEFL test is primarily intended to evaluate the English proficiency of nonnative speakers who wish to study in colleges or universities in English-speaking countries. Section 1 (Listening Comprehension) measures the ability to recognize and understand English as it is spoken in North America. Section 2 (Structure and Written Expression) measures the ability to recognize selected structural and grammatical points in English. Section 3 (Reading Comprehension) measures the ability to read and understand short passages similar in topic and style to those that students are likely to encounter in North American universities and colleges.

During the 1994-95 testing year, more than 845,000 persons in more than 180 countries and regions registered to take the TOEFL test.

TWE Developmental Research

Early TOEFL research studies (Pike, 1976; Pitcher & Ra, 1967) showed that performance on the TOEFL Structure and Written Expression section correlated positively with scores on direct measures of writing ability. However, some TOEFL score users expressed concern about the validity of Section 2 as a measure of a nonnative speaker's ability to write for academic purposes in English. The perception among many graduate faculty was that there might be little actual relationship between the *recognition* of correct written expression, as measured by Section 2, and the *production* of an organized essay or report (Angelis, 1982).

In surveys conducted in a number of studies (Angelis, 1982; Hale and Hinofotis, 1981; Kane, 1983) college and university administrators and faculty, as well as English as a second language (ESL) teachers, requested the development of an essay test to assess directly the academic writing skills of foreign students.

As an initial step in exploring the development of an essay component for the TOEFL test, Bridgeman and Carlson (1983) surveyed faculty in undergraduate and graduate departments with large numbers of foreign students at 34 major universities. The purpose of their study was to identify the types of academic writing tasks and skills required of college and university students.

Following the identification of appropriate writing tasks and skills, a validation study investigating the relationship of TOEFL scores to writing performance was conducted (Carlson, Bridgeman, Camp, and Waanders, 1985). It was found that, while scores on varied writing samples and TOEFL scores were moderately related, the writing samples and the TOEFL test reliably measured some aspect of English language proficiency not assessed by the other. The researchers also found that holistic scores, discourse-level scores, and sentence-level scores of the writing samples were all closely related. Finally, the researchers reported that correlations of scores were as high across writing topic types as within the topic types, suggesting that the different topic types used in the study comparably assessed overall competency in academic composition.

These research studies provided the foundation for the development of the Test of Written English. Early TWE topics were based on the types of writing tasks identified in the Bridgeman and Carlson (1983) study. Based on the findings of the validation study, a single holistic score is reported for the TWE test. This score is derived from a criterion-referenced scoring guide that encompasses relevant aspects of communicative competence.

TWE ITEM DEVELOPMENT

The TWE Committee

Tests developed by Educational Testing Service must meet requirements for fair and accurate testing, as outlined in the *ETS Standards for Quality and Fairness* (Educational Testing Service, 1987). These standards advise a testing program to:

> Obtain substantive contributions to the test development process from qualified persons who are not on the ETS staff and who represent valid perspectives, professional specialties, population subgroups, and institutions.
>
> Have subject matter and test development specialists who are familiar with the specifications and purpose of the test and with its intended population review the items for accuracy, content appropriateness, suitability of language, difficulty, and the adequacy with which the domain is sampled. (pp. 10-11)

In accordance with these ETS standards, in July 1985 the TOEFL program established the TWE Core Reader Group, now known as the TWE Committee. The committee is a consultant group of college and university faculty and administrators who are experienced with the intended test population, current writing assessment theory and practice, pedagogy, and large-scale essay testing management. The committee develops the TWE essay questions, evaluates their pretest performance using the TWE scoring criteria, and approves the items for administration. Members also participate in TWE essay readings throughout the year.

TWE Committee members are rotated on a regular basis to ensure the continued introduction of new ideas and perspectives related to the assessment of English writing. Appendix A lists current and former committee members.

Test Specifications

Test specifications outline what a test purports to measure and how it measures the identified skills. The purpose of TWE is to give examinees whose native language is not English an opportunity to demonstrate their ability to express ideas in acceptable written English in response to an assigned topic. Topics are designed to be fair, accessible, and appropriate to all members of the international TOEFL population. Each essay is judged according to lexical and syntactic standards of English and the effectiveness with which the examinee, organizes, develops, and expresses ideas in writing. A criterion-referenced scoring guide ensures that a level of consistency in scoring is maintained from one administration to another.

Development of the TWE Scoring Guide

The TWE Scoring Guide (see Appendix B) was developed to provide concise descriptions of the general characteristics of essays at each of six points on the criterion-referenced scale. The scoring guide also serves to maintain consistent scoring standards and high interrater reliability within and across administrations. As an initial step in developing these guidelines, a specialist in applied linguistics examined 200 essays from the Carlson et al. (1985) study — analyzing the rhetorical, syntactic, and communicative characteristics at each of the six points — and wrote brief descriptions of the strengths and weaknesses of the group of essays at each level. This analysis, the TWE Committee's analysis of pretest essays, and elements of scoring guides used by other large-scale essay reading programs at ETS and elsewhere were used to develop the TWE Scoring Guide.

The guide was validated on the aforementioned research essays and on pretest essays before being used to score the first TWE essays in July 1986. To maintain consistency in the interpretation and application of the guide, before each TWE essay reading TWE essay reading managers review a sample of essays that are anchored to the original essays from the first TWE administration. This review helps to ensure that a given score will consistently represent the same proficiency level across test administrations.

In September 1989 the TWE Scoring Guide was revised by a committee of TWE essay reading managers who were asked to refine it while maintaining the comparability of scores assigned at previous TWE essay readings. The revisions were based on feedback from TWE essay readers, essay reading managers, and the TWE Committee.

The primary purpose of the revision was to make the guide a more easily internalized tool for scoring TWE essays during a reading. After completing the revisions, the committee of essay reading managers rescored essays from the first TWE administration to see that no shift in scoring occurred.

The revised scoring guide was reviewed, used to score pretest essays, and approved by the TWE Committee in February 1990. It was introduced at the March 1990 TWE reading.

TWE Essay Questions

The TWE test requires examinees to produce an essay in response to a brief question or topic. The writing tasks presented in TWE topics have been identified by research as typical of those required for college and university course work. The topics and tasks are designed to give examinees the opportunity to develop and organize ideas and to express those ideas in lexically and syntactically appropriate English. Because TWE aims to measure composition skills rather than reading comprehension skills, topics are brief, simply worded, and not based on reading passages. Samples of TWE essay questions used in past administrations are included in Appendix D.

TWE questions are developed in two stages. The TWE Committee writes, reviews, revises, and approves essay topics for pretesting. In developing topics for pretesting, the committee considers the following criteria:

- the topic (prompt) should be accessible to TOEFL examinees from a variety of linguistic, cultural, and educational backgrounds
- the task to be performed by examinees should be explicitly stated
- the wording of the prompt should be clear and unambiguous
- the prompt should allow examinees to plan, organize, and write their essays in 30 minutes

Once approved for pretesting, each TWE question is further reviewed by ETS test developers and sensitivity reviewers to ensure that it is not biased, inflammatory, or misleading, and that it does not unfairly advantage or disadvantage any subgroup within the TOEFL population.

As more is learned about the processes and domains of academic writing, TWE test developers and researchers will explore the use of different kinds of writing topics and tasks in the TWE test.

TWE Pretesting Procedures

Each potential TWE item or prompt is pretested with international students (both undergraduate and graduate) studying in the United States and Canada who represent a variety of native languages and English proficiency levels. Pretesting is conducted primarily in English language institutes and university composition courses for nonnative speakers of English.

Each pretest item is sent to a number of institutions in order to obtain a diverse sample of examinees and essays. The pretest sites are chosen on the basis of geographic location, type of institution, foreign student population, and English language proficiency levels of the students at the site. The goal is to obtain a population similar to the TOEFL/TWE test population.

During a pretest administration, writers have 30 minutes to plan and write an essay under standardized testing procedures similar to those used in operational TWE administrations. The essays received for each item are then prepared for the TWE Committee to evaluate. When evaluating pretest essays, the committee is given detailed information on the examinees (native language, undergraduate/graduate status, language proficiency test scores, if known) as well as feedback received on each essay question from pretest supervisors and examinees.

After a representative sample of pretest essays has been obtained, the sample is reviewed by the TWE Committee to evaluate the effectiveness of each prompt. An effective prompt is one that is easily understood by examinees at a range of language proficiencies and that elicits essays that can be validly and consistently scored according to the TWE scoring guide. The committee is also concerned that the prompt engage the writers, and that the responses elicited by the prompt be varied and interesting enough to engage readers. If the committee approves a prompt after reading the sample of pretest essays, it may be used in an operational TOEFL/TWE test administration.

TWE ESSAY READINGS

Reader Qualifications

Readers for the TWE test are primarily English and ESL writing specialists affiliated with accredited colleges, universities, and secondary schools in the United States and Canada. In order to be invited to serve as a reader, an individual must have read successfully for at least one other ETS program or qualify at a TWE reader training session.

TWE reader training sessions are conducted as needed. During these sessions, potential readers receive intensive training in holistic scoring procedures using the TWE Scoring Guide and TWE essays. At the conclusion of the training, participants independently rate 50 TWE essays that were scored at an operational reading. To qualify as a TWE rater, participants must demonstrate their ability to evaluate TWE essays reliably and accurately using the TWE Scoring Guide.

Scoring Procedures

All TWE essay readings are conducted in a central location under standardized procedures to ensure the accuracy and reliability of the essay scores.

TWE essay reading managers are English or ESL faculty who represent the most capable and experienced readers. In preparation for a TWE scoring session, the essay reading managers prepare packets of sample essays illustrating the six points on the scoring guide. Readers score and discuss these sets of sample essays with the essay reading managers prior to and throughout the reading to maintain scoring accuracy.

Small groups of readers work under the direct supervision of reading managers, who monitor the performance of each scorer throughout the reading. Each batch of essays is scrambled between the first and second readings to ensure that readers are not unduly influenced by the sequence of essays.

Each essay is scored by two readers working independently. The score assigned to an essay is derived by averaging the two independent ratings or, in the case of a discrepancy of more than one point, by the adjudication of the score by a reading manager. For example, if the first reader assigns a score of 5 to an essay and the second reader also assigns it a score of 5, 5 is the score reported for that essay. If the first reader assigns a score of 5 and the second reader assigns a score of 4, the two scores are averaged and a score of 4.5 is reported. However, if the first reader assigns a score of 5 to an essay and the second reader assigns it a 3, the scores are considered discrepant. In this case, a reading manager scores the essay to adjudicate the score.

Using the scenario above of first and second reader scores of 3 and 5, if the reading manager assigns a score of 4, the three scores are averaged and a score of 4 is reported. However, if the reading manager assigns a score of 5, the discrepant score of 3 is discarded and a score of 5 is reported. To date, more than 2,500,000 TWE essays have been scored, resulting in some 5,000,000 readings. Discrepancy rates for the TWE readings have been extremely low, usually ranging from 1 to 2 percent per reading.

TWE SCORES

Six levels of writing proficiency are reported for the TWE test. TWE scores range from 6 to 1 (see Appendix B). A score between two points on the scale (5.5, 4.5, 3.5, 2.5, 1.5) can also be reported (see "Scoring Procedures" above). The following codes and explanations may also appear on TWE score reports:

1NR	Examinee did not write an essay.
OFF	Examinee did not write on the assigned topic.
*	TWE not offered on this test date.
**	TWE score not available.

Because language proficiency can change considerably in a relatively short period, the TOEFL office will not report TWE scores that are more than two years old. Therefore, individually identifiable TWE scores are retained in a database for only two years from the date of the test. After two years, information that could be used to identify an individual is removed from the database. Information such as score data and essays that may be used for research or statistical purposes may be retained indefinitely; however, this information does not include any individual examinee identification.

TWE scores and all information that could identify an examinee are strictly confidential. An examinee's official TWE score report will be sent only to those institutions or agencies designated by the examinee on the answer sheet on the day of the test, or on a Score Report Request Form submitted by the examinee at a later date, or by other written authorization from the examinee.

Examinees receive their test results on a form titled **Examinee's Score Record**. These are not official TOEFL score reports and should not be accepted by institutions. If an examinee submits a TWE score to an institution or agency and there is a discrepancy between that score and the official TWE score recorded at ETS, ETS will report the official score to the institution or agency. Examinees are advised of this policy in the ***Bulletin of Information for TOEFL, TWE, and TSE***.

A TWE rescoring service is available to examinees who would like to have their essays rescored. Further information on this rescoring process can also be found in the ***Bulletin of Information for TOEFL, TWE, and TSE***.

GUIDELINES FOR USING TWE TEST SCORES

An institution that uses TWE scores should consider certain factors in evaluating an individual's performance on the test and in determining appropriate TWE score requirements. The following guidelines are presented to assist institutions in arriving at reasonable decisions.

1. Use the TWE score as an indication of English writing proficiency only and in conjunction with other indicators of language proficiency, such as TOEFL section and total scores. **Do not use the TWE score to predict academic performance.**

2. Base the evaluation of an applicant's readiness to begin academic work on all available relevant information and recognize that the TWE score is only one indicator of academic readiness. The TWE test provides information about an applicant's ability to compose academic English. Like TOEFL, TWE is **not** designed to provide information about scholastic aptitude, motivation, language learning aptitude, field specific knowledge, or cultural adaptability.

3. Consider the kinds and levels of English writing proficiency required at different levels of study in different academic disciplines. Also consider the resources available at the institution for improving the English writing proficiency of students for whom English is not the native language.

4. Consider that examinee scores are based on a single 30-minute essay that represents a first-draft writing sample.

5. Use the TWE Scoring Guide and writing samples illustrating the guide as a basis for score interpretation (see Appendix B and E). Score users should bear in mind that a TWE score level represents a range of proficiency and is not a fixed point.

6. Avoid decisions based on small score differences. Small score differences (i.e., differences less than approximately two times the standard error of measurement) should not be used to make distinctions among examinees. Based upon the average standard error of measurement for the past 10 TWE administrations, distinctions among individual examinees should not be made unless their TWE scores are **at least** one point apart.

7. Conduct a local validity study to assure that the TWE scores required by the institution are appropriate.

As part of its general responsibility for the tests it produces, the TOEFL program is concerned about the interpretation and use of TWE test scores by recipient institutions. The TOEFL office encourages individual institutions to request its assistance with any questions related to the proper use of TWE scores.

STATISTICAL CHARACTERISTICS OF THE TWE TEST

Reliability

The reliability of a test is the extent to which it yields consistent results. A test is considered reliable if it yields similar scores across different forms of the test, different administrations, and, in the case of subjectively scored measures, different raters.

There are several ways to estimate the reliability of a test, each focusing on a different source of measurement error. The reliability of the TWE test has been evaluated by examining interrater reliability, that is, the extent to which readers agree on the ratings assigned to each essay. To date, it has not been feasible to assess alternate-form and test-retest reliability, which focus on variations in test scores that result from changes in the individual or changes in test content from one testing situation to another. To do so, it would be necessary to give a relatively large random sample of examinees two different forms of the test (alternate-form reliability) or the same test on two different occasions (test-retest reliability). However, the test development procedures that are employed to ensure TWE content validity (discussed later in this section) would be expected to contribute to alternate-form reliability.

Two measures of interrater reliability are reported for the TWE test. The first measure reported is the Pearson product-moment correlation between first and second readers, which reflects the overall agreement (across all examinees and all raters) of the pairs of readers who scored each essay. The second measure reported is coefficient alpha, which provides an estimate of the internal consistency of the final scores based upon two readers per essay. Because each reported TWE score is the average of two separate ratings, the reported TWE scores are more reliable than the individual ratings. Therefore, coefficient alpha is generally higher than the simple correlation between readers, except in those cases where the correlation is equal to 0 or 1. (If there were perfect agreement on each essay across all raters, coefficient alpha would equal 1.0; if there were no relationship between the scores given by different raters, coefficient alpha would be 0.0.)

Table 1 contains summary statistics and interrater reliability statistics for the 10 TWE administrations from August 1993 through May 1995. The interrater correlations and coefficients alpha indicate that reader reliability is acceptably high, with correlations between first and second readers ranging from .77 to .81, and the values for coefficient alpha ranging from .87 to .89.

Table 1 also shows the reader discrepancy rate for each of the 10 TWE administrations. This value is simply the proportion of essays for which the scores of the two readers differed by two or more points. These discrepancy rates are quite low, ranging from 0.2 percent to 1.1 percent. (Because all essays with ratings that differed by two or more points were given a third reading, the discrepancy rates also reflect the proportions of essays that received a third reading.)

Table 1
Reader Reliabilities
(Based on scores assigned to 606,883 essays in the 10 TWE administrations from August 1993 through May 1995)

Admin. Date	N	TWE Mean	TWE S.D.	Discrepancy Rate[1]	Correlation 1st & 2nd Readers	Alpha	SEM[2]	
							Indiv. Scores	Score Diffs.
Aug. 1993	56,240	3.66	0.84	.011	.780	.876	.30	.42
Sept. 1993	27,951	3.69	0.78	.004	.788	.881	.27	.38
Oct. 1993	87,616	3.68	0.85	.010	.782	.877	.30	.42
Feb. 1994	48,694	3.65	0.89	.010	.799	.888	.30	.42
May 1994	74,972	3.73	0.83	.010	.767	.868	.30	.43
Aug. 1994	56,553	3.66	0.80	.007	.770	.870	.29	.41
Sept. 1994	28,282	3.71	0.78	.002	.807	.893	.26	.36
Oct. 1994	89,656	3.72	0.84	.009	.783	.878	.29	.41
Feb. 1995	54,783	3.65	0.84	.010	.777	.874	.30	.42
May 1995	82,136	3.65	0.84	.009	.777	.875	.30	.42

[1] Proportion of papers in which the two readers differed by two or more points. (When readers differed by two or more points, the essay was adjudicated by a third reader.)

[2] Standard errors of measurement listed here are based upon the extent of interrater agreement and do not take into account other sources of error, such as differences between test forms. Therefore, these values probably underestimate the actual error of measurement.

Standard Error of Measurement

Any test score is only an estimate of an examinee's knowledge or ability, and an examinee's test score might have been somewhat different if the examinee had taken a different version of the test, or if the test had been scored by a different group of readers. If it were possible to have someone take all the editions of the test that could ever be made, and have those tests scored by every reader who could ever score the test, the average score over all those test forms and readers presumably would be a completely accurate measure of the examinee's knowledge or ability. This hypothetical score is often referred to as the "true score." Any difference between this true score and the score that is actually obtained on a given test is considered to be measurement error.

Because an examinee's hypothetical true score on a test is obviously unknown, it is impossible to know exactly how large the measurement error is for any individual examinee. However, it is possible statistically to estimate the average measurement error for a large group of examinees, based upon the test's standard deviation and reliability. This statistic is called the Standard Error of Measurement (SEM).

The last two columns in Table 1 show the standard errors of measurement for individual scores and for score differences on the TWE test. The standard errors of measurement that are reported here are estimates of the average differences between obtained scores and the theoretical true scores that would have been obtained if each examinee's performance *on a single test form* had been scored by all possible readers. For the 10 test administrations shown in the table, the average standard error of measurement was approximately .29 for individual scores and .41 for score differences.

The standard error of measurement can be helpful in the interpretation of test scores. Approximately 95 percent of all examinees are expected to obtain scores within 1.96 standard errors of measurement from their true scores and approximately 90 percent are expected to obtain scores within 1.64 standard errors of measurement. For example, in the May 1995 administration (with SEM = .30), less than 10 percent of examinees with true scores of 3.0 would be expected to obtain TWE scores lower than 2.5 or higher than 3.5; of those examinees with true scores of 4.0, less than 10 percent would be expected to obtain TWE scores lower than 3.5 or higher than 4.5.

When the scores of two examinees are compared, the difference between the scores will be affected by errors of measurement in each of the scores. Thus, the standard errors of measurement for score differences are larger than the corresponding standard errors of measurement for individual scores (about 1.4 times as large). In approximately 95 percent of all cases, the difference between obtained scores is expected to be within 1.96 standard errors above or below the difference between the examinees' true scores; in approximately 80 percent of all cases, the difference between obtained scores is expected to be within 1.28 standard errors above or below the true difference. This information allows the test user to evaluate the probability that individuals with different obtained TWE scores actually differ in their true scores. For example, among all pairs of examinees with the same true scores (i.e., with true-score differences of zero) in the May 1995 administration, more than 20 percent would be expected to obtain TWE scores that differ from one another by one-half point or more; however, fewer than 5 percent (in fact, only about 1.7 percent) would be expected to obtain TWE scores more than one point apart.

Validity

Beyond being reliable, a test should be valid; that is, it should actually measure what it is intended to measure. It is generally recognized that validity refers to the usefulness of inferences made from a test score. The process of validation is necessarily an ongoing one, especially in the area of written composition, where theorists and researchers are still in the process of defining the construct.

To support the inferences made from test scores, validation should include several types of evidence. The nature of that evidence should depend upon the uses to be made of the test. The TWE test is used to make inferences about an examinee's ability to compose academically appropriate written English.

Two types of validity evidence are available for the TWE test: (1) construct-related evidence and (2) content-related evidence. Construct-related evidence refers to the extent to which the test actually measures the particular construct of interest, in this case, English-language writing ability. Content-related evidence refers to the extent to which the test provides an adequate and representative sample of the particular content domain that the test is designed to measure.

Construct-related Evidence. One source of construct-related evidence for the validity of the TWE test is the relationship between TWE scores and TOEFL scaled scores. Research suggests that skills such as those intended to be measured by both the TOEFL and TWE tests are part of a more general construct of English language proficiency (Oller, 1979). Therefore, in general, examinees who demonstrate high ability on TOEFL would not be expected to perform poorly on TWE, and examinees who perform poorly on TOEFL would not be expected to perform well on TWE.

This expectation is supported by the data collected over several TWE administrations. Table 2 displays the frequency distributions of TWE scores for five different TOEFL score ranges over 10 administrations.

Table 2
Frequency Distribution of TWE Scores for TOEFL Total Scaled Scores

(Based on 607,350 examinees who took the TWE test from August 1993 through May 1995)

TWE Score	TOEFL Scores Below 477		TOEFL Scores Between 477 and 523		TOEFL Scores Between 527 and 573		TOEFL Scores Between 577 and 623		TOEFL Scores Above 623	
	N	Percent	N	Percent	N	Percent	N	Percent	N	Percent
6.0	5	0.0+	55	0.04	402	0.23	1,703	1.54	4,338	10.36
5.5	27	0.02	205	0.13	1,224	0.71	3,612	3.27	5,190	12.40
5.0	564	0.43	2,949	1.94	10,962	6.36	19,415	17.57	13,276	31.71
4.5	1,634	1.25	6,695	4.39	16,877	9.80	18,783	17.00	7,275	17.38
4.0	20,429	15.68	50,451	33.10	75,860	44.03	47,286	42.79	9,594	22.92
3.5	18,910	14.51	29,066	19.07	28,956	16.81	10,951	9.91	1,383	3.30
3.0	49,948	38.34	47,702	31.30	31,838	18.48	7,804	7.06	721	1.72
2.5	17,161	13.17	9,203	6.04	4,096	2.38	685	0.62	57	0.14
2.0	15,771	12.11	5,182	3.40	1,785	1.04	228	0.21	27	0.06
1.5	2,979	2.29	518	0.34	165	0.10	23	0.02	2	0.0+
1.0	2,857	2.19	372	0.24	118	0.07	30	0.03	1	0.0+

As the data in Table 2 indicate, across the 10 TWE administrations from August 1993 through May 1995 it was rare for examinees to obtain either very high scores on the TOEFL test and low scores on the TWE test or very low scores on TOEFL and high scores on TWE. It should be pointed out, however, that ***the data in Table 2 do not suggest that TOEFL scores should be used as predictors of TWE scores.***

Although there are theoretical grounds for expecting a positive relationship between TOEFL and TWE scores, there would be no point in administering the TWE test to examinees if it did not measure an aspect of English language proficiency distinct from what is already measured by TOEFL. Thus, the correlations between TWE scores and TOEFL scaled scores should be high enough to suggest that TWE is measuring the appropriate construct, but low enough to support the conclusion that the test also measures abilities that are distinct from those measured by TOEFL. The extent to which TWE scores are independent of TOEFL scores is an indication of the extent to which the TWE test measures a distinct skill or skills.

Table 3 presents the correlations of TWE scores with TOEFL scaled scores for examinees within each of the three geographic regions in which TWE was administered at the 10 administrations. The correlations between the TOEFL total scores and TWE scores range from .57 to .68, suggesting that the productive writing abilities assessed by TWE are somewhat distinct from the proficiency skills measured by the multiple-choice items of the TOEFL test.

Table 3
Correlations between TOEFL and TWE Scores[1]

(Based on 606,883 examinees who took the TWE test from August 1993 through May 1995)

Admin. Date	Geographic Region[2]	N	Total r	TOEFL Section 1 r	TOEFL Section 2 r	TOEFL Section 3 r
Aug. 1993[3]	Region 1	27,807	.64	.66	.58	.57
	Region 2	12,072	.68	.66	.65	.62
	Region 3	16,361	.62	.60	.60	.57
Sept. 1993[3]	Region 1	6,662	.65	.66	.63	.53
	Region 2	10,961	.64	.62	.62	.59
	Region 3	10,328	.59	.55	.58	.53
Oct. 1993[3]	Region 1	41,638	.66	.65	.62	.62
	Region 2	16,288	.67	.65	.66	.60
	Region 3	29,690	.64	.63	.63	.58
Feb. 1994	Region 1	16,555	.65	.65	.59	.60
	Region 2	11,305	.60	.54	.60	.56
	Region 3	20,834	.61	.59	.58	.56
May 1994	Region 1	35,290	.60	.62	.55	.54
	Region 2	14,239	.59	.53	.59	.51
	Region 3	25,443	.64	.61	.62	.57
Aug. 1994	Region 1	36,137	.63	.64	.59	.54
	Region 2	4,010	.64	.56	.66	.60
	Region 3	16,406	.62	.58	.60	.54
Sept. 1994	Region 1	14,436	.62	.64	.57	.55
	Region 2	3,623	.66	.62	.66	.61
	Region 3	10,223	.57	.55	.55	.51
Oct. 1994	Region 1	48,628	.68	.68	.63	.62
	Region 2	10,289	.58	.52	.58	.54
	Region 3	30,739	.62	.58	.59	.58
Feb. 1995	Region 1	22,102	.65	.64	.60	.59
	Region 2	11,562	.61	.52	.64	.56
	Region 3	21,119	.59	.55	.57	.54
May 1995	Region 1	43,450	.65	.65	.62	.59
	Region 2	13,825	.64	.57	.66	.56
	Region 3	24,861	.63	.58	.62	.56

[1] Correlations have been corrected for unreliability of TOEFL scores.

[2] Geographic Region 1 includes Asia, the Pacific (including Australia), and Israel; Geographic Region 2 includes Africa, the Middle East, and Europe; Geographic Region 3 includes North America, South America, and Central America.

[3] For these administrations, some examinees from test centers in Asia are included in Region 2 and/or Region 3.

Table 3 also shows the correlations of TWE scores with each of the three TOEFL section scores. Construct validity would be supported by higher correlations of TWE scores with TOEFL Section 2 (Structure and Written Expression) than with Section 1 (Listening Comprehension) or Section 3 (Reading Comprehension) scores. In fact, this pattern is generally found in TWE administrations for Regions 2 and 3. In Region 1, however, TWE scores correlated more highly with TOEFL Section 1 scores than with Section 2 scores in all 10 administrations. These correlations are consistent with those found by Way (1990), who noted that correlations between TWE scores and TOEFL Section 2 scores were generally lower for examinees from selected Asian language groups than for other examinees.

Content-related Evidence. As a test of the ability to compose in standard written English, TWE uses writing

tasks similar to those required of college and university students in North America. As noted earlier, the TWE Committee develops items/prompts to meet detailed specifications that encompass widely recognized components of written language facility. Thus, each TWE item is constructed by subject-matter experts to assess the various factors that are generally considered crucial components of written academic English. Each item is pretested, and results of each pretested item are evaluated by the TWE Committee to ensure that the item is performing as anticipated. Items that do not perform adequately in a pretest are not used for the TWE test.

Finally, the actual scoring of TWE essays is done by qualified readers who have experience teaching English writing to native and nonnative speakers of English. The TWE readers are guided in their ratings by the TWE Scoring Guide and the standardized training and scoring procedures used at each TWE essay reading.

Performance of TWE Reference Groups

Table 4 presents the overall frequency distribution of TWE scores based on the 10 administrations from August 1993 through May 1995.

Table 5 lists the mean TWE scores for examinees tested at the 10 administrations, classified by native language. Table 6 lists the mean TWE scores for examinees classified by native country. These tables may be useful in comparing the test performance of a particular student with the average performance of other examinees who are from the same country or who speak the same native language.

It is important to point out that the data do not permit any generalizations about differences in the English writing proficiency of the various national and language groups. The tables are based simply on the performance of those examinees who have taken the TWE test. Because different selective factors may operate in different parts of the world to determine who takes the test, the samples on which the tables are based are not necessarily representative of the student populations from which the samples came. In some countries, for example, virtually any high school, university, or graduate student who aspires to study in North America may take the test. In other countries, government regulations permit only graduate students in particular areas of specialization, depending on national interests, to do so.

Table 4
Frequency Distribution of TWE Scores for All Examinees
(Based on 607,350 examinees who took the TWE test from August 1993 through May 1995)

TWE Score	N	Percent	Percentile Rank
6.0	6,503	1.07	99.47
5.5	10,258	1.69	98.09
5.0	47,166	7.77	93.36
4.5	51,264	8.44	85.25
4.0	203,620	33.53	64.28
3.5	89,266	14.70	40.16
3.0	138,013	22.72	21.45
2.5	31,202	5.14	7.52
2.0	22,993	3.79	3.06
1.5	3,687	0.61	0.87
1.0	3,378	0.56	0.28

Table 5

TWE Score Means –– All Examinees Classified by Native Language[1]

(Based on 594,536 examinees[2] who took the TWE test from August 1993 through May 1995)

Language	N	Mean
Afrikaans	295	3.72
Akan	336	4.54
Amharic	835	3.56
Arabic	22,969	3.46
Armenian	255	3.76
Assamese	129	3.99
Azerbaijani	103	3.78
Bashkir	3	—
Basque (Euskara)	52	4.08
Belorussian	59	3.90
Bemba	46	4.34
Bengali	5,594	4.11
Berber	61	3.43
Bikol	39	4.04
Bulgarian	1,444	4.20
Burmese	593	3.67
Catalan (Provencal)	332	3.95
Cebuano (Visayan)	488	4.05
Chichewa	142	4.48
Chinese	163,728	3.69
Chuvash	2	—
Czech	829	4.10
Danish	740	4.27
Dutch	1,397	4.30
Efik-Ibibio	55	4.50
English	3,726	4.64
Estonian	142	4.12
Ewe	160	4.44
Farsi (Persian)	3,002	3.52
Fijian	21	—
Finnish	1,122	4.20
French	13,161	3.97
Fula (Peulh)	91	3.70
Ga	97	4.59
Galician	25	—
Ganda (Luganda)	96	4.66
Georgian	190	3.54
German	12,710	4.29
Greek	6,277	3.92
Guarani	7	—
Gujarati	3,020	4.26
Hausa	94	4.27
Hebrew	1,549	4.01
Hindi	6,823	4.74
Hungarian (Magyar)	1,252	4.24
Ibo (Igbo)	489	4.51
Icelandic	394	4.09
Ilocano	156	3.94
Indonesian	15,508	3.55
Italian	4,997	3.77
Japanese	120,746	3.36
Javanese	796	3.44
Kannada (Kanarese)	1,396	4.66
Kanuri	7	—
Kashmiri	84	4.55
Kazakh	124	3.70
Khalkha (Mongolian)	83	3.46
Khmer (Kampuchean)	187	3.70
Kikuyu	923	4.71
Kirundi	44	3.85
Konkani	326	4.90
Korean	53,128	3.29
Kurdish	63	3.56
Kurukh (Oraon)	3	—
Kusaiean	29	—
Lao	151	3.69
Latvian	119	3.96
Lingala	83	3.69
Lithuanian	288	3.95
Luba-Lulua	27	—
Luo	295	4.76
Madurese	47	3.56
Malagasy	63	3.67
Malay	9,812	4.11
Malayalam	1,394	4.70
Malinke-Bambara-Dyula	191	3.66
Maltese	16	—
Marathi	1,358	4.90
Marshallese	87	3.49
Mende	39	4.42
Minankabau	20	—
More	16	—
Nepali	1,202	4.17
Norwegian	1,278	4.17
Nyanja	27	—
Oriya	203	4.63
Oromo (Galla)	63	3.76
Palauan	92	3.82
Panay-Hilligaynon	145	4.20
Pidgin	64	4.54
Polish	2,407	4.03
Ponapean	32	3.91
Portuguese	5,589	3.77
Punjabi	1,394	4.36
Pushtu	274	4.11
Romanian	1,463	4.16
Ruanda	55	4.02
Russian	7,009	3.90
Samar-Leyte	44	4.18
Samoan	237	4.13
Santali	2	—
Serbo-Croatian	1,497	3.97
Sesotho	97	4.56
Setswana	320	4.52
Shona	254	4.84
Sindhi	406	4.37
Sinhalese	1,023	4.25
Siswati	57	4.58
Slovak	413	4.01
Somali	187	3.79
Spanish	30,657	3.87
Sundanese	102	3.59
Swahili	749	4.41
Swedish	2,507	4.11
Tagalog	3,201	4.27
Tamil	5,108	4.63
Tatar	12	—
Telugu	6,386	4.36
Thai	30,074	3.20
Tibetan	64	4.29
Tigrinya	187	3.72
Tongan	15	—
Trukese	137	3.47
Tulu	82	4.82
Turkish	8,953	3.88
Turkmen	9	—
Twi-Fante	178	4.60
Ukrainian	776	3.91
Ulithian	7	—
Urdu	7,902	4.21
Vietnamese	2,852	3.70
Wolof	315	3.59
Xhosa	43	4.74
Yapese	16	—
Yiddish	8	—
Yoruba	703	4.65
Zulu	80	4.82

[1] Because of the unreliability of statistics based on small samples, means are not reported for groups with fewer than 30 examinees.
[2] Excludes 12,814 examinees who did not specify native language.

Table 6
TWE Score Means –– Nonnative English-Speaking Examinees Classified by Country[1]

(Based on 597,526 examinees[2] who took the TWE test from August 1993 through May 1995)

Country*	N	Mean
Afghanistan	241	3.58
Albania	175	3.94
Algeria	313	3.46
American Samoa	254	4.09
Andorra	11	—
Angola	50	3.73
Anguilla	4	—
Antigua and Barbuda	3	—
Argentina	1,566	3.94
Armenia	150	3.68
Aruba	69	3.79
Australia	73	3.98
Austria	974	4.35
Azerbaijan	196	3.63
Azores	5	—
Bahamas	6	—
Bahrain	255	3.59
Bangladesh	4,318	3.88
Barbados	6	—
Belarus	293	3.88
Belgium	729	4.27
Belize	31	3.95
Benin	53	3.96
Bermuda	6	—
Bhutan	19	—
Bolivia	633	3.85
Bosnia and Herzegovina	262	3.99
Botswana	300	4.47
Brazil	5,056	3.75
British Virgin Islands	11	—
Brunei Darussalam	43	4.12
Bulgaria	1,455	4.20
Burkina Faso	45	3.74
Burundi	49	3.84
Cambodia	205	3.73
Cameroon	297	4.01
Canada	1,395	3.99
Cape Verde	43	3.67
Cayman Islands	6	—
Central African Republic	8	—
Chad	30	3.53
Chile	599	3.79
China, People's Republic of	79,461	3.77
Colombia	3,547	3.82
Comoros	7	—
Congo	58	3.69
Costa Rica	577	4.10
Cote D'Ivoire (Ivory Coast)	316	3.72
Croatia	424	4.02
Cuba	179	3.76
Cyprus	2,671	3.83
Czech Republic	460	4.09
Czech and Slovak Federal Republic	573	4.10
Denmark	747	4.28
Djibouti	13	—
Dominican Republic	485	3.69
Ecuador	1062	3.89
Egypt	2,981	3.74
El Salvador	429	4.05
England	94	4.16
Equatorial Guinea	7	—
Eritrea	59	3.69
Estonia	161	4.06
Ethiopia	1,014	3.60
Federated States of Micronesia	220	3.65
Fiji	56	4.39
Finland	1,167	4.20
Former Yugoslav Republic of Macedonia	117	3.92
France	9,935	3.98
French Guiana	4	—
French Polynesia	10	—

Country*	N	Mean
Gabon	37	3.97
Gambia	163	4.06
Georgia	239	3.58
Germany	10,688	4.29
Ghana	723	4.63
Greece	4,054	3.97
Greenland (Kalaallit Nunaat)	5	—
Guadaloupe	47	3.79
Guam	16	—
Guatemala	891	4.01
Guinea	121	3.37
Guinea-Bissau	28	—
Guyana	5	—
Haiti	515	3.55
Honduras	397	3.92
Hong Kong	27,299	3.79
Hungary	1,110	4.24
Iceland	409	4.10
India	27,937	4.61
Indonesia	15,725	3.55
Iran	3,005	3.51
Iraq	361	3.64
Ireland	3	—
Israel	1,841	3.95
Italy	4,889	3.77
Jamaica	14	—
Japan	122,537	3.36
Jordan	3,472	3.52
Kazakhstan	263	3.81
Kenya	2,095	4.71
Kiribati	3	—
Korea (DPR)	921	3.30
Korea (ROK)	52,944	3.29
Kuwait	2,300	3.00
Kyrgyzstan	56	3.66
Laos	167	3.69
Latvia	283	3.93
Lebanon	4,492	3.71
Lesotho	40	4.30
Liberia	122	3.98
Libya	242	3.43
Liechtenstein	10	—
Lithuania	312	3.97
Luxembourg	68	4.01
Macau	1,225	3.72
Madagascar	69	3.96
Madeira Islands	2	—
Malawi	174	4.49
Malaysia	15,567	4.05
Maldives	17	—
Mali	188	3.64
Malta	27	—
Mariana Islands	31	3.92
Marshall Islands	92	3.50
Martinique	37	3.74
Mauritania	26	—
Mauritius	180	4.72
Mexico	8,411	3.81
Moldova	240	3.81
Monaco	14	—
Mongolia	85	3.52
Morocco	1,238	3.56
Mozambique	63	3.67
Myanmar	611	3.67
Namibia	24	—
Nauru	6	—
Nepal	1,197	4..19
Netherlands	993	4.25
Netherlands Antilles	52	4.13
New Caledonia	67	3.75
New Zealand	21	—
Nicaragua	458	3.81

[1] Because of the unreliability of statistics based on small samples, means are not reported for groups with fewer than 30 examinees.
[2] Excludes 6,098 examinees who did not specify country.

Table 6 (continued)

Country*	N	Mean
Niger	45	3.72
Nigeria	1,591	4.57
Niue Island	4	—
Northern Ireland	5	—
Norway	1,292	4.17
Oman	456	3.56
Pakistan	8,141	4.17
Panama	684	3.87
Papua New Guinea	55	4.36
Paraguay	156	3.74
Peru	1,972	3.77
Philippines	4,159	4.24
Poland	2,442	4.03
Portugal	472	3.94
Puerto Rico	1,572	4.11
Qatar	252	3.28
Reunion	24	—
Romania	1,525	4.17
Russia	4,697	3.93
Rwanda	69	4.09
San Marino	3	—
Sao Tome and Principe	8	—
Saudi Arabia	3,247	3.10
Scotland	3	—
Senegal	373	3.45
Seychelles	2	—
Sierra Leone	120	4.31
Singapore	1,424	4.51
Slovak Republic	226	4.03
Slovenia	96	4.28
Solomon Islands	7	—
Somalia	174	3.70
South Africa	374	4.70
Spain	4,633	3.94
Sri Lanka	1,814	4.14
St. Lucia	3	—
Sudan	518	3.69
Suriname	49	3.91
Swaziland	68	4.57
Sweden	2,525	4.10
Switzerland	1,721	4.14
Syria	1,298	3.29
Tahiti	26	—
Taiwan	47,839	3.47
Tajikistan	32	3.81
Tanzania	382	4.21
Thailand	30,210	3.20
Togo	92	3.77
Tonga	19	—
Trinidad and Tobago	6	—
Tunisia	321	3.53
Turkey	8,576	3.88
Turkmenistan	13	—
USSR	34	3.71
Uganda	274	4.67
Ukraine	1925	3.90
United Arab Emirates	1,221	3.35
United Kingdom	21	—
United States of America	1,009	4.13
Uruguay	144	4.11
Uzbekistan	257	3.50
Venezuela	2,744	3.82
Vietnam	2,979	3.71
Wales	4	—
Western Samoa	6	—
Yemen	310	3.41
Yugoslavia	988	3.90
Zaire	244	3.77
Zambia	123	4.50
Zimbabwe	290	4.83

[1] Because of the unreliability of statistics based on small samples, means are not reported for groups with fewer than 30 examinees.
[2] Excludes 6,098 examinees who did not specify country.

Table 7
TWE Score Means –– Applicants to Undergraduate and Graduate Programs

(Based on 518,671 examinees who took the TWE test from August 1993 through May 1995)

	N	Mean
Undergraduate	217,644	3.69
Graduate	301,027	3.72

Table 7 shows the mean TWE scores and numbers of examinees who indicated that they were taking the TWE test for admission to undergraduate or graduate degree programs. As the table indicates, there was no substantial difference between the performance of self-identified undergraduate and graduate applicants. Zwick and Thayer (1995) found, however, that after matching undergraduate and graduate examinees on TOEFL total score, undergraduate TWE means were higher than graduate means in 63 of 66 data sets analyzed.

Of the 301,027 examinees who indicated that they were applying to graduate programs, 146,299 requested at the time of testing that their scores be sent to specific graduate departments in the United States and Canada. Table 8 shows the mean TWE scores for examinees who requested that their scores be sent to graduate departments in the United States and Canada, classified by major field of study. (Examinees who requested their scores be sent to programs in more than one major field were classified on the basis of the first department code they specified.) The mean TWE score for this subgroup of graduate applicants was 3.99, which is somewhat higher than the mean for the total group of graduate applicants in Table 7. Within this subgroup, there were notable differences among departments. On average, applicants in the humanities and social sciences scored higher than applicants in biological and physical sciences.

Table 8
TWE Score Means –– Graduate School Applicants Classified by Department*

(Based on 146,299 examinees who took the TWE test from August 1993 through May 1995)

Department	N	Mean
Humanities		
Archaeology	90	3.79
Architecture	1,464	3.98
Art History	184	3.95
Classical Languages	46	4.27
Comparative Literature	335	4.21
Dramatic Arts	158	3.95
English	1,460	4.04
Far Eastern Languages and Literature	192	3.93
Fine Arts, Art, Design	1,594	3.66
French	165	4.20
German	115	4.36
Linguistics	688	4.15
Music	975	3.54
Near Eastern Languages and Literature	78	3.98
Philosophy	263	4.10
Religious Studies or Religion	605	4.08
Russian/Slavic Studies	100	4.02
Spanish	178	4.16
Speech	38	3.96
Other Foreign Languages	118	3.78
Other Humanities	324	3.99
Social Sciences		
American Studies	147	4.09
Anthropology	313	4.04
Business and Commerce	9,988	3.95
Communications	1,364	4.00
Economics	4,273	4.03
Education (including M.A. in Teaching)	2,383	4.03
Educational Administration	449	3.92
Geography	400	3.97
Government	824	4.21
History	410	4.05
Industrial Relations and Personnel	147	3.97
International Relations	1,304	4.13
Journalism	607	4.08
Library Science	414	4.97
Physical Education	304	3.59
Planning (City, Community, Urban, Regional)	359	4.04
Psychology, Clinical	338	4.18
Psychology, Educational	263	4.02
Psychology, Experimental/ Developmental	163	4.07
Psychology, Social	212	4.00
Psychology, Other	440	4.14
Public Administration	443	4.02
Public Health	778	3.97
Social Work	307	3.99
Sociology	617	4.06
Other Social Sciences	666	3.98

Department	N	Mean
Biological Sciences		
Agriculture	1,249	3.83
Anatomy	149	3.80
Audiology	44	3.81
Bacteriology	55	3.79
Biochemistry	2,042	4.01
Biology	1,624	3.91
Biomedical Sciences	753	4.07
Biophysics	204	4.02
Botany	300	3.87
Dentistry	829	3.98
Entomology	148	3.83
Environmental Science	840	3.92
Forestry	387	3.83
Genetics	447	4.02
Home Economics	135	3.80
Hospital & Health Services Administration	209	3.87
Medicine	2,141	3.95
Microbiology	1,051	4.03
Molecular and Cellular Biology	1,184	4.06
Nursing	1,085	3.73
Nutrition	653	3.87
Occupational Therapy	89	3.95
Pathology	356	3.93
Pharmacy	1,657	4.02
Physical Therapy	344	3.96
Physiology	416	3.89
Speech-Language Pathology	83	4.30
Veterinary Medicine	309	3.92
Zoology	185	3.91
Other Biological Sciences	771	4.00
Physical Sciences		
Applied Mathematics	527	3.93
Astronomy	140	3.93
Chemistry	5,573	3.90
Computer Sciences	10,824	4.05
Engineering, Aeronautical	764	4.07
Engineering, Chemical	3,138	4.18
Engineering, Civil	4,244	4.02
Engineering, Electrical	10,287	4.12
Engineering, Industrial	1,825	4.09
Engineering, Mechanical	6,285	4.01
Engineering, Other	3,723	3.99
Geology	910	3.91
Mathematics	2,270	3.89
Metallurgy	455	4.08
Oceanography	205	3.82
Physics	3,460	3.96
Statistics	657	3.89
Other Physical Sciences	607	3.90
Other Departments	20,019	3.83

* Because of the unreliability of statistics based on small samples, means are not reported for groups with fewer than 30 examinees.

TWE RESEARCH

Ongoing research studies related to the TWE test continue to address issues of importance to the TWE program. This research, reviewed and approved by outside specialists from the academic and testing communities, is essential to continual evaluation and improvement of the technical quality and utility of the test. To date 11 TWE-related research projects have been completed, and two projects are in progress; others are under consideration. The results of research efforts are published as reports and are available to anyone interested in the TWE program by writing to TOEFL Research Reports (L03), P.O. Box 6161, Princeton, NJ 08541-6161.

Research Reports Available (by date of completion)

- ***Survey of Academic Writing Tasks Required of Graduate and Undergraduate Foreign Students.*** Brent Bridgeman and Sybil Carlson. September 1983. TOEFL Research Report 15. This report describes a survey of faculty in 190 departments at 34 US and Canadian universities with high foreign student enrollments; respondents indicated a desire to use scores on a direct writing sample to supplement admissions and placement decisions.

- ***Relationship of Admissions Test Scores to Writing Performance of Native and Nonnative Speakers of English.*** Sybil Carlson, Brent Bridgeman, Roberta Camp, and Janet Waanders. August 1985. TOEFL Research Report 19. This study investigated the relationship between essay writing skills and scores on the TOEFL test and the GRE General Test obtained from applicants to US institutions.

- ***A Preliminary Study of the Nature of Communicative Competence.*** Grant Henning and Eduardo Cascallar. February 1992. TOEFL Research Report 36. This study was conducted to survey the theoretical literature related to communicative competence; to identify major variables said to comprise the construct(s); to test for comparative presence and measurability of such variables as in typical native/nonnative speaker university academic communication; to propose a tentative model of communicative competence as a synthesis of these variables; and to examine the relationship of TOEFL, TSE®, and TWE scores with the various elements of the tentative model. Results provide information on the comparative contributions of some theory-based communicative competence variables to domains of linguistic, discourse, sociolinguistic, and strategic competencies. In turn, these competency domains were investigated for their relation to components of language proficiency as assessed by the TOEFL, TSE, and TWE tests. Twelve oral and 12 written communication tasks were also analyzed and rank ordered for suitability in eliciting communicative language performance.

- ***An Investigation of the Appropriateness of the TOEFL Test as a Matching Variable to Equate TWE Topics.*** Gerald DeMauro. May 1992. TOEFL Research Report 37. This study explored the feasibility of using linear and equipercentile equating methods to equate forms of the TWE test by using the TOEFL test as an anchor. 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 TOEFL and TWE do not measure the same skills and the 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.

- ***Scalar Analysis of the Test of Written English.*** Grant Henning. August 1992. TOEFL Research Report 38. This study investigated the psychometric characteristics of the TWE rating scale employing Rasch model scalar analysis methodology with more than 4,000 scored TWE essays across two prompts. Results suggest that the intervals between TWE scale steps were surprisingly uniform, and the size of the intervals was appropriately larger than the error associated with assignment of individual ratings. The proportion of positively misfitting essays was small and approximately equal to the proportion of essays that required adjudication by a third reader. This latter finding, along with the low proportion of misfitting readers detected, provides preliminary evidence of the feasibility of employing Rasch rating scale analysis methodology for the equating of TWE essays prepared across prompts.

- ***Effects of Amount of Time Allowed on the Test of Written English.*** Gordon Hale. June 1992. TOEFL Research Report 39. This study examined students' performance on TWE prompts under two time limits – 30 minutes, as on the current TWE, and 45 minutes. Mean scores on the six-point TWE scale were found to be significantly higher by about 1/4 to 1/3 point under the 45-minute condition, indicating that allowing additional time produced a modest but reliable increase in scores. The magnitude of the effect was roughly comparable for students of low versus high proficiency, and for students in intensive English programs versus students in academic coursework. The correlation between scores for both time conditions was relatively high; both parallel-form reliability and interrater reliability were approximately the same for the two time conditions.

Provision of additional time apparently had little effect on the relative standing of students on the test. Results are discussed in relation to the literature on time effects and to practical implications for the TWE test.

- ***Topic and Topic Type Comparability on the Test of Written English.*** Marna Golub-Smith, Clyde Reese, and Karin Steinhaus. March 1993. TOEFL Research Report 42. This study addressed the question of how comparable scores are for TWE essays written on different topics and/or different topic types, particularly compare-contrast and chart-graph topic types. It compared TWE mean scores across eight equivalent groups of examinees in an operational TWE administration and also reported on differences observed across prompts in the number of examinees at each score level. Additional analyses by gender were also conducted.

- ***A Comparison of Performance of Graduate and Undergraduate School Applicants on the Test of Written English.*** Rebecca Zwick and Dorothy T. Thayer. May 1995. TOEFL Research Report 50. The performance of graduate and undergraduate school applicants on the Test of Written English was compared for each of 66 data sets, dating from 1988 to 1993. The analyses compared the average TWE score for graduates and undergraduates after matching examinees on the TOEFL total score. The main finding was that, for matched examinees, undergraduate TWE means were higher than graduate means in 63 of the 66 data sets. Although these standardized mean differences (SMDs) never exceeded 0.3 of a TWE score point (with standard errors that were typically between 0.01 and 0.02), the results are noteworthy because they give a different picture than do simple comparisons of means for unmatched graduates and undergraduates, which showed higher mean TWE scores for graduate applicants in the majority of cases.

- ***Reader Calibration and Its Potential Role in Equating for the Test of Written English.*** Carol Myford, Diana Marr, and J. Michael Linacre. Spring 1996. TOEFL Research Report 52. When judges use a rating scale to rate performances, some may rate more severely than others, giving lower ratings. Judges may also differ in the consistency with which they apply rating criteria. This study pilot tested a quality control procedure that provides a means for monitoring and adjusting for differences in reader performance. FACETS, a Rasch-based rating scale analysis procedure, was employed to calibrate readers within and across two TWE administrations. The study had four general foci: (1) to determine the extent to which individual readers can be considered interchangeable, both within and across TWE administrations; (2) to investigate reader characteristics and their relationships to the volume and quality of ratings; (3) to examine the efficacy of the use of a third reading to adjudicate rating discrepancies; and (4) to make a preliminary determination of the feasibility of using FACETS Reader Severity Measures as a first step toward equating TWE scores across different topics.

- ***A Study of Writing Tasks Assigned in Academic Degree Programs.*** Gordon Hale, Carol Taylor, Brent Bridgeman, Joan Carson, Barbara Kroll, and Robert Kantor. Spring 1996. TOEFL Research Report 54. Writing tasks assigned in 162 undergraduate and graduate courses in several disciplines at eight universities were collected. Using a sample of the assignments, key dimensions of difference were identified, and a classification scheme based on those dimensions was developed. Application of the classification scheme provided data on the prevalence of various types of assignments and, for essay tasks, showed the degree to which the assignments were characterized by each of several features. Differences among the kinds of writing tasks assigned in different groups of disciplines were examined.

- ***Adjustment for Reader Rating Behavior in the Test of Written English.*** Nicholas T. Longford. Spring 1996. TOEFL Research Report 55. This report evaluated the impact of a potential scheme for score adjustment using data from the administrations of the Test of Written English in 1994. It is shown that, assuming noninformative assignment of readers to essays, the adjustment due to reader differences would reduce the mean squared error for all essays except those graded by readers with small workloads. The quality of the rating process as described by the variances due to true scores, severity, and inconsistency, as well as the distribution of workloads was similar across the administrations. This would enable a reliable prediction of the optimal score adjustment in future administrations. Two approximations to the optimal adjustment are proposed, and an array of diagnostic procedures for the engaged raters are presented. The report highlights the relevance of shrinkage estimators to problems in which a large number of quantities is to be estimated and indicates how combining information across rating exercises could lead to further gains in the precision of assigned scores.

Research in Progress

- ***Computer Analysis of the Test of Written English.*** Lawrence Frase and Joseph Faletti with consultants Doug Biber, Ulla Connor, Gerard Dalgish, and Joy Reid. Seeks to conduct a variety of automated text analyses of TWE essays to summarize, analyze, and compare linguistic properties of TWE essays written by examinees from different language groups and to determine how TWE scores relate to linguistic text properties. Database of analyzed essays is now being used in other studies.
- ***Reliability Study of the Test of Written English Using Generalizability Theory.*** Gwyneth Boodoo. Investigates use of generalizability theory (G-theory) to explore and develop methods for estimating the reliability of the TWE test; will take into account sources of variation in scores associated with the fact that different pairs of readers rate different subsets of papers within a prompt as well as variation associated with the use of different prompts.

Other Relevant Documents

- Educational Testing Service. 1987. ***ETS Guidelines for Developing and Scoring Free-Response Tests.*** Princeton, NJ.
- Educational Testing Service. 1987. ***ETS Standards for Quality and Fairness.*** Princeton, NJ.
- Livingston, S. A., and Zieky, M. J. 1982. ***Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests.*** Educational Testing Service: Princeton, NJ.

REFERENCES

American Educational Research Association, American Psychological Association, and National Council for Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

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

Bridgeman, B., and Carlson, S. (1983). *Survey of academic writing tasks required of graduate and undergraduate foreign students* (TOEFL Research Report No. 15). Princeton, NJ: Educational Testing Service.

Carlson, S. B., Bridgeman, B., Camp, R., and Waanders, J. (1985). *Relationship of admission test scores to writing performance of native and nonnative speakers of English* (TOEFL Research Report No. 19). Princeton, NJ: Educational Testing Service.

Conlan, G. (1976). *Suggestions for writing essay questions.* Princeton, NJ: Educational Testing Service.

Educational Testing Service. (1987). *ETS guidelines for developing and scoring free-response tests***.** Princeton, NJ: Author.

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

Educational Testing Service. (1991). *TWE reading management guidelines*. Princeton, NJ: Author.

Hale, G. A., and Hinofotis, F. (1981). *New directions in English language testing.* Internal report submitted to the TOEFL Research Committee. Princeton, NJ: Educational Testing Service.

Kane, H. (1983). *A study of practices and needs associated with intensive English language programs: report of findings.* Internal report submitted to the TOEFL Program Office. New York: Kane, Parsons, and Associates, Inc.

Oller, J. W. (1979). *Language tests at school.* London: Longman Group Ltd.

Pike, L. (1976). *An evaluation of alternate item formats for testing English as a foreign language* (TOEFL Research Report No. 2). Princeton, NJ: Educational Testing Service.

Pitcher, B., and Ra, J. B. (1967). *The relationship between scores on the Test of English as a Foreign Language and ratings of actual theme writing* (Statistical Report 67-9). Princeton, NJ: Educational Testing Service.

Way, W. D. (1990). *TOEFL 2000 and Section II: Relationships between structure, written expression, and the Test of Written English* (Internal Report, March 1990). Princeton, NJ: Educational Testing Service.

Zwick, R., and Thayer, D. T. (1995). *A comparison of performance of graduate and undergraduate school applicants on the Test of Written English* (TOEFL Research Report No. 50). Princeton, NJ: Educational Testing Service.

APPENDIX A

TWE COMMITTEE MEMBERS*
(1995-96)

Louis Arena	University of Delaware
Dwight Atkinson	Auburn University
Diane Belcher	Ohio State University
Cherry Campbell	Monterey Institute of International Studies
Joan Carson	Georgia State University
Melinda Erickson	University of California, Berkeley
Dennie Rothschild	Vancouver Community College

FORMER MEMBERS

George Braine	University of South Alabama
Milton Clark	California State University, San Bernardino
Ulla Connor	Purdue University
William Gaskill	San Diego State University
Lynn Goldstein	Monterey Institute of International Studies
Roseann Duenas Gonzales	University of Arizona, Tucson
Kay Grandage	North York Board of Education, Ontario
Robert Kantor	Ohio State University
Jane Hughey	Texas A & M University
Barbara Kroll	California State University, Northridge
Ilona Leki	University of Tennessee
Vivian McDonough	University of Toronto
Joy Reid	University of Wyoming
Marian Tyacke	University of Toronto

TWE CHIEF READERS
(1994-95)

Mary Bly	University of California, Davis
Milton Clark	California State University, San Bernardino
Lynn Goldstein	Monterey Institute of International Studies
John White	California State University, Fullerton
Agnes Yamada	California State University, Dominguez Hills

***Formerly known as the TWE Core Reader Group**

APPENDIX B

TEST OF WRITTEN ENGLISH (TWE) SCORING GUIDE

Revised 2/90

Readers will assign scores based on the following scoring guide. Though examinees are asked to write on a specific topic, parts of the topic may be treated by implication. Readers should focus on what the examinee does well.

Scores

6 **Demonstrates clear competence in writing on both the rhetorical and syntactic levels, though it may have occasional errors.**
A paper in this category
—effectively addresses the writing task
—is well organized and well developed
—uses clearly appropriate details to support a thesis or illustrate ideas
—displays consistent facility in the use of language
—demonstrates syntactic variety and appropriate word choice

5 **Demonstrates competence in writing on both the rhetorical and syntactic levels, though it will probably have occasional errors.**
A paper in this category
—may address some parts of the task more effectively than others
—is generally well organized and developed
—uses details to support a thesis or illustrate an idea
—displays facility in the use of language
—demonstrates some syntactic variety and range of vocabulary

4 **Demonstrates minimal competence in writing on both the rhetorical and syntactic levels.**
A paper in this category
—addresses the writing topic adequately but may slight parts of the task
—is adequately organized and developed
—uses some details to support a thesis or illustrate an idea
—demonstrates adequate but possibly inconsistent facility with syntax and usage
—may contain some errors that occasionally obscure meaning

3 **Demonstrates some developing competence in writing, but it remains flawed on either the rhetorical or syntactic level, or both.**
A paper in this category may reveal one or more of the following weaknesses:
—inadequate organization or development
—inappropriate or insufficient details to support or illustrate generalizations
—a noticeably inappropriate choice of words or word forms
—an accumulation of errors in sentence structure and/or usage

2 **Suggests incompetence in writing.**
A paper in this category is seriously flawed by one or more of the following weaknesses:
—serious disorganization or underdevelopment
—little or no detail, or irrelevant specifics
—serious and frequent errors in sentence structure or usage
—serious problems with focus

1 **Demonstrates incompetence in writing.**
A paper in this category
—may be incoherent
—may be undeveloped
—may contain severe and persistent writing errors

Papers that reject the assignment or fail to address the question must be given to the Table Leader. Papers that exhibit absolutely no response at all must also be given to the Table Leader.

APPENDIX C

TWE TEST BOOK COVER

Form: 3RTF12

Topic **A**

Test of Written English

TWE® Test Book

Do NOT open this test book until you are told to do so.

Read the directions that follow.

1. The TWE essay question is printed on the inside of this test book. You will have **30 minutes** to plan, write, and make any necessary changes to your essay. Your essay will be graded on its overall quality.
2. Read the topic carefully. You may want to read it more than once to be sure you understand what you are asked to write about.
3. Think before you write. Making notes may help you to organize your essay. Below the essay topic is a space marked NOTES. You may use this area to outline your essay or make notes.
4. Write only on the topic printed on the inside. If you write on a different topic, your essay will not be scored. Write clearly and precisely. How well you write is much more important than how much you write, but to cover the topic adequately, you may want to write more than one paragraph.
5. Start writing your essay on the first line of Side 3 of the TWE answer sheet. Use Side 4 if you need more space. Extra paper will not be provided. Write neatly and legibly. Do not skip lines. Do not write in very large letters or leave large margins.
6. Check your work. Allow a few minutes **before** time is called to read over your essay and make any changes.
7. After 30 minutes, you will be instructed to stop and put your pencil down. You MUST stop writing. **If you continue to write, it will be considered cheating.**

Do NOT break the seal on this book until you are told to do so.

When you have finished reading the directions, look up.