

TOEFL[®]

TEST OF ENGLISH AS
A FOREIGN LANGUAGE

COMPUTER-
BASED
TOEFL

SCORE
USER
GUIDE

www.toefl.org



*Educational
Testing Service*



Listening • Structure • Reading • Writing



2000-2001
EDITION

The *Computer-Based TOEFL Score User Guide* was prepared for deans, admissions officers and graduate department faculty, administrators of scholarship programs, ESL instructors, foreign student advisers, overseas advisers, and others who have an interest in the TOEFL test. The *Guide* provides information specifically about the computer-based TOEFL test: the new format, the new score scale and information about the interpretation of the new scores, the administration of the test, and program research activities and new testing developments.

In July of 1998, the computer-based TOEFL test was introduced in many areas of the world. It was introduced in Asia in October 2000, and will become available in the People's Republic of China at a later date. Thus, institutions will receive scores from both the paper- and computer-based tests.

The information in this *Guide* is designed to supplement information provided in the 1997 edition of the *TOEFL Test and Score Manual* and the 1999-00 edition of the *TOEFL Test and Score Data Summary*, both of which refer specifically to the paper-based test. More information about score interpretation for the computer-based test will appear on the TOEFL Web site at www.toefl.org as it becomes available. To receive electronic updates on the computer-based TOEFL test, join our Internet mailing list by completing the requested information at www.toefl.org/cbtindex.html. To be added to our mailing list, fill out the form on page 43.

*TOEFL Programs and Services
International Language Programs
Educational Testing Service*



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

Copyright © 2000 by Educational Testing Service. All rights reserved.

EDUCATIONAL TESTING SERVICE, ETS, the ETS logos, GRADUATE RECORD EXAMINATIONS, GRE, SLEP, SPEAK, THE PRAXIS SERIES: PROFESSIONAL ASSESSMENTS FOR BEGINNING TEACHERS, TOEFL, the TOEFL logo, TSE, and TWE are registered trademarks of Educational Testing Service.

The Test of English as a Foreign Language, Test of Spoken English, and Test of Written English are trademarks of Educational Testing Service.

COLLEGE BOARD and SAT are registered trademarks of the College Entrance Examination Board.

GMAT and GRADUATE MANAGEMENT ADMISSION TEST are registered trademarks of the Graduate Management Admission Council.

Prometric is a registered trademark of Thomson Learning.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the publisher. Violators will be prosecuted in accordance with both United States and international copyright and trademark laws.

Permissions requests may be made online at www.toefl.org/copyright.html or sent to

Proprietary Rights Office
Educational Testing Service
Rosedale Road
Princeton, NJ 08541-0001, USA
Phone: 1-609-734-5032

Contents

Overview	4	Additional Score Reports	15
The Test of English as a Foreign Language	4	Confidentiality of TOEFL Scores	15
Educational Testing Service	4	Calculation of TOEFL Scores	16
TOEFL Board	4	Rescoring Essays	16
		Scores of Questionable Validity	17
		Examinees with Disabilities	17
TOEFL Developments	5		
Evolution of the Computer-Based TOEFL Test	5	Use of TOEFL Test Scores	18
The Switch to Computer-Based Testing	6	Who Should Take the TOEFL Test?	18
The Added Value of the Computer-Based TOEFL Test	6	How to Use the Scores	18
Computer Familiarity and Test Performance	6	Preparing Your Institution for Computer-Based	
Test Center Access	6	TOEFL Test Scores	19
		Standard Setting	19
Description of the Computer-Based TOEFL Test ..	7	Staff Training	19
Test Design	7	Faculty Training	19
Linear Testing	7	Advising Students	20
Computer-Adaptive Testing	7	Institutional Publications and Web Sites	20
Timing of the Test	7	Guidelines for Using TOEFL Test Scores	20
Tutorials	8		
Listening Section	8	Statistical Characteristics of the Test	24
Structure Section	8	The Computer-Based Test	24
Reading Section	8	The Computer-Based Population Defined	24
Writing Section	9	Intercorrelations Among Scores	24
New Score Scales	9	The New Scores	25
		Calculation of the Structure/Writing Score	25
Administration of the Computer-Based		Adequacy of Time Allowed	26
TOEFL Test	10	Essay Data	27
Where the Test Is Offered	10	Reliabilities	27
Test Security	10	The Standard Error of Measurement	27
Procedures at Test Centers	10	Reliability of Gain Scores	28
Identification Requirements	10	Validity	29
Checking Names	11		
Photo Score Reporting	11	References	32
Testing Irregularities	11	Appendix A: Standard-setting Procedures,	
Preventing Access to Test Items	11	Concordance Tables	34
		Appendix B: Database Modification Options	37
TOEFL Test Results	12	Appendix C: Essay Ratings	39
Release of Test Results	12	Appendix D: Computer-familiarity Studies	40
Test Score Data Retention	12	Where to Get More TOEFL Information	42
Image Score Reports	12	Order Form	43
Official Score Report from ETS	12		
Features of the Image Reports	12		
Information in the Official Score Report	14		
Examinee's Score Record	14		
Acceptance of Test Results Not Received from ETS	14		
How to Recognize an Unofficial Score Report	14		

Overview

The Test of English as a Foreign Language

The purpose of the Test of English as a Foreign Language (TOEFL®) is to evaluate the English proficiency of people whose native language is not English. The test was initially developed to measure the English proficiency of international students wishing to study at colleges and universities in the United States and Canada, and this continues to be its primary function. A number of academic institutions in other countries, as well as certain independent organizations, agencies (including medical certification and licensing agencies), and foreign governments, have also found the test scores useful. Overall, more than 4,200 institutions, agencies, and organizations in over 80 countries use TOEFL test scores.

A National Council on the Testing of English as a Foreign Language was formed in 1962, composed of representatives of more than 30 private organizations and government agencies concerned with the English proficiency of foreign nonnative speakers of English who wished to study at colleges and universities in the United States. The Council supported the development of the TOEFL test for use starting in 1963-64. Financed by grants from the Ford and Danforth Foundations, the program was, at first, attached administratively to the Modern Language Association. In 1965, the College Entrance Examination Board and Educational Testing Service® (ETS®) assumed joint responsibility for the program. Because many who take the TOEFL test are potential graduate students, a cooperative arrangement for the operation of the program was entered into by Educational Testing Service, the College Board®, and the Graduate Record Examinations® (GRE®) Board in 1973. Under this arrangement, ETS is responsible for administering the TOEFL program according to policies determined by the TOEFL Board.

TOEFL-related tests include

- the Test of Written English (TWE®), a writing assessment administered with the paper-based TOEFL test
- the Test of Spoken English (TSE®), which provides a reliable measure of proficiency in spoken English
- SPEAK® (Speaking Proficiency English Assessment Kit), an institutional form of the TSE test
- the Institutional Testing Program (ITP), which permits approved institutions to administer previously used forms of the paper-based TOEFL test on dates convenient for them using their own facilities and staff

- the Secondary Level English Proficiency (SLEP®) test, designed for students entering grades 7 through 12.

Educational Testing Service

ETS is a nonprofit organization committed to the development and administration of testing programs; the creation of advisory and instructional services; research on techniques and uses of measurement, human learning, and behavior; and educational development and policy formation. In addition to developing tests, it supplies related services; for example, it scores the tests; records, stores, and reports test results; performs validity and other statistical studies; and undertakes program research. All ETS activities are governed by a 16-member board of trustees composed of persons from various fields, including education and public service.

In addition to the Test of English as a Foreign Language and the Graduate Record Examinations, ETS develops and administers a number of other tests, including the Graduate Management Admission Test® (GMAT®), The Praxis Series: Professional Assessments for Beginning Teachers®, and the College Board SAT® and Achievement tests.

The Chauncey Group International Ltd., a wholly owned subsidiary of ETS, provides assessment, training, and guidance products and services in the workplace, military, professional, and adult educational environments.

TOEFL Board

Policies governing the TOEFL program are formulated by the 15-member TOEFL Board. The College Board and the GRE Board each appoint three members to the Board. These six members constitute the Executive Committee and elect the remaining nine members. Some Board members are affiliated with such institutions and agencies as secondary schools, undergraduate and graduate schools, community colleges, nonprofit educational exchange organizations, and other public and private agencies with an interest in international education. Other members are specialists in the field of English as a foreign or second language.

The Board has six standing committees, each responsible for specific areas of program activity: the Committee of Examiners, Finance Committee, Grants and Awards Committee, Nominating Committee, Committee of Outreach and Services, and TSE Committee. There are also several ad hoc committees: Access, Community College Constituents Committee, Products and Services, Secondary School Advisory Group, and Technology.

TOEFL Developments

Evolution of the Computer-Based TOEFL Test

In recent years, various constituencies,¹ including TOEFL committees and score users, have called for a new TOEFL test that (1) is more reflective of communicative competence models; (2) includes more constructed-response tasks and direct measures of writing and speaking; (3) includes tasks that integrate the language modalities tested; and (4) provides more information than current TOEFL scores do about the ability of international students to use English in an academic environment. Accordingly, in 1993, the TOEFL Board initiated the TOEFL 2000 project, a broad effort to further strengthen TOEFL's validity. The introduction of the computer-based TOEFL test is the first incremental step in this broad test-improvement effort.

The impetus for the redesigned TOEFL is drawn from several sources. Many in the language teaching and testing communities associate the TOEFL test with discrete-point testing, which is based on the structuralist, behaviorist model of language learning and testing. Discrete-point tests contain items that target only one element of a skill, such as grammar or vocabulary.² Some teachers of English as a second language (ESL) and English as a foreign language (EFL) are concerned that discrete-point test items, and the exclusive use of traditional, multiple-choice items to assess receptive skills, have a negative impact on instruction. Although direct measures of speaking and writing abilities appeared in the 1980s in the TOEFL program's TSE and TWE tests, these have been much less widely used than the TOEFL test.

Because the TOEFL test is frequently used in making admissions decisions about international students, the motivation to revise the test has been strong. In 1995, certain aspects of the test were changed as a step toward a more integrative approach to language testing. For example, single-statement listening comprehension items were eliminated, the academic lectures and longer dialogues were increased in number, and vocabulary tasks were embedded in reading comprehension passages. Still, these changes reflected relatively minor progress toward an integrative approach to language testing.

¹ Primary constituencies are score users from the North American higher education admissions community, applied linguists, language testers, and second language teachers. While fewer in number, other users of TOEFL scores represent a diverse range of groups: public and private high schools, overseas colleges and universities, embassies, foundations and commissions, medical and professional boards, government agencies, language institutes, and a small number of private businesses.

² As defined by Carroll, 1961, and Oller, 1979.

Recently, in an extensive effort to address the concerns mentioned above, TOEFL staff undertook several parallel, interrelated efforts with the advice and ongoing review of the TOEFL Board and committees. Specifically, project staff systematically considered three broad areas: the needs of test users, feasible technology relevant to test design and international delivery, and test constructs. With respect to test users, TOEFL staff profiled examinees and score users, conducted a number of score user focus groups and surveys, and prepared reports on trends in international student admissions and intensive English program enrollments. These activities helped to clarify and elaborate on constituents' concerns and needs. With respect to technology, staff examined existing and developing technologies available worldwide, and anticipated technological developments that could facilitate implementation of computer-based testing.

The majority of their initial efforts, however, focused on test constructs and the development of prototype tests. Project staff systematically reviewed the literature on communicative competence and communicative language testing.

The project continues with efforts designed to establish a solid foundation for the next generation of computer-based TOEFL tests. Project teams of ETS test developers, researchers, and external language testing experts have produced framework documents for assessing reading, writing, listening, and speaking as both independent and interdependent skills. These frameworks lay out (1) how respective domains will be defined, taking into account recent research and thinking among the language learning and testing communities; (2) what the operational constraints for delivering the test internationally will be; (3) what research is needed to refine and validate the frameworks; and (4) what criteria should be used to judge whether the project produces a better test.

Prototyping trials of test tasks, item types, and scoring rubrics are currently being carried out. The trials will provide evidence of the redesigned test's measurement claims and purpose. This effort will be followed by further trials during which performance definitions will be refined at the test-form level.

A compilation of related research and development reports is available in a new monograph series and can be ordered through *The Researcher* newsletter (see page 43) or the TOEFL Web site at www.toefl.org/edpubs.html#researcher.

The Switch to Computer-Based Testing (CBT)

By introducing computer-based testing, ETS is taking a critical step toward a long-term goal of enhancing its assessments by using electronic technology to test more complex skills. Institutions can now obtain scores faster, an advantage they have long sought. This initiative has not been undertaken casually. Rather, it began in the 1980s and continued into the 1990s as, one by one, tests such as GRE, the Praxis Series for teacher licensing, the national test for licensing nurses, and GMAT were computerized. These tests and the TOEFL test are currently administered by a global network of computerized testing centers.

Today, the growing popularity of distance learning has raised the prospect of global testing. This prospect will achieve fruition with developments in such technologies as information security, encryption, and transmission speed. Thus, the day when ETS assessments are delivered via the Internet may be imminent.

In 1995, the TOEFL program decided to introduce an enhanced, computer-based test in 1998. New test design features were identified, computer-based prototypes were created and piloted, a study of TOEFL examinees' computer familiarity and performance on computer-based test tasks was completed,³ and implementation plans were developed. In 1996, the development of the very large pools of test items needed to deliver a computer-based test worldwide began.

The initial pretesting of the computer-based questions was conducted in phases between the spring of 1997 and the spring of 1998 with test takers who had previously taken the paper-based TOEFL test. Pretest questions are included in the operational test, as they have always been with the paper-based test.

The Added Value of the Computer-Based TOEFL Test

With computer-based testing, the TOEFL program is endeavoring to provide more extensive information about candidates' English proficiency than it has in the past. In response to institutions' requests to include a productive writing measure, the program added a Writing section as part of each test administration. This addition is one step toward a more communicative test. Essay ratings are integrated into section and total scores, but are also reported separately on official score reports for informational purposes. New types of questions have been added to the Listening and Reading sections; these new question types move beyond

³Eignor, Taylor, Kirsch, & Jamieson, 1998; Kirsch, Jamieson, Taylor, & Eignor, 1998; Taylor, Jamieson, Eignor & Kirsch, 1998.

single-selection multiple-choice questions. Visuals have also been added to the Listening section, providing a significant enhancement to that portion of the test. Because two sections (Listening and Structure) are computer-adaptive, the test is tailored to each examinee's performance level. (See pages 7-9 for more information on test enhancements.)

Computer Familiarity and Test Performance

Developing the computer-based TOEFL test led to questions of access and equity, especially considering that examinees with little or no prior computer experience would be taking the test with examinees who were highly familiar with this technology. In 1995, the TOEFL Research Committee (now part of the Committee of Examiners) recommended that the TOEFL program undertake formal research to ascertain empirically whether variation in familiarity with computers would affect test takers' performance on a computer-delivered test.

Given this concern that measurement of examinees' English proficiency might be confounded by their computer proficiency, a group of researchers conducted a two-phase study of (1) TOEFL examinees' access to and familiarity with computers and (2) the examinees' performance on a set of computerized TOEFL test items following a computer-based tutorial especially designed for this population. Examinees who had previously taken the paper-based TOEFL were invited to participate in the study, which was conducted in the spring to early fall of 1996. The results of the study indicated that there was no meaningful relationship between computer familiarity and the examinees' performance on computer-based test questions once they had completed the computer tutorial. For further discussion of the study, see Appendix D.

Test Center Access

All ETS's computer-based tests are offered at Prometric® Testing Centers, at computer test centers at specified colleges and universities, at selected USIS posts and advising centers overseas, and at ETS offices in the United States. Both permanent CBT and paper-based centers are used to meet the needs of international test takers. Permanent centers are established in large population areas. In low-volume testing areas, supplemental paper-based testing is provided.

A network of Regional Registration Centers (RRCs) disseminates information about ETS's computer-based tests and handles test appointments overseas. The RRC list and a list of test centers are printed in the *TOEFL Information Bulletin for Computer-Based Testing*. Updates to the *Bulletin* test center list can be found on the TOEFL Web site at www.toefl.org/cbt_tclandfees.html.

Description of the Computer-Based TOEFL Test

Test Design

The process of creating a test and its individual items is determined by the overall test design, which is much like a blueprint. Test design determines the total number and types of items asked, as well as the subject matter presented. TOEFL test design specifies that each item be coded for content and statistical characteristics. Content coding assures that each examinee will receive test questions that assess a variety of skills (e.g., comprehension of a main idea, understanding of inferences) and cover a variety of subject matter (topics for lectures and passages). Statistical characteristics, including estimates of item difficulty and the ability of an item to distinguish (or discriminate) between higher or lower ability levels, are also coded.

The TOEFL test utilizes two types of computer-based testing: linear and computer-adaptive. Two of the sections (Listening and Structure) are computer-adaptive, and one section (Reading) is linear.

The TOEFL program is often asked to describe the language skills associated with certain TOEFL scores. An assumption of such questions is often that the TOEFL test and its score scale are based on formal criteria of what it means to know a language. One can easily imagine a test in which the designer would say: an examinee who can do tasks a, b, and c has achieved a certain level of competency in the language. However, the TOEFL is not designed to be that kind of test. It is a norm-referenced test, which means that it was designed to compare individuals in their English language proficiencies.

Linear Testing

In a linear test, examinees are presented with questions that cover the full range of difficulty (from easy to difficult) as well as the content specifications designated by the test design. In the Reading section, the computer selects for each examinee a combination of passages with accompanying sets of questions that meet both the content and the statistical designs of the test. The questions are selected without consideration of examinee performance on the previous questions. The result is a section much like the one on the paper-based test, but each examinee receives a unique set of passages and questions.

Computer-Adaptive Testing

A computer-adaptive test is tailored to an individual examinee. Each examinee receives a set of questions that

meet the test design and are generally appropriate for his or her performance level.

The computer-adaptive test starts with questions of moderate difficulty. As examinees answer each question, the computer scores the question and uses that information, as well as the responses to previous questions, to determine which question is presented next. As long as examinees respond correctly, the computer typically selects a next question of greater or equal difficulty. In contrast, if they answer a question incorrectly, the computer typically selects a question of lesser or equal difficulty. The computer is programmed to fulfill the test design as it continuously adjusts to find questions of appropriate difficulty for test takers of all performance levels.

In other words, in a computer-adaptive test, the computer is programmed to estimate an examinee's ability and chooses items that will provide the most information to refine the ability estimate.

In computer-adaptive tests, only one question is presented at a time. Because the computer scores each question before selecting the next one, examinees must answer each question as it is presented. For this reason, examinees cannot skip questions, and once they have entered and confirmed their answers, they cannot return to questions; nor can they return to any earlier part of the test. In the linear Reading section, however, examinees are allowed to skip questions and return to previously answered questions.

Timing of the Test

The overall administration time is approximately 4 hours, which includes time to view unofficial scores, to choose score recipients, and to answer a short posttest questionnaire. However, this administration time may be shorter in many cases because examinees can advance through the test at their own pace. The chart below outlines the time limits and number of questions for each part of the test.

Computer-Based TOEFL Test Format		
Test Portion	# Questions	Time Limit
Tutorials	7 Tutorials	Untimed
Listening	30-50	40-60 minutes
Structure	20-25	15-20 minutes
BREAK		5 minutes
Reading	44-55	70-90 minutes
Writing	1 prompt	30 minutes

All test takers obtain test scores based on the same number of questions. Because pretest questions used for research purposes may be randomly distributed in some test forms, the number of questions and the times allowed for each section may vary. However, the number of questions and the total amount of time allowed for a section is stated in the on-screen directions.

Tutorials

The computer-based TOEFL test begins with tutorials that show examinees how to take the test using basic computer tools and skills (e.g., use a mouse to point, click, and scroll). Tutorials at the beginning of the first three sections (Listening, Structure, and Reading) demonstrate how to answer the questions in those sections. These tutorials are mandatory but untimed. There is also an untimed Writing tutorial for those who plan to type their essays.

Listening Section

The Listening section measures the examinee’s ability to understand English as it is spoken in North America. Conversational features of the language are stressed, and the skills tested include vocabulary and idiomatic expression as well as special grammatical constructions that are frequently used in spoken English. The stimulus material and questions are recorded in standard North American English.

This section includes various stimuli, such as dialogues, short conversations, academic discussions, and minilectures, and poses questions that test comprehension of main ideas, the order of a process, supporting ideas, important details, and inferences, as well as the ability to categorize topics/objects. The section consists of 30-50 questions and is 40-60 minutes in length.

Type of Stimulus	Number of Stimuli	Number of Questions
Dialogues	11-17	1 each
Short Conversations	2-3	2-3 each
Minilectures/ Academic Discussions	4-6	3-5 each

The test developers have taken advantage of the multimedia capability of the computer by using photos and graphics to create context and support the content of the minilectures, producing stimuli that more closely approximate “real world” situations in which people do more than just listen to voices. The listening stimuli are often

accompanied by either context-setting or content-based visuals. All dialogues, conversations, academic discussions, and minilectures include context visuals to establish the setting and role of the speakers.

Four types of questions are included in the Listening section: (1) traditional multiple-choice questions with four answer choices; (2) questions that require examinees to select a visual or part of a visual; (3) questions for which examinees must select two choices, usually out of four; and (4) questions that require examinees to match or order objects or text.

The Listening section is computer-adaptive. In addition, there are other new features in this section: test takers control how soon the next question is presented, they have headphones with adjustable volume control, and they both see and hear the questions before the answer choices appear.

Structure Section

The Structure section measures an examinee’s ability to recognize language that is appropriate for standard written English. The language tested is formal, rather than conversational. The topics of the sentences are associated with general academic discourse so that individuals in specific fields of study or from specific national or linguistic groups have no particular advantage. When topics have a national context, it is United States or Canadian history, culture, art, or literature. However, knowledge of these contexts is not needed to answer the questions.

This section is also computer-adaptive, with the same two types of questions used on the paper-based TOEFL test. These are questions in which examinees must (1) complete an incomplete sentence using one of four answers provided and (2) identify one of four underlined words or phrases that would not be accepted in English. The two question types are mixed randomly rather than being separated into two subsections as in the paper-based TOEFL test. There are 20-25 questions in this section, which is 15-20 minutes long.

Reading Section

The Reading section measures the ability to read and understand short passages similar in topic and style to academic texts used in North American colleges and universities. Examinees read a variety of short passages on academic subjects and answer several questions about each passage. Test items refer to what is stated or implied in the passage, as well as to words used in the passage. To avoid creating an advantage for individuals in any one field of

study, sufficient context is provided so that no specific familiarity with the subject matter is required to answer the questions.

The questions in this section assess the comprehension of main ideas, inferences, factual information stated in a passage, pronoun referents, and vocabulary (direct meaning, synonym, antonym). In all cases, the questions can be answered by reading and understanding the passages. This section includes: (1) traditional multiple-choice questions; (2) questions that require examinees to click on a word, phrase, sentence, or paragraph to answer; and (3) questions that ask examinees to “insert a sentence” where it fits best.

The Reading section includes 44-55 questions and is 70-90 minutes long. The section consists of four to five passages of 250-350 words, with 11 questions per passage. The Reading section is not computer-adaptive, so examinees can skip questions and return to previous questions.

Writing Section

The Writing section measures the ability to write in English, including the ability to generate, organize, and develop ideas, to support those ideas with examples or evidence, and to compose a response to one assigned topic in standard written English. Because some examinees may not be accustomed to composing an essay on computer, they are given the choice of handwriting or typing the essay in the 30-minute time limit. Topics that may appear on the test are published in the *TOEFL Information Bulletin for Computer-Based Testing* and on the TOEFL Web site at www.toefl.org/tstprpmt.html. The rating scale used to score the essay is published in the *Bulletin*, on the TOEFL Web site, and on score reports. Also see Appendix C.

The essay in the Writing section is scored by trained readers in the ETS Online Scoring Network (OSN). Prospective TOEFL readers are trained to (1) interpret TOEFL standards, (2) score across multiple topics, and (3) use the OSN software. At the conclusion of training, prospective readers take an online certification test. Those who pass the test can be scheduled to score operational essays.

Certified readers are scheduled to score essays on specific days and times and are always monitored by a scoring leader, who provides guidance and support. After logging into the scoring network, readers must first score a set of calibration essays, to ensure that they are scoring accurately. Both typed essays and handwritten essay images are displayed on-screen, and readers enter their ratings by clicking on the appropriate score points. Support materials (which include the scoring

guide and training notes) are available online at all times during scoring. Each essay is rated independently by two readers. Neither reader knows the rating assigned by the other.

An essay will receive the average of the two ratings unless there is a discrepancy of more than one point: in that case, a third reader will independently rate the essay. The essay will thus receive a final rating of 6.0, 5.5, 5.0, 4.5, 4.0, 3.5, 3.0, 2.5, 2.0, 1.5, or 1. A score of 0 is given to papers that are blank, simply copy the topic, are written in a language other than English, consist only of random keystroke characters, or are written on a topic different from the one assigned.

The essay rating is incorporated into the Structure/Writing scaled score, and constitutes approximately 50 percent of that combined score. (See page 25 for information about score calculation.) The rating is also reported separately on the official score report to help institutions better interpret examinees' Structure/Writing scores.

New Score Scales

New score scales for the computer-based TOEFL test have been introduced because of the addition of the essay and the new types of questions in the Listening and Reading sections. The paper-based score scales have been truncated at the lower end to prevent overlap and to further differentiate between the scales for the two tests. These differences immediately identify which type of test the examinee has taken. (See page 13 for an example of a computer-based TOEFL score report.)

Computer-Based TOEFL Test Score Scale	
Section	Score Scale
Listening	0 to 30
Structure/Writing	0 to 30
Reading	0 to 30
Total Score	0 to 300

Paper-Based TOEFL Test Score Scale	
Section	Score Scale
Listening Comprehension	31 to 68
Structure/Written Expression	31 to 68
Reading Comprehension	31 to 67
Total Score	310 to 677

Administration of the Computer-Based TOEFL Test

Where the Test Is Offered

The computer-based TOEFL test, offered in many locations since July 1998, was introduced in most of Asia in October 2000. The paper-based TOEFL is administered in the People's Republic of China. Supplemental paper-based testing and the Test of Written English (TWE) are offered in areas where computer-based testing is not available. The Test of Spoken English (TSE) continues to be administered worldwide. In addition, the Institutional Test Program (ITP) will continue to be paper-based for the next several years.

Test Security

In administering a worldwide testing program in more than 180 countries, ETS and the TOEFL program consider the maintenance of security at testing centers to be of utmost importance. The elimination of problems at test centers, including test taker impersonations, is a continuing goal.

To offer score users the most valid, and reliable measurements of English language proficiency available, the TOEFL office continuously reviews and refines procedures to increase the security of the test before, during, and after administrations.

Procedures at Test Centers

Standard, uniform procedures are important in any testing program, but are essential for an examination that is given worldwide. Each test center is staffed by trained and certified test center administrators. Working with test center administrators, ETS ensures that uniform practices are followed at all centers. Test administrators are instructed to exercise extreme vigilance to prevent examinees from giving or receiving assistance in any way. Two test administrators are present at every test center, and observation windows at permanent test centers enable them to observe the test takers at all times. In addition, there is video monitoring of the examinees at each test center.

To prevent copying from notes or other aids, examinees may not have anything on their desks during the first half of the test. After the break, a center administrator provides examinees with scratch paper. Official answer sheets are given out at the start of the Writing section if examinees choose to handwrite their essays.

A test center administrator who is certain that an examinee gave or received assistance has the authority to dismiss the individual from the test center. Scores for dismissed examinees will not be reported. If a test administrator suspects behavior that may lead to an invalid score, he or she submits an electronic "irregularity report" to ETS. Both suspected and confirmed cases of misconduct are investigated by the Test Security Office at ETS.

Identification Requirements

Strict admission procedures are followed at test centers to prevent attempts by some examinees to have others with higher English proficiency impersonate them at a TOEFL administration. To be admitted to a test center, every examinee must present an official identification document with a recognizable photograph. Although a passport is acceptable at all test centers, other specific photobearing documents may be acceptable for individuals who do not have passports because, for example, they are taking the test in their own countries.

Examinees are told what form of identification is needed in the *Information Bulletin for Computer-Based Testing*;⁴ the information is also confirmed when they call to schedule an appointment or send in their registration forms. Test center administrators must follow the TOEFL identification requirement policy and procedures and **cannot make exceptions**. Examinees will not be admitted if they do not have the proper identification or if the validity of their identification is questionable.

Through embassies in the United States and ETS representatives, the TOEFL program office continually verifies the names of official, secure, photobearing identification documents used in all countries/areas in which the test is given.

⁴The *Information Bulletin for Computer-Based Testing* explains the procedures a candidate should follow to make an appointment for the test, and lists required fees, test center locations, and identification requirements. It also provides information about the tutorials that are included in every test session, as well as sample questions to help familiarize candidates with the computerized format of the test. The *Bulletin* may be downloaded from the TOEFL Web site www.toefl.org/infobull.html. It can also be ordered by calling 1-609-771-7100.

Checking Names

To further prevent impersonation, examinees are required to write their signatures in a sign-in log. These names and signatures are compared to those on the official identification documents as well as to the names on the appointment record. Examinees must also sign their names when they leave or reenter the testing room, and their signatures are compared each time.

Photo Score Reporting

As an additional procedure to help eliminate the possibility of impersonation at test centers, the official score reports routinely sent to institutions and the examinee's own copy of the score report bear an electronically reproduced photo image of the examinee. Examinees are advised in the *Information Bulletin* that the score reports will contain these photo images. Key features of the image score reports are discussed on page 12. In addition to strengthening security through this deterrent to impersonation, the report form provides score users with the immediate information they may need to resolve issues of examinee identity. If there is a clear discrepancy in photo identification for test takers with multiple test dates, ETS will cancel the test scores.

Testing Irregularities

"Testing irregularities" refers to irregularities in connection with the administration of a test, such as equipment failure, improper access to test content by individuals or groups of test takers, and other disruptions of test administrations (natural disasters and other emergencies). When testing irregularities occur, ETS gives affected test takers the opportunity to take the test again as soon as possible without charge.

Preventing Access to Test Items

A number of measures are taken to ensure the security of test items. First, very large question pools are created. The computer selects different questions for each examinee according to the examinee's ability level. Once the predetermined exposure rate for a question is reached, that question is removed from the pool. All questions are encrypted and decrypted only on the screen at the test center; examinee responses are also encrypted and decrypted only once they reach ETS. Data are stored in servers that can only be accessed by the certified test center administrators.

TOEFL Test Results

Release of Test Results

After finishing the test, examinees view their unofficial test scores on the Listening and Reading sections. Because the essay will not yet have been read and rated, examinees can see only ranges for the Structure/Writing scaled score and the total score. Once test takers have reviewed their scores, they can decide whether they want their official scores sent to up to four institutions/agencies or they can decide to cancel their scores, in which case, the scores are not reported to the test takers or any institutions. Score reports are typically mailed within two weeks of the test date if the essay was composed on the computer. If the essay was handwritten, score reports are mailed approximately five weeks after the test date.

Each examinee is entitled to five copies of the test results: one examinee's score record is sent to the examinee,⁵ and up to four official score reports are sent by ETS directly to the institutions designated by the examinee as recipients. A list of the most frequently used institutions and agencies is printed in the *Information Bulletin*.

The most common reason that institutions do not receive score reports is that examinees do not properly identify institutions as score report recipients by choosing them from a list provided on the computer (i.e., because the examinees are not familiar with the spelling of the names of those institutions). An institution whose code number is not listed in the *Bulletin* should give applicants the spelling of its official name so that they can indicate it accurately at the test center.

Test Score Data Retention

Language proficiency can change considerably in a relatively short period. Therefore, individually identifiable TOEFL scores are retained on the TOEFL database for only two years from the test date and scores more than two years old are not reported. Individuals who took the TOEFL test more than two years ago must take it again if they want scores sent to an institution. While all information that could be used to identify an individual is removed from the database after two years, anonymous score data and other information that can be used for research or statistical purposes are retained.

⁵ The test score is not the property of the examinee. A TOEFL score is measurement information and subject to all the restrictions noted in this *Guide*. (These restrictions are also noted in the *Bulletin*.)

Image Score Reports

The image-processing technology used to produce official score reports allows ETS to capture the examinee's image, as well as other identifying data submitted by the examinee at the testing site, and to reproduce it, together with the examinee's test results, directly on score reports. If a photograph is so damaged that it cannot be accepted by the image-processing system, "Photo Not Available" is printed on the score report. Steps have been taken to minimize tampering with examinee score records that are sent directly to applicants. To be sure of receiving valid score records, however, admissions officers and others responsible for the admissions process should accept only official score reports sent directly from ETS.

Official Score Report from ETS

TOEFL score reports⁶ give the score for each of the three test sections, the total score, and, in a separate field, the essay rating, which has already been incorporated into the Structure/Writing and total scaled scores. See page 14 for the computer-based TOEFL score report codes.

Features of the Image Reports

1. The blue background color quickly identifies the report as being an official copy sent from ETS.
2. The examinee's name and scores are printed in red fields.
3. The examinee's photo, taken on the day of the test administration, is reproduced on the score report.
4. A red serial number, located in the center, bleeds through to the back of the score report as a security measure to prevent tampering.
5. To distinguish the official score reports for the computer-based TOEFL from those for the paper-based TOEFL, a computer icon and the words "Computer-Based Test" are printed on the top of the reports.
6. A rule across the bottom of the score report contains microprinting that spells out "TOEFL." It can be seen on the original form by using a magnifying glass. When photocopied, the line will appear to be solid.

⁶ TOEFL Magnetic Score Reporting Service provides a convenient method of entering TOEFL scores into an institution's fields. The TOEFL program offers examinee score data on disk, cartridge, or magnetic tape twice a month. See page 43 to order a request form for this service. Institutions can also download scores electronically via the Internet. For updates on this service, visit the TOEFL Web site at www.toefl.org/edservcs.html.

Information in the Official Score Report

In addition to test scores, native country, native language, and birth date, the score report includes other pertinent data about the examinee and information about the test.

INSTITUTION CODE. The institution code provided on score reports designates the recipient college, university, or agency. A list of the most frequently used institution and agency codes is printed in the *Bulletin*. An institution that is not listed should give applicants the spelling of its official name so that they can indicate it at the test center. (This information should be included in application materials prepared for international students.)

Note: An institution that does not know its TOEFL code number or wishes to obtain one should call 1-609-771-7975 or write to ETS Code Control, P.O. Box 6666, Princeton, NJ 08541-6666, USA.

DEPARTMENT CODE. The department code number identifies the professional school, division, department, or field of study in which the applicant plans to enroll. The department code list shown below is also included in the graduate *Bulletin*. The department code for all business schools is (02), for law schools (03), and for unlisted departments (99).

Fields of Graduate Study Other Than Business or Law

HUMANITIES

- 11 Archaeology
- 12 Architecture
- 26 Art History
- 13 Classical Languages
- 28 Comparative Literature
- 53 Dramatic Arts
- 14 English
- 29 Far Eastern Languages and Literature
- 15 Fine Arts, Art, Design
- 16 French
- 17 German
- 04 Linguistics
- 19 Music
- 57 Near Eastern Languages and Literature
- 20 Philosophy
- 21 Religious Studies or Religion
- 22 Russian/Slavic Studies
- 23 Spanish
- 24 Speech
- 10 Other foreign languages
- 98 Other humanities

SOCIAL SCIENCES

- 27 American Studies
- 81 Anthropology
- 82 Business and Commerce
- 83 Communications
- 84 Economics
- 85 Education (including M.A. in Teaching)
- 01 Educational Administration
- 70 Geography
- 92 Government
- 86 History
- 87 Industrial Relations and Personnel
- 88 International Relations
- 18 Journalism
- 90 Library Science
- 91 Physical Education
- 97 Planning (City, Community, Regional, Urban)
- 92 Political Science
- 93 Psychology, Clinical
- 09 Psychology, Educational
- 58 Psychology, Experimental/Developmental
- 79 Psychology, Social
- 08 Psychology, other
- 94 Public Administration
- 50 Public Health
- 95 Social Work
- 96 Sociology
- 80 Other social sciences

BIOLOGICAL SCIENCES

- 31 Agriculture
- 32 Anatomy
- 05 Audiology
- 33 Bacteriology
- 34 Biochemistry
- 35 Biology
- 45 Biomedical Sciences
- 36 Biophysics
- 37 Botany
- 38 Dentistry
- 39 Entomology
- 46 Environmental Science
- 40 Forestry
- 06 Genetics
- 41 Home Economics
- 25 Hospital and Health Services Administration
- 42 Medicine
- 07 Microbiology
- 74 Molecular and Cellular Biology
- 43 Nursing
- 77 Nutrition
- 44 Occupational Therapy
- 56 Pathology
- 47 Pharmacy
- 48 Physical Therapy
- 49 Physiology
- 55 Speech-Language Pathology
- 51 Veterinary Medicine
- 52 Zoology
- 30 Other biological sciences

PHYSICAL SCIENCES

- 54 Applied Mathematics
- 61 Astronomy
- 62 Chemistry
- 78 Computer Sciences
- 63 Engineering, Aeronautical
- 64 Engineering, Chemical
- 65 Engineering, Civil
- 66 Engineering, Electrical
- 67 Engineering, Industrial
- 68 Engineering, Mechanical
- 69 Engineering, other
- 71 Geology
- 72 Mathematics
- 73 Metallurgy
- 75 Oceanography
- 76 Physics
- 59 Statistics
- 60 Other physical sciences

Use 99 for any department not listed.

Examinee's Score Record

Examinees receive their test results on a form entitled Examinee's Score Record. **These are NOT official score reports and should not be accepted by institutions.**

Acceptance of Test Results Not Received from ETS

Because some examinees may attempt to alter score records, institution and agency officials are urged to verify all TOEFL scores supplied by examinees. TOEFL/TSE Services will either confirm or deny the accuracy of the scores submitted by examinees. If there is a discrepancy between the official scores recorded at ETS and those submitted in any form by an examinee, the institution will be requested to send ETS a copy of the score record supplied by the examinee. At the written request of an official of the institution, ETS will report the official scores, as well as all previous scores recorded for the examinee within the last two years. Examinees are advised of this policy in the *Bulletin*, and, in signing their completed test scheduling forms, they accept these conditions. (Also see "Test Score Data Retention" on page 12.)


How to Recognize an Unofficial Examinee's Score Report

1. * * *Examinee's Score Record* * * appears at the top of the form.
2. An Examinee's Score Record is printed on white paper.
3. The last digit of the total score should end in 0, 3, or 7.
4. There should be no erasures, blurred areas, or areas that seem lighter than the rest of the form.

1

TOEFL[®]

**Computer-Based Test
Examinee's Score Record
for the Test of English as a Foreign Language**



4

Appointment Number 0000 0000 0123 4567			
NAME SALVAN KIM TORRES (Family or Surname, Given, Middle)			


02/27/1998 Month/Day/Year Test Date	S3225 Test Center Number	06/30/1963 Month/Day/Year Date of Birth	F Sex
1726 XXXX XXXX XXXX	00 XX XX	PHILIPPINES Native Country	
Institution Code		TAGALOG Native Language	

TOEFL Scaled Scores			
19	17	17	177
Listening	Structure/ Writing	Reading	Total Score
Essay Rating 3.0			

Examinee's Mailing Address:

SALVAN KIM TORRES
PO BOX 1234
ANYTOWN, OH 12345-6789

2



3

Test of English as a Foreign Language • PO Box 6151 • Princeton, NJ 08541-6151 • USA

Facsimile reduced.

DOs and DON'Ts

DO verify the information on an examinee's score record by calling TOEFL/TSE Services at 1-800-257-9547.

DON'T accept scores that are more than two years old.

DON'T accept score reports from other institutions that were obtained under the TOEFL Institutional Testing Program.

DON'T accept photocopies of score reports.

Additional Score Reports

TOEFL examinees may request that official score reports be sent to additional institutions at any time up to two years after their test date.

There are two score reporting services: (1) mail and (2) phone service (1-888-TOEFL-44 in the United States or 1-609-771-7267 outside the U.S.). Additional score reports (\$12 each) are mailed within two weeks after receipt of Score Report Request Form. For an additional fee of \$12, examinees can use the phone service to request that score reports be sent to institutions within four working days after a request is processed. See the TOEFL Web site at www.toefl.org/cbscrsvc.html#services for more information about TOEFL scores by phone.

Confidentiality of TOEFL Scores

Information retained in TOEFL test files about an examinee's native country, native language, and the institutions to which the test scores were sent, as well as the actual scores, is the same as the information printed on the examinee's score record and on the official score reports. An official score report will be sent only at the consent of the examinee to those institutions or agencies designated by the examinee on the day of the test, on a Score Report Request Form submitted at a later date, or otherwise specifically authorized by the examinee through the phone service.⁷

To ensure the authenticity of scores, the TOEFL program office urges that institutions accept only official copies of TOEFL scores received directly from ETS.

Score users are responsible for maintaining the confidentiality of an individual's privacy with respect to score information. Scores are not to be released by an institutional recipient without the explicit permission of the examinee. Dissemination of score records should be kept to a minimum, and all staff with access to them should be informed of their confidential nature.

⁷ Institutions or agencies that are sponsoring an examinee and have made prior arrangements with the TOEFL office by completing an Examinee Fee Voucher Request Form (see page 43 to order the form) will receive copies of examinees' official score reports if the examinees have given permission to the TOEFL office to send them.

The TOEFL program recognizes the right of institutions and individuals to privacy with respect to information stored in ETS files. The program is committed to safeguarding this information from unauthorized disclosure. As a consequence, information about an individual or institution will only be released by prior agreement, or with the explicit consent of the individual or institution.

Calculation of TOEFL Scores

The scoring for a computer-adaptive test or section is based on the difficulty level of the items (which is determined through pretesting), the examinee's performance on the items, and the number of items answered. The examinee typically gets more credit for correctly answering a difficult question than for correctly answering an easy question. The computer then uses the data gained from information about the items presented, and the item responses, to compute an estimate of the examinee's ability. This ability estimate is refined following each item response.

It is important to remember that the scoring of a computer-adaptive section is cumulative. If the last item of a section is relatively easy and is answered incorrectly, it does not mean the examinee will receive a low score. The computer considers the examinee's performance on all questions to determine the score.

For two test takers who have the same number of correct responses on a computer-adaptive test, generally the one who answers the more difficult questions correctly will receive the higher score. Similarly, for two test takers who answer questions of equivalent difficulty on average, the examinee who answers fewer questions will receive a lower score.

This is very different from the paper-based TOEFL test, in which scores are determined solely by the number of correct answers for a section and in which a correct answer to an easy question counts as much as a correct answer to a difficult question.

The Structure/Writing composite score is obtained by combining the Structure score with the essay rating, and then converting it to a scale score that consists of approximately 50 percent Structure and 50 percent Writing.

In the linear Reading test section, the computer selects questions on the basis of test design, not on the actual responses to the items presented. The scoring of this linear section takes into account the examinee's performance on the items, the number of items answered, and the difficulty level of the items answered. (See page 24 for a more technical explanation of score calculation procedures.)

TOEFL section scale scores are reported on a scale ranging from 0 to 30. TOEFL total scale scores are reported on a scale from 0 to 300. Paper-based total scale scores currently range from 310 to 677 to avoid overlap. Because of the calculation method used, the rounded total scores for both paper- and computer-based tests can only end in 0, 3, or 7.

In 1997-98, TOEFL completed a concordance study that established the relationship between scores on the paper- and computer-based tests. (See Appendix A for concordance tables.) The chart below shows actual ranges of observed scores for all sections of the computer-based test (except the essay) for all examinees tested between July 1999 and June 2000.

Minimum and Maximum Observed Section and Total Scores on Computer-Based TOEFL		
Section	Minimum	Maximum
Listening	0	30
Structure/Writing	0	30
Reading	1	30
Total Score	10	300

Rescoring Essays

Examinees who question the accuracy of reported essay ratings may request to have the essays rated again. Requests must be received within six months of the test date, and there is a fee for this service.

In the rereading process, the TOEFL essay is rated independently by two people who have not seen it previously. If the process confirms the original rating, the examinee is notified by letter. If the process results in a rating change, the Structure/Writing score and the total score will be affected, and as a result the examinee will receive a revised score record and a refund of the rescore fee. The revised rating becomes the official rating, and revised official score reports are sent to the institutions previously selected by the examinee. Experience has shown that very few score changes result from this procedure.

Scores of Questionable Validity

Examinees are likely to achieve improved scores over time if, between tests, they study English or increase their exposure to native spoken English. Thus, improvement in scores may not indicate an irregularity in the test itself or its administration. However, institutions and other TOEFL score recipients that note such inconsistencies as high TOEFL scores and apparent weak English proficiency should refer to the photo on the official score report for evidence of impersonation. Institutions should notify the TOEFL office if there is any such evidence or they believe the scores are questionable for other reasons.

Apparent irregularities, reported by institutions or brought to the attention of the TOEFL office by examinees or test center administrators who believe that misconduct has taken place, are investigated. Such reports are reviewed, statistical analyses are conducted, and scores may be canceled by ETS as a result. In some cases, the ETS Test Security Office assembles relevant documents, such as previous score reports, CBT Voucher Requests or International Test Scheduling Forms, and test center logs and videos. When evidence of handwriting differences or possible collaboration is found, the case is referred to the ETS Board of Review, a group of senior professional staff members. After an independent examination of the evidence, the Board of Review directs appropriate action.

ETS policy and procedures are designed to provide reasonable assurance of fairness to examinees in both the identification of suspect scores and the weighing of information leading to possible score cancellation. These procedures are intended to protect both score users and examinees from inequities that could result from decisions based on fraudulent scores and to maintain the test's integrity.

Examinees with Disabilities

The TOEFL program is committed to serving test takers with disabilities by providing services and reasonable accommodations that are appropriate given the purpose of the test. Some accommodations that may be approved are enlarged print or Braille formats, omission of the Listening section, a test reader, an amanuensis or keyboard assistant, other customarily used aids, sign language interpreter (for spoken directions only), a separate testing room, and extended time and/or rest breaks during the test administration. For those familiar with their use, a Kensington Trackball Mouse, Headmaster Mouse, Intellikeys Keyboard, and ZOOMTEXT can be made available. Security procedures are the same as those followed for standard administrations.

For certain special accommodations (e.g., extended time or omission of Listening), the test may not provide a valid measure of the test taker's proficiency, even though the testing conditions were modified to minimize adverse effects of the test taker's disability on test performance. Alternative methods of evaluating English proficiency are recommended for individuals who cannot take the test under standard conditions. Criteria such as past academic record (especially if English has been the language of instruction), recommendations from language teachers or others familiar with the applicant's English proficiency, and/or a personal interview or evaluation can often supplement TOEFL scores to give a fuller picture of a candidate's proficiency.

Because nonstandard administrations vary widely and the number of examinees tested under nonstandard conditions is small, the TOEFL program cannot provide normative data to interpret scores obtained in such administrations.

Use of TOEFL Test Scores

Who Should Take the TOEFL Test?

Most educational institutions at which English is the language of instruction require international applicants who are nonnative English speakers to provide evidence of their English proficiency prior to beginning academic work. TOEFL scores are frequently required for the following categories of applicants:

- Individuals from countries where English is one of the official languages, but not necessarily the first language of the majority of the population or the language of instruction at all levels of schooling. These countries include, but are not limited to, British Commonwealth and United States territories and possessions.
- Persons from countries where English is not a native language, even though there may be schools or universities at which English is the language of instruction.

NOTE: The TOEFL test is recommended for students at the eleventh-grade level or above; the test content is considered too difficult for younger students.

Many institutions report that they frequently do not require TOEFL test scores of certain kinds of international applicants. These include:

- Nonnative speakers who hold degrees or diplomas from postsecondary institutions in English-speaking countries (e.g., the United States, Canada, England, Ireland, Australia, New Zealand) and who have successfully completed at least a two-year course of study in which English was the language of instruction.
- Transfer students from institutions in the United States or Canada whose academic course work was favorably evaluated in relation to its demands and duration.
- Nonnative speakers who have taken the TOEFL test within the past two years and who have successfully pursued academic work at schools where English was the language of instruction in an English-speaking country for a specified period, generally two years.

How to Use the Scores

The TOEFL test is a measure of general English proficiency. It is not a test of academic aptitude or of subject matter competence; nor is it a direct test of English speaking ability.⁸ TOEFL test scores can help determine whether an applicant has attained sufficient proficiency in English to study at a college or university. However, even applicants who achieve high TOEFL scores may not succeed in a given program of study if they are not broadly prepared for academic work. Therefore, the admissibility of nonnative English speakers depends not only on their levels of English proficiency but also on other factors, such as their academic records, the schools they have attended, their fields of study, their prospective programs of study, and their motivations.

If a nonnative English speaker meets an institution's academic requirements, official TOEFL test scores may be used in making distinctions such as the following:

- The applicant may begin academic work with no restrictions.
- The applicant may begin academic work with some restrictions on the academic load and in combination with concurrent work in English language classes. This implies that the institution can provide the appropriate English courses to complement the applicant's part-time academic schedule.
- The applicant is eligible to begin an academic program within a stipulated period of time but is assigned to a full-time program of English study. Normally, such a decision is made when an institution has its own English as a second language (ESL) program.
- The applicant's official status cannot be determined until he or she reaches a satisfactory level of English proficiency. Such a decision will require that the applicant pursue full-time English training at the same institution or elsewhere.

The above decisions presuppose that an institution is able to determine what level of English proficiency is sufficient to meet the demands of a regular or modified program of study. Such decisions should never be based on TOEFL scores alone, but on an analysis of all relevant information available.

⁸ The Test of Spoken English was developed by ETS under the direction of the TOEFL Board and TSE Committee to provide a reliable measure of proficiency in spoken English. For more information on TSE, visit the TOEFL Web site at www.toefl.org/edabtse.html.

Preparing Your Institution for Computer-Based TOEFL Test Scores

It is important that each institution prepares to receive the new computer-based scores. Following are some suggestions for standard-setting, faculty and staff training, advising students, and updating institutional publications and Web sites.

Standard Setting

Institutional decision makers need to set new score standards for the computer-based TOEFL. Where departmental standards differ from institutional minimums, new cut-score ranges also should be established.

The following information should be consulted during the standard-setting process:

- Concordance tables and the standard error of measurement; see Appendix A
- Standard-setting suggestions in Appendix A. The concordance tables include range and point scores for the section scores as well as for the total score. When comparing paper- and computer-based scores, keep in mind that the range-to-range concordance table should be used to set cut-score ranges. In other words, long-standing advice against using absolute cut scores for admission should be observed.
- Guidelines for Using TOEFL Scores on pages 20-23.

Staff Training

Using this *Guide*, the *TOEFL Sampler* CD-ROM,⁹ and the *Information Bulletin for Computer-Based Testing*, your institution might want to familiarize admissions and advising staff with the changes to the test, the new score scale, and the concordance tables. They can also learn to recognize score reports for both the computer- and paper-based tests by looking at samples. Copies of these reports were sent to

⁹The *TOEFL Sampler* CD-ROM, designed by test developers at ETS, contains the computerized tutorials that all examinees take before starting the test as well as practice questions for each section. Animated lessons show test takers how to use a mouse, scroll, and use testing tools. Interactive test tutorials provide instructions for answering questions in the four sections of the test. The *Sampler* also includes 67 practice questions to help examinees become familiar with the directions, formats, and question types in the test. Copies of the CD-ROM can be ordered by calling 1-800-446-3319 (United States) 1-609-771-7243 (elsewhere) or by accessing the TOEFL Web site at www.toefl.org.

institutions in the summer of 1998; additional copies can be obtained by calling the computer-based TOEFL hotline at 1-609-771-7091.

To understand the reported scores on the new computer-based TOEFL test, a case study approach may be helpful. This approach involves asking staff members to review admissions profiles of typical applicants that include computer-based TOEFL test scores. These profiles could consist of either real but disguised or fabricated information. By reviewing them, staff members can become more familiar with the process of using computer-based TOEFL scores in combination with other information to make a decision about a candidate.

By the time the staff training is planned, your institution should have already set its new score requirements for the computer-based TOEFL test. (See “Standard Setting” above.) It is advisable that the new requirements be publicized on your Web site and in printed literature. Your staff members might be able to suggest other places where the information should be included or if there are any other departments or divisions that also need to receive training. In addition, it is important to provide training updates as needed.

Faculty Training

It is important to provide ongoing training for faculty involved in making decisions about applicants’ scores on the computer-based TOEFL test. Admissions and nonacademic advising staff can work with the faculty and academic advisers in the following ways.

- At department meetings, inform faculty members about the new test and provide them with copies of Appendices A, B, and C, and the guidelines on pages 20-23.
- Meet with academic advisers on a regular basis to clarify this information.
- Identify the success rates of students with computer-based scores, and compare these rates to those of students with paper-based scores.

Advising Students

Institutions can designate staff members to address student questions on the computer-based TOEFL test. A booklet entitled *Preparing Students for the Computer-Based TOEFL Test* can be found at www.toefl.org/tflannc.html.

Institutional Publications and Web Sites

All sources of public information, including catalogs, admission materials, departmental brochures, and Web sites, should be updated to include information on the computer-based TOEFL test and your institution's new score requirements. Institutions are encouraged to create links from their Web pages to key TOEFL Web site pages, including www.toefl.org/concords1.html and www.toefl.org/tflannc.html.

Guidelines for Using TOEFL Test Scores

As a result of ETS's general concern with the proper use of test data, the TOEFL program encourages institutions that use TOEFL scores to provide this *Guide* to those who use the scores and to advise them of changes in the test and performance data that affect interpretation. The TOEFL program disseminates information about test score use in conference presentations, regional meetings, and campus events. The TOEFL program also urges institutions to request the assistance of TOEFL staff when the need arises.

An institution that uses TOEFL test scores should be cautious in evaluating an individual's performance on the test and determining appropriate score requirements.

The following guidelines are intended to help institutions do that reasonably and effectively.

- Base the evaluation of an applicant's readiness to begin academic work on all available relevant information, not solely on TOEFL test scores.

The TOEFL test measures an individual's ability in several areas of English language proficiency, but it does not test that proficiency comprehensively. Nor does it provide information about scholastic aptitude or skills, motivation, language-learning aptitude, or cultural adaptability. An estimate of an examinee's proficiency can be fully established only with reference to a variety of measures, including the institution's informed sense of the proficiency needed to succeed in its academic programs. Many institutions utilize local tests, developed and administered in their English language programs, to supplement TOEFL results.

- Do not use rigid cut scores to evaluate applicants' performance on the TOEFL test.

Because test scores are not perfect measures of ability, the rigid use of cut scores should be avoided. The standard error of measurement should be understood and taken into consideration in making decisions about an individual's test performance or in establishing appropriate cut-score ranges for the institution's academic demands. Good practice includes looking at additional evidence of language proficiency (e.g., use of local tests, interviews) for applicants who score near the cut-score ranges. See Appendix A for information on the standard error of measurement for the computer-based TOEFL test.

- Take TOEFL section scores, as well as total scores, into account.

The total score on the TOEFL test is derived from scores on the three sections of the test. Though applicants achieve the same total score, they may have different section score profiles that could significantly affect subsequent academic performance. For example, an applicant with a low score on the Listening section but relatively high scores on the other sections may have learned English mainly through the written medium; this applicant might have more trouble in lecture courses than students with higher scores. Similarly, students who score high in Listening but low in Structure/Writing or Reading may have developed their proficiencies in nonacademic settings that required little literacy. Therefore, they are good candidates for compensatory courses in those skills. Applicants whose scores in Reading are much lower than their scores on the other two sections might be advised to take a reduced academic load or postpone enrollment in courses that involve a significant amount of reading. The information section scores yield can be used to advise students about their options and place them appropriately in courses.

- Consider the kinds and levels of English proficiency required in various fields and levels of study and the resources available for improving the English language skills of nonnative speakers.

Applicants' fields of study determine the kind and level of language proficiency needed. Students pursuing studies in fields requiring high verbal ability, such as journalism, may need a stronger command of English, particularly of its grammar and the conventions of written expression, than those in fields such as mathematics. Furthermore, many institutions require a higher range of TOEFL test scores for graduate than for undergraduate applicants. Institutions offering courses in English as a second language (ESL) for nonnative speakers can modify academic course loads to allow students to have concurrent language training, and thus may be able to consider applicants with a lower range of scores than institutions that do not offer supplemental language training.

- Consider TOEFL test scores an aid in interpreting an applicant's performance on other standardized tests.

Interpreting the relationship between the TOEFL test and achievement tests and tests of dependent abilities in verbal areas can be complex. Few of the most qualified foreign applicants approach native proficiency in English. Factors such as cultural differences in educational programs may also affect performance on tests of verbal ability. Nonetheless, international applicants are often among the most qualified applicants available in terms of aptitude and preparation. They are frequently required to take standardized tests in addition to the TOEFL test to establish their capacity for academic work. In some cases, TOEFL scores may prove useful in interpreting these scores. For example, if applicants' TOEFL scores *and* their scores on other assessments of verbal skills are low, it may be that performance on the latter tests is impaired because of deficiencies in English. However, the examination records of students with high verbal scores but low TOEFL scores should be carefully reviewed. It may be that high verbal scores are not valid and that TOEFL scores may be artificially low because of some special circumstance.

The TOEFL program has published research reports that can help institutions evaluate the effect of language proficiency on an applicant's performance on specific standardized tests. Although these studies were conducted with data from paper-based administrations, they may offer some help with the cross-interpretation of different measures.

- ❖ *The Performance of Nonnative Speakers of English on TOEFL and Verbal Aptitude Tests* (Aurelis, Swinton, and Cowell, 1979) gives comparative data about foreign student performance on the TOEFL test and either the GRE verbal or the SAT verbal and the Test of Standard Written English (TSWE). It provides interpretive information about how combined test results might best be evaluated by institutions that are considering foreign students.
- ❖ *The Relationship Between Scores on the Graduate Management Admission Test and the Test of English as a Foreign Language* (Powers, 1980) provides a similar comparison of performance on the GMAT and TOEFL tests.
- ❖ *Language Proficiency as a Moderator Variable in Testing Academic Aptitude* (Alderman, 1981) and *GMAT and GRE Aptitude Test Performance in Relation to Primary Language and Scores on TOEFL* (Wilson, 1982) contain information that supplements the other two studies.

- Do not use TOEFL test scores to predict academic performance.

The TOEFL test is designed as a measure of English language proficiency, not of developed academic abilities. Although there may be some overlap between language proficiency and academic abilities, other tests have been designed to measure those abilities more overtly and more precisely than the TOEFL test. Therefore, the use of TOEFL scores to predict academic performance is inappropriate. Numerous predictive validity studies,¹⁰ using grade-point averages as criteria, have been conducted. These have shown that correlations between TOEFL test scores and grade-point averages are often too low to be of any practical significance. Moreover, low correlations are a reasonable expectation where TOEFL scores are concerned. If an institution admits only international applicants who have demonstrated a high level of language competence, academic success performance cannot be attributed to English proficiency because they all possess high proficiency. Rather, other factors, such as these applicants' academic preparation or motivation, may be paramount.

¹⁰ Chase and Stallings, 1966; Heil and Aleamoni, 1974; Homburg, 1979; Hwang and Dizney, 1970; Odunze, 1980; Schrader and Pitcher, 1970; Sharon, 1972

The English proficiency of international applicants is not as stable as their mathematical skills. Proficiency in a language is subject to change over relatively short periods. If considerable time has passed between the date on which an applicant takes the TOEFL test and the date on which he or she begins academic study, loss of language proficiency may have a greater impact on academic performance than anticipated. On the other hand, students who are at a linguistic disadvantage in the first term of study might find themselves at less of a disadvantage in subsequent terms.

- Assemble information about the validity of TOEFL test score requirements at the institution.

It is important to establish appropriate standards of language proficiency. Given the wide variety of educational programs, it is impossible for TOEFL to design a validity study that is relevant for all institutions. Rather, the TOEFL program strongly encourages score users to design and carry out institutional validity studies to validate a relationship between the established cut-score range and the levels of performance on classroom tasks.¹¹ Validity evidence may provide support for raising or lowering requirements or for retaining current requirements should their legitimacy be challenged.

An important source of validity evidence for TOEFL scores is students' performance in English or ESL courses. Scores can be compared to such criterion measures as teacher or adviser ratings of English proficiency, graded written presentations, grades in ESL courses, and self-ratings of English proficiency. However, using data obtained solely from individuals who have met a high admissions standard may be problematic. If the standard is so high that only those with a high degree of proficiency are admitted, there may be no relationship between TOEFL scores and criterion measures. Because there will be little important variability in English proficiency among the group, variations in success

on the criterion variable will more likely be due to knowledge of the subject matter, academic aptitude, study skills, cultural adaptability, or financial security than English proficiency.

On the other hand, if the English proficiency requirement is low, many students will be unsuccessful because of an inadequate command of the language, and there will be a relatively high correlation between their TOEFL scores and criterion measures such as those given above. With a requirement that is neither too high nor too low, the correlation between TOEFL scores and subsequent success will be only moderate, and the magnitude of the correlation will depend on a variety of factors. These factors may include variability in scores on the criterion measures and the reliability of the raters, if raters are used.

Additional methodological issues should be considered before conducting a standard-setting or validation study. Because language proficiency can change within a relatively short time, student performance on a criterion variable should be assessed during the first term of enrollment. Similarly, if TOEFL scores are not obtained immediately prior to admission, gains or losses in language skills may reduce the relationship between the TOEFL test and the criterion. Another issue is the relationship between subject matter or level of study and language proficiency. Not all academic subjects require the same level of language proficiency for acceptable performance in the course. For instance, the study of mathematics may require a lower proficiency in English than the study of philosophy. Similarly, first-year undergraduates who are required to take courses in a wide range of subjects may need a different profile of language skills from graduate students enrolled in specialized courses of study.

Section scores should also be taken into consideration in the setting and validating of score requirements. For fields that require a substantial amount of reading, the Reading score may be particularly important. For fields that require little writing, the Structure/Writing score may be less important. Assessment of the relationship of section scores to criterion variables can further refine the process of interpreting TOEFL scores.

¹¹ A separate publication, *Guidelines for TOEFL Institutional Validity Studies*, provides information to assist institutions in the planning of local validity studies. To order this publication, fill out the form on page 43. To support institutions that wish to conduct validity studies on their cut scores for computer-based TOEFL test, the TOEFL program plans to fund a limited number of local validation studies. For more information, contact the TOEFL program office. Additional information about the setting and validation of test score standards is available in a manual by Livingston and Zieky, 1982.

To be useful, data about subsequent performance must be collected for relatively large numbers of students over an extended period of time. Because the paper-based test will eventually be eliminated, it might be advisable for institutions to convert paper-based scores to computer-based scores to establish trends. However, to do this, institutions must convert the score of every student, not just the means of certain groups. Institutions that have only begun to require TOEFL scores or that have few foreign applicants may not find it feasible to conduct their own validity studies. Such institutions might find it helpful to seek information and

advice from colleges and universities that have had more extensive experience with the TOEFL test. The TOEFL program recommends that institutions evaluate their TOEFL score requirements regularly to ensure that they are consistent with the institutions' own academic requirements and the language training resources they offer nonnative speakers of English.

For additional information on score user guidelines, visit the TOEFL Web site at www.toefl.org/useofscr.html#guidelines.

Statistical Characteristics of the Test

The Computer-Based Test

The computer-based test is conceptually different from the paper-based test; it contains new item types in the Listening and Reading section, a mandatory essay rating that is incorporated into the Structure/Writing and total scores, and a new delivery system for the test and its adaptive nature that affect the experience of taking the test.

The relationship between scores on the two tests was researched by means of a concordance study, which is described in the 1998-99 edition of this *Guide* and found at www.toefl.org/dloadlib.html. pubs.

The Computer-Based Population Defined

The tables below summarize the demographic characteristics of the paper- and computer-based testing populations for the 1999-2000 testing year. It should be noted that the paper-based test was predominantly offered in selected Asian countries with some supplemental testing in other parts of the world during the 1999-2000 testing year. This accounts for most of the differences observed between the two groups. Computer- and paper-based summary statistics from previous testing years can be found on the TOEFL Web site at www.toefl.org/edsumm.html.

Table 1 shows the gender breakdowns for both groups.

Gender	Computer Percent	Paper Percent
Male	53.2	51.4
Female	46.7	48.6

*Percentages may not add up to 100 percent due to rounding.

Table 2 shows the most frequently represented native language groups in the computer-based test group and their corresponding percentages in the paper-based test group.

Native Language	Computer Percent	Paper Percent
Spanish	10.5	0.5
Chinese	9.9	33.3
Arabic	9.0	0.1
Japanese	6.8	23.9
Korean	6.3	20.8
French	4.5	0.2

Table 3 shows that a lower percentage of computer-based test examinees took the test for graduate admissions than did the paper-based test group, and a higher percentage of computer-based test examinees took the test for undergraduate admissions and for professional license than did the paper-based test group.

Reason for Taking TOEFL	Computer Percent	Paper Percent
Undergraduate Student	39.7	22.8
Graduate Student	42.4	61.9
Other School	2.0	2.3
Professional License	7.1	1.5

Intercorrelations Among Scores

The three sections of the TOEFL test (Listening, Structure/Writing, and Reading) are designed to measure different skills within the domain of English language proficiency. It is commonly recognized that these skills are interrelated; persons who are highly proficient in one area tend to be proficient in the other areas as well. If this relationship were perfect, there would be no need to report scores for each section. The scores would represent the same information repeated several times.

Table 4 gives the correlation coefficients measuring the extent of the relationships among the three sections for the computer- and paper-based scores. A correlation coefficient of 1.0 would indicate a perfect relationship between the two scores, and 0.0 would indicate the lack of a relationship.

	Listening	Structure**	Reading
Listening		.69	.71
Structure	.68		.76
Reading	.69	.80	

*The values above the diagonal are the computer-based section intercorrelations; those below the diagonal are the paper-based section intercorrelations.

**Structure includes both the computer-adaptive score and the essay rating for the computer-based test scores.

The table shows average correlations for the testing period from July 1999 to June 2000. The observed correlations, ranging from .69 to .76 for the computer-based test, and from .68 to .80 for the paper-based test, indicate that there is a fairly strong relationship among the skills tested by the three sections of the test, but that the section scores provide some unique information. The slightly lower correlation between Structure/Writing and Reading for the computer-based test is most likely due to the inclusion of the essay rating in the Structure/Writing score.

The New Scores

A TOEFL score is not just the number of correct answers. On the TOEFL test, the same number of correct responses on different tests will not necessarily result in the same score because the difficulty levels of the questions vary according to the ability level of the examinees. In other words, if one examinee receives a slightly easier test and another examinee receives a slightly harder test, the same number of correct responses would result in different scores.

Historically, examinees taking the paper-based TOEFL exam on different test dates have received different test forms. Scores calculated from different test forms are made comparable by means of a statistical process known as score equating, which adjusts the scores for slight differences in overall difficulty.

Calculating reported scores for computer-based tests is similar, in that the number of questions answered correctly is adjusted according to the difficulty level of the questions on every test. Furthermore, on an adaptive test, questions are

chosen sequentially to match an examinee's ability level. Thus, examinees with different abilities will take tests that have different levels of difficulty. As with paper-based tests, the design of the test ensures that different types of questions and a variety of subject matter are presented proportionally the same for each examinee. As with paper-based scores, statistical adjustments are made to the scores so that they can be compared. Item response theory provides the mathematical basis for this adjustment. The statistics used in the process are derived from pretesting.

Scale scores on a computer-based test are derived from ability estimates.

- The computer estimates ability based on the difficulty of the questions answered. This estimate is updated after each question is answered.
- At the end of the test, an examinee's ability estimate on each section is converted to a scale score that enables one to compare the scale scores on different computer-based tests.

The total scale score for each examinee, which is reported on a 0 to 300 scale, is determined by adding the scale scores for all the sections and multiplying that figure by ten thirds (10/3).

Sample calculation:

$$\begin{array}{r}
 \text{Listening} + \text{Structure/Writing} + \text{Reading} = \text{Total} \\
 21 + 22 + 21 = 64 \\
 64 \times 10 \div 3 = 213
 \end{array}$$

The Structure adaptive score and the essay rating each contribute approximately 50 percent to the Structure/Writing composite score.

Calculation of the Structure/Writing Score

The composite Structure/Writing score is not a combination of the number correct on Structure and a rating on the essay. The score on the adaptive Structure section is calculated as a function of the difficulty of the questions given and the examinee's performance on those questions. The essay rating is weighted to account for approximately 50 percent of the composite score. Because these separate scores (adaptive Structure and essay) are both unrounded decimal values, their sum (the composite) is actually on a continuous scale, which is then converted to a scale score (also a decimal value) and rounded. As a result of this summing, scaling, and rounding, the same rounded Structure-only score viewed on

screen at the test center and an unweighted essay rating can result in slightly different final composite scores. For this reason, it is not possible to provide a table illustrating the exact conversion of Structure and Writing scores to composite scores.

The maximum scale score on Structure is 13. This is the official score examinees would receive if they had a perfect score on Structure and a zero (0) on the essay. An essay rating of 1 would add approximately 3 points to an examinee's composite Structure/Writing scale score and approximately 10 points to the total scale score. Each successive 0.5 increase in the essay rating adds approximately 1 to 2 points to the composite Structure/Writing scale score and approximately 3 to 7 points to the total scale score. Thus, examinees' scores on the essay greatly affect not only their Structure/Writing scores but also their total scores. Scores are most dramatically affected if examinees do not write an essay at all or if they write an essay that is off topic.

The following example further illustrates how this procedure works. The examinee below viewed these unofficial scores on screen:

Listening	22
Structure/Writing	6 to 25
Reading	22
Total	167 to 230

In this sample, the possible Structure/Writing score of 6 is based on the examinee's performance on Structure and an essay rating of 0. This would result in a total score of 167. The score of 25 is based on the performance on Structure and an essay rating of 6. This would result in a total score of 230. The total score represents the sum of the three section scores multiplied by 10/3.

Given this examinee's performance on Structure, the official scores would be as follows for each possible essay rating. Note that because of the rounding described previously, two examinees with the same unofficial Structure score and essay rating might receive different official scores once the Structure and essay scores were combined.

Essay rating	Structure/Writing scale score*	Total scale score*
0.0	6	167
1.0	9	177
1.5	11	183
2.0	13	190
2.5	14	193
3.0	16	200
3.5	18	207
4.0	19	210
4.5	21	217
5.0	22	220
5.5	24	227
6.0	25	230

* Total scores end with 0, 3, or 7 only because of the averaging of the three section scores.

For a more technical explanation of this procedure, e-mail the TOEFL statisticians at toefl@ets.org.

Adequacy of Time Allowed

Time limits for the computer-based TOEFL test are as follows: Listening, 15 to 25 minutes (excluding audio); Structure, 15 to 20 minutes; and Reading, 70 to 90 minutes. No single statistic has been widely accepted as a measure of the adequacy of time allowed for a separately timed section. In computer-based testing, the best indicator of sufficient time allowed is the proportion of examinees completing the test.

The data contained in Table 6 indicate that virtually all examinees tested between July 1999 and June 2000 were able to complete the test within the time limit. Indeed, the data show that speededness is not an issue.

	Listening	Structure	Reading
Without Pretests	96.1	94.9	97.7
With Pretests	95.8	93.5	97.5

Essay Data

The essay section, which represents one-half of the Structure/Writing score,¹² presents a different set of issues from the rest of the test because it is the only section that requires the productive use of language. Furthermore, the computer-based essay departs from the Test of Written English (TWE), on which it is based, by giving examinees the option of keying in or handwriting their essays. Table 7 shows the proportion of examinees opting for each method between July 1999 and June 2000.

Essay Mode	Percent
Keyed	63.7
Handwritten	36.3

Reliabilities

The paper-based TOEFL test has been shown to be an accurate and dependable measure of proficiency in English as a foreign language. However, no test score, whether from a paper- or computer-based test, is entirely without measurement error. This does not mean that a mistake has been made in constructing or scoring the test. It means only that examinees' scores are not perfectly consistent (from one test version to another, or from one administration to another), for any of a number of reasons. The estimate of the extent to which test scores are free of variation or error in the measurement process is called reliability. Reliability describes the tendency of a set of scores to be ordered identically on two or more tests, and it can be estimated by a variety of statistical procedures. The reliability coefficient and the standard error of measurement are the two most commonly used statistical indices.

The term "reliability coefficient" is generic, but a variety of coefficients exist because errors in the measurement process can arise from a number of sources. For example, errors can stem from variations in the tasks required by the test or from the way examinees respond during the course of a single administration. Reliability coefficients that quantify these variations are known as measures of internal consistency, and they refer to the reliability of a measurement instrument at a single point in time. It is also possible to obtain reliability coefficients that take additional sources of

error into account, such as changes in the performance of examinees from day to day and/or variations in test forms. Typically, the latter measures of reliability are difficult to obtain because they require that a group of examinees be retested with the same or a parallel test form on another occasion.

In numerical value, reliability coefficients can range from .00 to .99, and generally fall between .60 and .95. The closer the value of the reliability coefficient to the upper limit, the greater the freedom of the test from error in measurement.

With regard to the essay section, because each examinee responds to only one essay prompt, it is not possible to estimate parallel-form reliability for the TOEFL Writing section from one administration. However, experience with essay measures for other ETS testing programs and populations has shown a high degree of consistency in the reliability estimates of the scores for these types of essays. The Structure/Writing composite reliability and standard error of measurement have been estimated based on this experience.

Data from simulations provide the best data on which to estimate section reliabilities and standard errors of measurement. ETS's experience with other computer-based tests that have been operational for a number of years has shown that reliability estimates based on data from simulations accurately reflect reliability estimates derived from actual test administrations. These estimates approximate internal consistency reliability, and quantify the variation due to tasks or items.

For the Structure/Writing and total composite score reliabilities and total score standard errors of measurement, however, observed score variances and correlations are used. Table 8 gives the section and total score reliabilities and standard errors of measurement for the 1998-99 testing year.

The Standard Error of Measurement

The standard error of measurement (SEM) is an estimate of how far off a score is from an examinee's actual proficiency as a result of the measurement error mentioned earlier. As an example, suppose that a number of persons all have exactly the same degree of English language proficiency. If they take the test, they are not necessarily going to receive exactly the same TOEFL scores. Instead, they will achieve scores that are probably close to each other and close to the scores that represent their actual proficiency. This variation in scores could be attributable to differences in motivation, attentiveness, the questions on the test, or other factors.

¹² The essay rating is also reported separately on score reports.

The standard error of measurement is an index of how much the scores of examinees with the same actual proficiency, or true score, can be expected to vary.

Table 8. Reliabilities and Standard Errors of Measurement (SEM)		
July 1998 – June 1999		
	Reliability	SEM
Listening	0.89	2.76
Structure/Writing	0.88	4.89
Reading	0.88	2.73
Total Score	0.95	10.8

Interpretation of the standard error of measurement is rooted in statistical theory. It is applied with the understanding that errors of measurement can be expected to follow a particular sampling distribution. That is, observed scores will vary around the true score in a predictable pattern. The score that an examinee would achieve if nothing — neither an external condition nor an internal state — interfered with the test is the “true score.” In the example above, the true score would be the score each of the examinees with the same proficiency would have achieved if there were no errors of measurement. That is, the true score is assumed to be the average of the observed scores. The standard deviation of this distribution is the standard error of measurement.

There is no way to determine how much a particular examinee’s actual proficiency may have been under- or overestimated in a single administration. However, the SEM can be useful in another way: it can be used to set score bands or confidence bands around true scores, which can then be used to determine cut-score ranges. If measurement errors are assumed to be normally distributed (which is almost always the case), an examinee’s observed score is expected to be within one SEM of his or her true score about 66 percent of the time and within two standard errors about 95 percent of the time.

In comparing total scores for two examinees, the standard errors of measurement also need to be taken into account. The standard error of the difference between TOEFL total scores for two examinees is $\sqrt{2}$ (or 1.414) times the standard error of measurement presented in Table 7 and takes into account the contribution of two error sources in the different scores. One should not conclude that one total score represents a significantly higher level of proficiency in English than another total score unless there is a difference of at least 15 points between them. In comparing section scores

for two persons, the difference should be at least 4 points for Listening and Reading, at least 7 points for Structure/Writing.¹³

Consideration of the standard error of measurement underscores the fact that no test score is entirely without measurement error, and that cut scores should not be used rigidly in evaluating an applicant’s performance on the TOEFL test. See Appendix A.

Reliability of Gain Scores

Some users of the TOEFL test are interested in the relationship between TOEFL scores that are obtained over time by the same examinees. For example, an English language instructor may be interested in the gains in TOEFL scores obtained by students in an intensive English language program. Typically, the available data will consist of differences calculated by subtracting TOEFL scores obtained at the beginning of the program from those obtained at the completion of the program. In interpreting gain scores, the reliability of the estimates of these differences must be considered. This difference is less reliable when examinees take the same version twice than when they take two versions of a test.¹⁴

The interpretation of gain scores in a local setting requires caution, because gains may reflect increased language proficiency, a practice effect, and/or a statistical phenomenon called “regression toward the mean” (which essentially means that, upon repeated testing, high scorers tend to score lower and low scorers tend to score higher).

Swinton (1983) analyzed data from a group of students at San Francisco State University that indicated that TOEFL paper-based test score gains decreased as a function of proficiency level at the time of initial testing. For this group, student scores were obtained at the start of an intensive English language program and at its completion 13 weeks later. Students whose initial scores were in the 353-400 range showed an average gain of 61 points; students whose initial scores were in the 453-500 range showed an average gain of 42 points.

As a part of the Swinton study, an attempt was made to remove the effects of practice and regression toward the mean by administering another form of the TOEFL test one week after the pretest. Initial scores in the 353-400 range

¹³ For additional information on the standard errors of score differences, see Anastasi, 1968, and Magnusson, 1967.

¹⁴ For further discussion on the limitations in interpreting gain scores, see Linn and Slinde, 1977, and Thorndike and Hagan, 1977.

increased about 20 points on the retest, and initial scores in the 453-500 range improved about 17 points on the retest. The greater part of these gains could be attributed to practice and regression toward the mean, although a small part might reflect the effect of one week of instruction.

Subtracting the retest gain (20 points) from the posttest gain (61 points), it was possible to determine that, within this sample, students with initial scores in the 353-400 range showed a real gain on the TOEFL test of 41 points during 13 weeks of instruction. Similarly, students in the 453-500 initial score range showed a 25-point gain in real language proficiency after adjusting for the effects of practice and regression. Thus, the lower the initial score, the greater the probable gain over a fixed period of instruction. Other factors, such as the nature of the instructional program, will also affect gain scores.

The TOEFL program has published a booklet¹⁵ that describes a methodology suitable for conducting local studies of gain scores. (The contents of the manual are equally relevant for the paper- and computer-based tests.) University-affiliated and private English language programs may wish to conduct gain score studies with their own students to determine the amount of time that is ordinarily required for a student to progress from one score level to another.

Validity

In addition to reliability, a test must establish its validity, that is, that scores on the test reflect what was intended to be measured, proficiency in English, for the TOEFL test. Although there are many types of validity, it is generally recognized that they are all forms of what is referred to as construct validity, or the validity of the design of the test itself, the set of behaviors it taps, and the scores it yields. Since establishing the validity of a test is one of the most difficult tasks facing test designers, validity is usually confirmed by analyzing the test from a number of perspectives, e.g., its content, the theory it embodies (the test construct), its aims and criteria. In general, such data are evaluated in the light of the test's use and purpose. That is, while a test may have a strong relationship with construct-irrelevant measures such as general ability or mathematical achievement, a language proficiency test should establish its relationship to other measures of language proficiency, even performance measures, if it is to claim validity.

One way to approach validity is to conduct a local study that establishes a relationship between TOEFL scores and the linguistic and curricular demands faced by students once they

have enrolled in classes. Given the variety of educational programs, it would be impossible to design a validity study that is universally relevant to all institutions. Rather, the TOEFL program has undertaken a large-scale study that assesses the linguistic needs of international students in a variety of discrete academic settings. Future studies will examine the relationship between scores and placements at various colleges and universities; studies like these are probably the most useful because of their close alignment with institutional needs and priorities.

Another way to approach validity is to compare sets of test scores from tests that purport to measure the same ability (Messick, 1987). For example, the validity of the TOEFL test could be estimated by giving it and another test of English proficiency to the same group of students. However, there was little validity evidence of this sort available when this manual was written. No computer-based TOEFL test had been given except for the version used in the concordance study and a 60-item version (comprising items in Listening, Structure, and Reading) developed for the familiarity studies. Furthermore, citing the scores achieved on the 60-item measure and their high correlation with paper-based scores as evidence of validity is problematic since the validity of that measure was never directly established.¹⁶

On the other hand, the concordance study contains clear evidence, in the form of correlations and intercorrelations, that the paper- and computer-based tests lead to comparable outcomes. This study employed a fully developed version of the computer-based test similar to those administered operationally. Furthermore, because an examination of the contents of the two tests suggests that they share many features, an argument could be made that the tests are closely related. Therefore, validity evidence developed over many years for the *paper*-based test, in the absence of contrary evidence, is a good source of information about the *computer*-based version. The studies of the paper-based test cited below are more fully described in the 1997 edition of the *TOEFL Test and Score Manual*. These studies are available from the TOEFL program at nominal cost, and can be ordered from the TOEFL Web site at www.toefl.org/rrpts.html.

- As early as 1985, a TOEFL research study by Duran, Canale, Penfield, Stansfield, and Liskin-Gasparro established that successful performance on the test requires a wide range of communicative competencies, including grammatical, sociolinguistic, and discourse competencies.

¹⁵ Swinton, 1983.

¹⁶ Taylor et al., 1998.

- TOEFL is not merely a test of an examinee's knowledge of sociolinguistic or cultural content associated with life in the United States. A study (Angoff, 1989) using one form of the TOEFL test with more than 20,000 examinees tested abroad and more than 5,000 in the United States showed that there was no cultural advantage for examinees who had resided more than a year in the United States.
- The test does stress academic language, however. Powers (1985) found that the kinds of listening comprehension questions used in the TOEFL test were considered highly appropriate by college faculty members.
- Bachman, Kunnan, Vanniarajan, and Lynch (1988) showed that the test's reading passages are almost entirely academic in their topics and contexts. Thus, the test has been shown to test academic English, as it was intended to.
- In 1965, Maxwell found a .87 correlation between total scores on the TOEFL test and the English proficiency test used for the placement of foreign students at Berkeley (University of California).
- Subsequently, Upshur (1966) conducted a correlational study of the TOEFL test and the Michigan Test of English Language Proficiency that yielded a correlation of .89.¹⁷ In the same year, a comparison of TOEFL scores and those on a placement test at Georgetown University (N = 104) revealed a correlation of .79 (American Language Institute, 1966). The Georgetown study also showed a correlation between TOEFL and teacher ratings for 115 students of .73; four other institutions reported similar correlations.
- In a 1976 study, Pike (1979) investigated the relationship between the TOEFL test and alternate criterion measures, including writing samples, Cloze tests, oral interviews, and sentence-combining exercises. Results suggested a close relationship, especially between subscores on the test and oral interviews and writing samples (essays). These correlations led eventually to a merger and revision of sections to form the current three-section version of the test.
- Henning and Cascallar (1992) provide similar evidence of validity in a study relating TOEFL test scores to independent ratings of oral and written communicative language ability over a variety of controlled academic communicative functions.
- Angoff and Sharon (1970) found that the mean TOEFL scores of native speakers in the United States were significantly and homogeneously higher than those of foreign students who had taken the same test. In this way, nonnative and native scores were distinguished, and the test's validity as a measure of nonnative English proficiency was substantiated.
- Clark (1977) conducted a more detailed study of native speaker performance on the TOEFL test with similar results. In this case, the mean raw score for the native speakers was 134 (out of 150), while the mean scores achieved by the nonnative group were 88 or 89.
- Angelis, Swinton, and Cowell (1979) compared the performance of nonnative speakers of English on the TOEFL test with their performance on the verbal portions of the GRE Aptitude (now General) Test (graduate-level students) or both the SAT and the Test of Standard Written English (undergraduates). The GRE verbal performance of the nonnative speakers was significantly lower and less reliable than the performance of native speakers. Similar results were reported for undergraduates on the SAT verbal and the Test of Standard Written English (TSWE).
- Between 1977 and 1979, Wilson (1982) conducted a similar study of all GRE, TOEFL, and GMAT examinees. These results, combined with those obtained in the earlier study by Angelis, et al., (1979), show that mid-range verbal aptitude test scores of nonnative examinees are significantly lower on average than the scores earned by native English speakers, whereas the scores on measures of quantitative aptitude are not greatly affected by English language proficiency.
- In a study comparing the performance of nonnative speakers of English on TOEFL and the Graduate Management Admission Test, Powers (1980) reported correlations that resemble those for Wilson's study. The fact that the correlations for the quantitative section of the GMAT are the lowest of all reinforces evidence from other sources for the power of the TOEFL test as a measure of verbal skills in contrast to quantitative skills.

¹⁷ Other studies found a similarly high correlation (e.g., Gershman, 1977; Pack, 1972).

- More recent evidence of the test's validity comes from a series of studies investigating the range of abilities the TOEFL test measures (Boldt, 1988; Hale, Rock, and Jirele, 1989; Oltman, Stricker, and Barrows, 1988), current and prospective listening and vocabulary item types (Henning, 1991a and b), and the reading and listening portions of the test (Freedle and Kostin, 1993, 1996; Nissan, DeVincenzi, and Tang, 1996; Schedl, Thomas, and Way, 1995).

These and all other studies cited in this section of the *Guide* are available from the TOEFL program. To order copies, see the TOEFL Web site www.toefl.org/rrpts.html or write to the TOEFL office:

TOEFL Program Office
P.O. Box 6155
Princeton, NJ 08541-6155
USA

References

Alderman, D. L. *TOEFL item performance across seven language groups* (TOEFL Research Report 9). Princeton, NJ: Educational Testing Service, 1981.

American Language Institute (Georgetown). *A report on the results of English testing during the 1966 Pre-University Workshop at the American Language Institute*. Unpublished manuscript. Georgetown University, 1966.

Anastasi, A. *Psychological testing* (3rd ed.). New York: Macmillan, 1968.

Angelis, P. J., Swinton, S. S., and Cowell, W. R. *The performance of nonnative speakers of English on TOEFL and verbal aptitude tests* (TOEFL Research Report 3). Princeton, NJ: Educational Testing Service, 1979.

Angoff, W. A., and Sharon, A. T. *A comparison of scores earned on the Test of English as a Foreign Language by native American college students and foreign applicants to United States colleges* (ETS Research Bulletin No. 70-8). Princeton, NJ: Educational Testing Service, 1970.

Bachman, L. F., Kunnan, A., Vanniarajan, S., and Lynch, B. Task and ability analysis as a basis for examining content and construct comparability in two EFL proficiency test batteries. *Language Testing*, 1988, 5(2), 128-159.

Boldt, R. F. *Latent structure analysis of the Test of English as a Foreign Language* (TOEFL Research Report 28). Princeton, NJ: Educational Testing Service, 1988.

Carroll, J. B. Fundamental considerations in testing for English proficiency of foreign students. *Testing the English proficiency of foreign students*. Washington, DC: Center for Applied Linguistics, 31-40, 1961

Chase, C. I., and Stallings, W. M. *Tests of English language as predictors of success for foreign students* (Indiana Studies in Prediction No. 8. Monograph of the Bureau of Educational Studies and Testing.) Bloomington, IN: Bureau of Educational Studies and Testing, Indiana University, 1966.

Clark, J. L. D. *The performance of native speakers of English on the Test of English as a Foreign Language* (TOEFL Research Report 1). Princeton, NJ: Educational Testing Service, 1977.

Eignor, D., Taylor, C., Kirsch, I., and Jamieson, J. *Development of a Scale for Assessing the Level of Computer Familiarity of TOEFL Examinees* (TOEFL Research Report 60). Princeton, NJ: Educational Testing Service, 1998.

Freedle, R., and Kostin, I. *The prediction of TOEFL reading comprehension item difficulty for expository prose passages for three item types: main idea, inference, and supporting idea items* (TOEFL Research Report 44). Princeton, NJ: Educational Testing Service, 1993.

Freedle, R., and Kostin, I. *The prediction of TOEFL listening comprehension item difficulty for minitalk passages: implications for construct validity* (TOEFL Research Report 56). Princeton, NJ: Educational Testing Service, 1996.

Gershman, J. *Testing English as a foreign language: Michigan/TOEFL study*. Unpublished manuscript. Toronto Board of Education, 1977.

Hale, G. A., Rock, D. A., and Jirele, T. *Confirmatory factor analysis of the Test of English as a Foreign Language* (TOEFL Research Report 32). Princeton, NJ: Educational Testing Service, 1989.

Heil, D. K., and Aleamoni, L. M. *Assessment of the proficiency in the use and understanding of English by foreign students as measured by the Test of English as a Foreign Language* (Report No. RR-350). Urbana: University of Illinois. (ERIC Document Reproduction Service No. ED 093 948), 1974.

Henning, G. *A study of the effects of variations of short-term memory load, reading response length, and processing hierarchy on TOEFL listening comprehension item performance* (TOEFL Research Report 33). Princeton, NJ: Educational Testing Service, 1991a.

Henning, G. *A study of the effects of contextualization and familiarization on responses to TOEFL vocabulary test items* (TOEFL Research Report 35). Princeton, NJ: Educational Testing Service, 1991b.

Henning, G., and Cascallar, E. *A preliminary study of the nature of communicative competence* (TOEFL Research Report 36). Princeton, NJ: Educational Testing Service, 1992.

Homburg, T. J. TOEFL and GPA: an analysis of correlations. In R. Silverstein, *Proceedings of the Third International Conference on Frontiers in Language Proficiency and Dominance Testing*. Occasional Papers on Linguistics, No. 6. Carbondale: Southern Illinois University, 1979.

Hwang, K. Y., and Dizney, H. F. Predictive validity of the Test of English as a Foreign Language for Chinese graduate students at an American university. *Educational and Psychological Measurement*, 1970 30, 475-477.

Kirsch, I., Jamieson, Taylor, C., and Eignor, D., *Computer Familiarity Among TOEFL Examinees* (TOEFL Research Report 59). Princeton, NJ: Educational Testing Service, 1998.

Linn, R., and Slinde, J. The determination of the significance of change between pre- and posttesting periods. *Review of Educational Research*, 1977, 47(1), 121-150.

Magnusson, D. *Test theory*. Boston: Addison-Wesley, 1967.

Maxwell, A. *A comparison of two English as a foreign language tests*. Unpublished manuscript. University of California (Davis), 1965.

Messick, S. *Validity* (ETS Research Bulletin No. 87-40). Princeton, NJ: Educational Testing Service, 1987. Also appears in R. L. Linn (Ed.), *Educational measurement* (3rd Ed.). New York: MacMillan, 1988.

Nissan, S., DeVincenzi, F., and Tang, K. L. *An analysis of factors affecting the difficulty of dialog items in TOEFL listening comprehension* (TOEFL Research Report 51). Princeton, NJ: Educational Testing Service, 1996.

Odunze, O. J. Test of English as a Foreign Language and first year GPA of Nigerian students (Doctoral dissertation, University of Missouri-Columbia, 1980.) *Dissertation Abstracts International*, 42, 3419A-3420A. (University Microfilms No. 8202657), 1982.

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

Pack, A. C. A comparison between TOEFL and Michigan Test scores and student success in (1) freshman English and (2) completing a college program. *TESL Reporter*, 1972, 5, 1-7, 9.

Pack, A. C. A comparison between TOEFL and Michigan Test scores and student success in (1) freshman English and (2) completing a college program. *TESL Reporter*, 1972.

Pike, L. *An evaluation of alternative item formats for testing English as a foreign language* (TOEFL Research Report 2). Princeton, NJ: Educational Testing Service, 1979.

Powers, D. E. *A survey of academic demands related to listening skills* (TOEFL Research Report 20). Princeton, NJ: Educational Testing Service, 1985.

Powers, D. E. *The relationship between scores on the Graduate Management Admission Test and the Test of English as a Foreign Language* (TOEFL Research Report 5). Princeton, NJ: Educational Testing Service, 1980.

Schedl, M., Thomas, N., and Way, W. *An investigation of proposed revisions to the TOEFL test* (TOEFL Research Report 47). Princeton, NJ: Educational Testing Service, 1995.

Schrader, W. B., and Pitcher, B. *Interpreting performance of foreign law students on the Law School Admission Test and the Test of English as a Foreign Language* (Statistical Report 70-25). Princeton, NJ: Educational Testing Service, 1970.

Sharon, A. T. *Test of English as a Foreign Language as a moderator of Graduate Record Examinations scores in the prediction of foreign students' grades in graduate school* (ETS Research Bulletin No. 71-50). Princeton, NJ: Educational Testing Service, 1971.

Swinton, S. S. *A manual for assessing language growth in instructional settings* (TOEFL Research Report 14). Princeton, NJ: Educational Testing Service, 1983.

Taylor, C., Jamieson, J., Eignor, D., and Kirsch, I., *The relationship between computer familiarity and performance on computer-based TOEFL test tasks* (TOEFL Research Report 61). Princeton, NJ: Educational Testing Service, 1998.

Thorndike, R. I. and Hagen, E. P., *Measurement and evaluation in education* (4th ed.). New York: Wiley, 1977.

Upshur, J. A., *Comparison of performance on "Test of English as a Foreign Language" and "Michigan Test of English Language Proficiency."* Unpublished manuscript. University of Michigan, 1966.

Wilson, K. M. *GMAT and GRE Aptitude Test performance in relation to primary language and scores on TOEFL* (TOEFL Research Report 12). Princeton, NJ: Educational Testing Service, 1982.

Appendix A:

Standard-Setting Procedures, Concordance Tables

Using Range-to-Range Concordance Tables to Establish a Cut-Score Range

The computer-based TOEFL test does not measure English language proficiency in the same manner as the paper-based test and there are different numbers of score points on each scale (0-300 for the computer-based test and 310-677 for the paper-based test). As a consequence, there is no one-to-one relationship between scores on these two highly related measures. Therefore, it is advisable to use the range-to-range concordance tables (pages 35-36) when establishing cut scores.¹ To use these range-to-range tables, find the score range that includes the cut score your institution requires on the paper-based TOEFL, and then look across the table to identify the comparable score range on the computer-based test.

Using the Standard Error of Measurement When Establishing a Cut-Score Range

Consideration of the standard error of measurement (SEM) underscores the fact that no test score is entirely without measurement error, and that cut scores should not be used in a completely rigid fashion in evaluating an applicant's performance on the TOEFL test.

The standard error of measurement is an index of how much the scores of examinees with the same actual proficiency can be expected to vary. In most instances, the SEM is treated as an average value and applied to all scores in the same way. It can be expressed in the same units as the reported score, which makes it quite useful in interpreting the scores of individuals. For the computer-based test the estimated SEMs are approximately 3 points for Listening and Reading, approximately 5 points for Structure/Writing, and approximately 11 points for the total score. (See Table 8 on page 28.) There is, of course, no way of knowing just how much a particular person's actual proficiency may have been under- or overestimated from a single administration. However, the SEM can be used to provide bands around true scores, which can be used in determining cut-score ranges.

If measurement errors are assumed to be normally distributed, a person's observed score is expected to be within one SEM of his or her true score about 66 percent of the time and within two standard errors about 95 percent of the time.

Some institutions do use a single cut score even though this practice is inadvisable. For example, the total score of 550 on the paper-based TOEFL is used as a cut score by a number of institutions. By using the score-to-score concordance tables (pages 35-36), one can see that a score of 213 on the computer-based TOEFL is comparable to 550 on the paper-based test. Since one SEM for the computer-based TOEFL total score is around 11 points, one could set a cut-score range of 202-224, which constitutes a band of 11 points on either side of 213. Similarly, by using 1 SEM on section scores one could set a band of 3 points on either side of a Listening or Reading score of 21, yielding a section cut-score range of 18-24, and a band of 5 points on either side of a Structure/Writing score of 21, yielding a cut-score range of 16-21.

¹ The score-to-score tables are provided for the convenience of institutions that rely on automated score processing for decisions about applicants and need to modify databases and define score-processing procedures that accommodate both types of scores (See Appendix B). It is not advisable to use these score-to-score tables when establishing cut-score ranges.

TOEFL[®]

Concordance Table Total Score

Score Comparison

Paper-based Total	Computer-based Total	Paper-based Total	Computer-based Total	Paper-based Total	Computer-based Total
677	300	523	193	370	77
673	297	520	190	367	73
670	293	517	187	363	73
667	290	513	183	360	70
663	287	510	180	357	70
660	287	507	180	353	67
657	283	503	177	350	63
653	280	500	173	347	63
650	280	497	170	343	60
647	277	493	167	340	60
643	273	490	163	337	57
640	273	487	163	333	57
637	270	483	160	330	53
633	267	480	157	327	50
630	267	477	153	323	50
627	263	473	150	320	47
623	263	470	150	317	47
620	260	467	147	313	43
617	260	463	143	310	40
613	257	460	140		
610	253	457	137		
607	253	453	133		
603	250	450	133		
600	250	447	130		
597	247	443	127		
593	243	440	123		
590	243	437	123		
587	240	433	120		
583	237	430	117		
580	237	427	113		
577	233	423	113		
573	230	420	110		
570	230	417	107		
567	227	413	103		
563	223	410	103		
560	220	407	100		
557	220	403	97		
553	217	400	97		
550	213	397	93		
547	210	393	90		
543	207	390	90		
540	207	387	87		
537	203	383	83		
533	200	380	83		
530	197	377	80		
527	197	373	77		

Range Comparison	
Paper-based Total	Computer-based Total
660-677	287-300
640-657	273-283
620-637	260-270
600-617	250-260
580-597	237-247
560-577	220-233
540-557	207-220
520-537	190-203
500-517	173-187
480-497	157-170
460-477	140-153
440-457	123-137
420-437	110-123
400-417	97-107
380-397	83- 93
360-377	70- 80
340-357	60- 70
320-337	47- 57
310-317	40- 47

TOEFL® Concordance Table Section Scaled Scores

Listening				Structure/Writing*				Reading					
Score-to-Score	Score-to-Score	Score-to-Score	Score-to-Score	Score-to-Score	Score-to-Score	Score-to-Score	Score-to-Score	Score-to-Score	Score-to-Score	Score-to-Score	Score-to-Score	Score-to-Score	Score-to-Score
Paper-based Listening Comprehension	Computer-based Listening	Paper-based Listening Comprehension	Computer-based Listening	Paper-based Structure and Written Expression	Computer-based Structure/Writing	Paper-based Structure and Written Expression	Computer-based Structure/Writing	Paper-based Structure and Written Expression	Computer-based Structure/Writing	Paper-based Reading Comprehension	Computer-based Reading	Paper-based Reading Comprehension	Computer-based Reading
68	30	40	7	68	30	68	30	67	11	67	30	39	9
67	30	39	6	67	29	67	29	66	10	66	29	38	9
66	29	38	6	66	28	66	28	65	9	65	28	37	8
65	28	37	5	65	28	65	28	64	9	64	28	36	8
64	27	36	5	64	27	64	27	63	8	63	27	35	7
63	27	35	4	63	27	63	27	62	8	62	26	34	7
62	26	34	4	62	26	62	26	61	7	61	26	33	6
61	25	33	3	61	26	61	26	60	7	60	25	32	6
60	25	32	3	60	25	60	25	59	6	59	25	31	5
59	24	31	2	59	25	59	25	58	6	58	24		
58	23			58	24	58	24	57		57	23		
57	22			57	23	57	23	56		56	22		
56	22			56	23	56	23	55		55	21		
55	21			55	22	55	22	54		54	21		
54	20			54	21	54	21	53		53	20		
53	19			53	20	53	20	52		52	19		
52	18			52	20	52	20	51		51	18		
51	17			51	19	51	19	50		50	17		
50	16			50	18	50	18	49		49	16		
49	15			49	17	49	17	48		48	16		
48	14			48	17	48	17	47		47	15		
47	13			47	16	47	16	46		46	14		
46	12			46	15	46	15	45		45	13		
45	11			45	14	45	14	44		44	13		
44	10			44	14	44	14	43		43	12		
43	9			43	13	43	13	42		42	11		
42	9			42	12	42	12	41		41	11		
41	8			41	11	41	11			40	10		
Range-to-Range				Range-to-Range				Range-to-Range					
Paper-based Listening Comprehension	Computer-based Listening	Paper-based Listening Comprehension	Computer-based Listening	Paper-based Structure and Written Expression	Computer-based Structure/Writing	Paper-based Structure and Written Expression	Computer-based Structure/Writing	Paper-based Reading Comprehension	Computer-based Reading	Paper-based Reading Comprehension	Computer-based Reading	Paper-based Reading Comprehension	Computer-based Reading
64-68	27-30	64-68	27-30	64-68	27-30	64-68	27-30	64-67	28-30	64-67	28-30	64-67	28-30
59-63	24-27	59-63	24-27	59-63	25-27	59-63	25-27	59-63	25-27	59-63	25-27	59-63	25-27
54-58	20-23	54-58	20-23	54-58	21-24	54-58	21-24	54-58	21-24	54-58	21-24	54-58	21-24
49-53	15-19	49-53	15-19	49-53	17-20	49-53	17-20	49-53	16-20	49-53	16-20	49-53	16-20
44-48	10-14	44-48	10-14	44-48	14-17	44-48	14-17	44-48	13-16	44-48	13-16	44-48	13-16
39-43	6-9	39-43	6-9	39-43	10-13	39-43	10-13	39-43	9-12	39-43	9-12	39-43	9-12
34-38	4-6	34-38	4-6	34-38	7-9	34-38	7-9	34-38	7-9	34-38	7-9	34-38	7-9
31-33	2-3	31-33	2-3	31-33	6-7	31-33	6-7	31-33	5-6	31-33	5-6	31-33	5-6

* Structure/Writing in the computer-based test includes multiple-choice items and an essay. The Structure and Written Expression section in the paper-based test consists of multiple-choice items only. Therefore, these section scores are derived differently.

Appendix B:

Database Modification Options

Database Modifications

The score-to-score concordances (Appendix A) can help institutions accept scores from both scales. Below is a list of frequently asked questions about modifying databases to accommodate the new computer-based TOEFL scores.

Q: Are there changes to the data layout for the TOEFL score report?

- A:** Yes, for magnetic tape and diskette users, the record length increases from 220 bytes to 239 bytes. In addition to field size and position changes, there are several new fields in the 1998-99 layout.
- Examinee Name was expanded from 21 characters to 30 characters to accommodate computer-based TOEFL.
 - Registration Number is now referred to as Appointment Number and was expanded to 16 characters to accommodate computer-based TOEFL. For paper-based score records, registration number is still 7 characters, left justified.
 - Date of Birth and Administration Date now contain century to accommodate Year 2000. The format is CCYYMMDD.
 - Center has been expanded to 5 characters to accommodate computer-based TOEFL. For paper-based score records, the center is still 4 characters, left justified.
 - Test Type is a new field to denote if the score record contains a paper-based score or a computer-based score.
 - All references to TSE except for the actual TSE score were removed.

The score area still shows the section and total scores. Added to the field descriptions are the computer-based test section names. Nonstandard indicators are used for both paper-based and computer-based tests: (L) if the listening portion was not administered and (X) if extra time was given (computer-based TOEFL test only). This is a change to the paper-based layout, where “9999” in Interpretive Information indicated a nonstandard test was given. Listed under “Fields” for paper-based only is an additional field, “truncated scale indicator”; a “Y” in this field means that the score has been adjusted to a 310 because of the truncation of the paper-based score scale.

Q: Where can I get a copy of the new data set layout?

A: Copies of the new data set layout can be obtained by calling the TOEFL office at 1-609-771-7091.

Q: How will I know that a newly received score report is in the new layout?

A: Tapes/diskettes received after August 1, 1998, present scores in the new layout. A special insert accompanies the first tape/diskette received in the new layout.

Q: Will paper-based and computer-based score reports share this new layout? Will the scores be reported on the same tape?

A: Institutions will receive paper-based scores and computer-based scores in the same format and on the same file. Test Type, Field 18, designates whether the student took the paper- or the computer-based test. If you wish to compare computer- and paper-based scores, refer to the concordance tables (Appendix A).

Q: What are the suggested system implementation options for the new layout and score scale?

Option A

Allocate to your database a new one-byte field to denote the test type being reported (e.g., “C” = computer-based TOEFL, “P” = paper-based TOEFL). When you receive a computer-based score report, use the concordance tables to convert the score to a paper-based score. Use the paper-based score in your initial automated processes (and store it if required by your institution). Store the original computer-based TOEFL score along with the appropriate code in the new test type field (e.g., C,P) in your database for long-term use.

Pros:

- Minimal changes to current data structures for storing the test scores (primarily the addition of the new one-byte field to denote the test type)
- Minimal changes to automated routines for processing of the score; conversion from the computer-based TOEFL scale to the paper-based scale should be the only modification (Your automated routines will continue to be driven by the paper-based test.)
- Staff to become familiar with the new computer-based score scale because those scores are the ones actually stored

- Easy access to the computer-based score when communicating with the test taker
- Will allow modification of automated routines to be driven by computer-based TOEFL instead of the paper-based test at some later date.

Cons:

- Reliance upon the TOEFL score-to-score comparison tables and not the range comparison tables
- Will require the conversion from the computer-based TOEFL score scale to the paper-based scale for display in on-line systems for printed reports (as deemed necessary by staff), and in automated routines where processes are driven based on paper-based TOEFL scores

Option B

Allocate to your database new data elements to house the new computer-based TOEFL scores, preserving your existing fields for storing paper-based results. Use each area exclusively for the test type being reported.

Pros:

- Keeps the automated processes pure because paper- and computer-based TOEFL scores are not converted
- Provides easy access to the original score when communicating with the test taker
- Positions both manual processes and automated routines for the time when the paper-based scale is phased out

Cons:

- Extensive changes to current data structures for storing computer-based TOEFL test score results
- Extensive changes to automated routines for supporting both the paper- and computer-based scores in your operating rules (i.e., determination of satisfactory scores on the test will require setting criteria for both scales)
- Both types of scores will have to be considered in manual and automated processes
- Extensive changes to on-line systems and printed reports to support the display of both paper- and computer-based scores

Appendix C

Essay Ratings/Explanations of Scores

- 6 An essay at this level
 - 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, though it may have occasional errors
- 5 An essay at this level
 - 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 the language
 - demonstrates some syntactic variety and range of vocabulary, though it will probably have occasional errors
- 4 An essay at this level
 - 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
 - displays adequate but possibly inconsistent facility with syntax and usage
 - may contain some errors that occasionally obscure meaning
- 3 An essay at this level 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 An essay at this level 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 An essay at this level
 - may be incoherent
 - may be undeveloped
 - may contain severe and persistent writing errors
- 0 An essay will be rated 0 if it
 - contains no response
 - merely copies the topic
 - is off topic, is written in a foreign language, or consists only of keystroke characters

Appendix D

Computer-familiarity Studies

Three studies have been conducted by the TOEFL program to assess the effect of asking the TOEFL examinee population, in all its diversity, to move from a paper-based mode of testing to a computer-based platform. Specifically, the issue of the examinees' familiarity or experience with using a personal computer (PC) for word processing and the other elementary operations necessary for the test was addressed. A lack of familiarity, or a low comfort level, was seen as a potential problem in the testing of individuals who had had little or no contact with the machine. If the delivery system so distracted examinees that their scores were affected, the program would have to explore the problem fully and resolve it before committing to computer-based testing.

The first step was to define the universe of computer familiarity. This effort took the form of a 23-item questionnaire distributed in April and May 1996 to TOEFL examinees (N = 89,620) who were found to be highly similar to the 1995-96 total examinee population in terms of gender, native language, test center region, TOEFL score ranges, and reason for taking the test. The aim of this study was to gather data on examinees' access to, attitudes toward, and experiences with the computer. Analysis of these data made it possible to categorize examinees into three subgroups: high, medium and low familiarity.

Overall, 16 percent of the TOEFL population was judged to have low computer familiarity, 34 percent to have moderate familiarity, and 50 percent to have high familiarity. In terms of background characteristics, computer familiarity was shown to be unrelated to age, but somewhat related to gender, native language, native region, and test region. Computer familiarity was also shown to be related to individuals' TOEFL scores and reasons for taking the test, but unrelated to whether or not the examinees had previously taken the test. The Phase I research further showed a small but significant relationship between computer familiarity and paper-based TOEFL test scores.

Phase II of the study examined the relationship between level of computer familiarity and level of performance on a set of computer-based test questions. More than 1,100 examinees classified as low or high computer familiarity from the survey results from 12 international sites were chosen from Phase I participants. The 12 sites were selected to represent both TOEFL volumes in various regions and broad geographic areas. They included Bangkok, Cairo,

Frankfurt, Karachi, Mexico City, Oakland (CA), Paris, Seoul, Taipei, Tokyo, Toronto, and Washington, DC. Participants were administered a computer tutorial and 60 computer-based TOEFL questions.

Differences between the low and high familiarity groups emerged, making the sample atypical, or at least different from the larger one that had participated in the questionnaire study. These differences included examinees' reasons for taking the test (between the original sample and the low familiarity group), and their English proficiency as established by their paper-based scores (again in the low familiarity group, but practically insignificant). Before the test scores could be weighted to account for differences in proficiency, three broad analyses of covariance were conducted to ensure that a clear-cut distinction, unencumbered by other variables, between computer familiarity and unfamiliarity could be made. The aim of these analyses was to isolate a sample that matched the original sample in key characteristics, and the method used was to adjust the sample while making sure that the distribution of scores achieved on the paper-based test continued to approximate that of the larger sample.

After identical distributions on the participants' paper-based test scores had been achieved, so that the two samples were virtually identical in their characteristics, **no meaningful relationship was found between computer familiarity and the examinees' performances on the 60 computer-based questions once the examinees had completed the computer tutorial.** In other words, those who had done poorly on the paper-based test because of low English proficiency also, as expected, performed poorly on the computer-based questions; those with high paper-based scores achieved high computer-based scores. Despite marginal interactions, the tutorial had leveled the playing field by eliminating familiarity as a confounding factor and enabling all participants to achieve as expected. Conclusion: there is no evidence of adverse effects on the computer-based TOEFL test performance because of prior computer experience.¹

¹ Taylor et al., 1998.

Three research reports on computer familiarity are listed below. These can be downloaded at no charge from the TOEFL Web site at www.toefl.org and can also be ordered in the published form for \$7.50 each at www.toefl.org/rrpts.html or by using the order form on page 43.

- ❖ *Computer Familiarity Among TOEFL Examinees.* TOEFL Research Report No. 59.
- ❖ *Development of a Scale for Assessing the Level of Computer Familiarity of TOEFL Examinees.* TOEFL Research Report No. 60.
- ❖ *The Relationship Between Computer Familiarity and Performance on Computer-Based TOEFL Test Tasks.* TOEFL Research Report No. 61.

Where to Get More TOEFL Information

Bulletins are usually available from local colleges and universities. In addition, *Bulletins* and study materials are also available at many of the locations listed below, at United States educational commissions and foundations, United States Information Service (USIS) offices, binational centers and private organizations, and directly from Educational Testing Service. Additional distribution locations for study materials are listed on the ETS Web site at www.ets.org/cbt/outsidus.html.

ALGERIA, OMAN, QATAR, SAUDI ARABIA, SUDAN

AMIDEAST
Testing Programs
1730 M Street, NW, Suite 1100
Washington, DC 20036-4505, USA
Telephone: 202-776-9600
www.amideast.org

EGYPT

AMIDEAST
23 Mossadak Street
Dokki, Giza, Egypt
Telephone: 20-2-337-8265
20-2-338-3877
www.amideast.org

or

AMIDEAST
American Cultural Center
3 Pharaana Street
Azarita, Alexandria
Egypt
Telephone: 20-3-482-9091
www.amideast.org

EUROPE, East/West

CITO
P.O. Box 1203
6801 BE Arnhem
Netherlands
Telephone: 31-26-352-1577
www.cito.nl

HONG KONG

Hong Kong Examinations
Authority
San Po Kong Sub-Office
17 Tseuk Luk Street
San Po Kong
Kowloon, Hong Kong
Telephone: 852-2328-0061, ext. 365
www.cs.ust.hk/hkea

INDIA/BHUTAN

Institute of Psychological and
Educational Measurement
119/25-A Mahatma Gandhi Marg
Allahabad, 211001, U.P. India
Telephone: 91-532-624881
or 624988
www.ipem.org

INDONESIA

International Education Foundation (IEF)
Menara Imperium, 28th Floor, Suite B
Metropolitan Kuningan
Superblok, Kav. 1
JI. H.R. Rasuna Said
Jakarta Selatan 12980
Indonesia
Telephone: 62-21-8317304 (hunting)
www.ief.org/ief

ISRAEL

AMIDEAST
Ahmad Abdelaziz Street
PO Box 1247
Gaza City, Palestinian National
Authority
Telephone: 972-2-7-286-9338
www.amideast.org

or

AMIDEAST

Nabulsi Bldg. 1st Floor
Corner of Road #1 and Anata Road
PO Box 19665
Shu'fat Jerusalem
Telephone: 972-2-581-1962
www.amideast.org

JAPAN

Council on International Educational
Exchange
TOEFL Division
Cosmos Aoyama B1
5-53-67 Jingumae, Shibuya-ku
Tokyo 150-8355, Japan
Telephone: (813) 5467-5520
www.cieej.or.jp

JORDAN

AMIDEAST
P.O. Box 1249
Amman, 11118 Jordan
Telephone: 962-6-586-2950 or
962-6-581-4023
www.amideast.org

KOREA

Korean-American Educational
Commission (KAEC)
M.P.O. Box 112
Seoul 121-600, Korea
Telephone: 02-3275-4000
www.fulbright.or.kr

KUWAIT

AMIDEAST
P.O. Box 44818
Hawalli 32063, Kuwait
Telephone: 965-532-7794 or 7795
www.amideast.org

LEBANON

AMIDEAST
Ras Beirut Center Bldg.
4th Floor
Sourati Street
P.O. Box 135-155
Ras Beirut, Lebanon
Telephone: 961-1-345-341 or
961-1-750-141
www.amideast.org

or

AMIDEAST
Sannine Bldg., 1st Floor
Antelias Highway
P.O. Box 70-744
Antelias, Beirut, Lebanon
Telephone: 961-4-419-342 or
961-4-410-438
www.amideast.org

MALAYSIA/SINGAPORE

MACEE
Testing Services
8th Floor Menara John Hancock
Jalan Gelenggang
Damansara Heights
50490 Kuala Lumpur, Malaysia
Telephone: 6-03-253-8107
www.macee.org.my/

MEXICO

Institute of International Education
Londres 16, 2nd Floor
Colonia Juarez, D.F., Mexico
Telephone: 525-209-9100,
ext. 3500, 3510, 4511
www.ief.org/latinamerica/

MOROCCO

AMIDEAST
15 rue Jabal El Ayachi, Agdal
Rabat, Morocco
Telephone: 212-7-675081
www.amideast.org

PEOPLE'S REPUBLIC OF CHINA

China International Examinations
Coordination Bureau
167 Haidian Road
Haidian District
Beijing 100080
People's Republic of China
Telephone: 86 (10) 6251-3994
www.neea.edu.cn

SYRIA

AMIDEAST/Syria
P.O. Box 2313
Damascus, Syria
Telephone: 963-11-333-2804
www.amideast.org

TAIWAN

The Language Training & Testing
Center
P.O. Box 23-41
Taipei, Taiwan 106
Telephone: (8862) 2362-6045
www.lttc.ntu.edu.tw

THAILAND, CAMBODIA, AND LAOS

Institute of International Education
G.P.O. Box 2050
Bangkok 10501, Thailand
Telephone: 66-2-639-2700
www.ief.org/seasia

VIETNAM

Institute of International Education
City Gate Building
104 Tran Hung Dao, 5th Floor
Hanoi, Vietnam
Telephone: (844) 822-4093
www.ief.org/ief/vietnam

TUNISIA

AMIDEAST
BP 351 Cite Jardins 1002
Tunis-Belvedere, Tunisia
Telephone: 216-1-790-559
www.amideast.org

UNITED ARAB EMIRATES

AMIDEAST
c/o Higher Colleges of Technology
P.O. Box 5464
Abu Dhabi, UAE
Telephone: 971-2-456-720
www.amideast.org

YEMEN

AMIDEAST
Algiers St. #66
P.O. Box 15508
Sana'a, Republic of Yemen
Telephone: 967-1-206-222 or
967-1-206-942
www.amideast.org

COMMONWEALTH OF INDEPENDENT STATES

Web Site Address and
Telephone Numbers of
ASPRIAL/AKSELS/ACET
Offices

www.actr.org

RUSSIA

Leninsky Prospect 2
Office 530
P.O. Box 1
Russia 117049, Moscow
Moscow – (095) 237-91-16
(095) 247-23-21
Novosibirsk – (3832) 34-42-93
St. Petersburg – (812) 311-45-93
Vladivostok – (4232) 22-37-98
Volgograd – (8442) 36-42-85
Ekaterinburg – (3432) 61-60-34
ARMENIA, Yerevan
(IREX Office)
(8852) 56-14-10
AZERBAIJAN, Baku
(99412) 93-84-88
BELARUS, Minsk
(10-37517) 284-08-52, 284-11-70
GEORGIA, Tbilisi
(10-995-32) 93-28-99, 29-21-06
KAZAKSTAN, Almaty
(3272) 63-30-06, 63-20-56
KYRGYSTAN, Bishkek
(10-996-312) 22-18-82
MOLDOVA, Chisinau
(10-3732) 23-23-89, 24-80-12
TURKMENISTAN, Ashgabat
(993-12) [within NIS (3632)] 39-90-65
39-90-66

UKRAINE

Kharkiv – (38-0572) 45-62-46 (temporary)
(38-0572) 18-56-06
Kyiv – (044) 221-31-92, 224-73-56
Lviv – (0322) 97-11-25
Odessa – (0487) 3-15-16

UZBEKISTAN

, Tashkent

(998-712) 56-42-44
(998-71) 152-12-81, 152-12-86

OTHER COUNTRIES AND AREAS

TOEFL Publications
P.O. Box 6154
Princeton, NJ 08541-6154, USA
Telephone: 609-771-7100
www.toefl.org

ORDER FORM

Priced publications can be ordered from the TOEFL Web site at www.toefl.org/cbprpmat.html.

For **free** publications, fill out the information below and mail to:

TOEFL Publications
P.O. Box 6161
Princeton, NJ 08541-6161 USA

Free Publications *(check the titles you want)*

<input type="checkbox"/>	988021	The Researcher
<input type="checkbox"/>	253884	Guidelines for TOEFL Institutional Validity Studies
<input type="checkbox"/>	988286	Fee Voucher Service Request Form
<input type="checkbox"/>	275152	Information Bulletin for Computer-Based Testing
<input type="checkbox"/>	275155	Information Bulletin for Computer-Based Testing - Asian Edition
<input type="checkbox"/>	275154	Information Bulletin for Supplemental TOEFL Administrations
<input type="checkbox"/>	987421	Information Bulletin for the Test of Spoken English
<input type="checkbox"/>	987690	Products and Services Catalog
<input type="checkbox"/>	988709	TOEFL Score Reporting Services
<input type="checkbox"/>	281018	CBT TOEFL Test Scores - Concordance Tables
<input type="checkbox"/>	281014	Computer Familiarity and Test Performance
<input type="checkbox"/>	988236	Test Center Reference List
<input type="checkbox"/>	678020	1999-00 Test & Score Data Summary (paper-based TOEFL)
<input type="checkbox"/>	678096	1997 Test & Score Manual (paper-based TOEFL)
<input type="checkbox"/>	987634	ITP Brochure U.S.
<input type="checkbox"/>	275061	TWE Guide
<input type="checkbox"/>	987256	The Computer-Based TOEFL Test - The New Scores
<input type="checkbox"/>		I would like to be added to the TOEFL mailing list.

Delivery Address *(please print)*

Institution: _____
Attention: _____
Address: _____

E-mail Address: _____



accurate and objective scores
careful monitoring of test
commitment to international education
comprehensive research program
exhaustive pre-testing
expert advisory committees
highly skilled test developers
Internet-based score reporting
longstanding reliability and consistency
meticulous review process
official score reports with photos
ongoing test improvements
required essay
sensitivity to institutions' concerns
standardized delivery procedures
uncompromising integrity
unparalleled test design standards

TOEFL[®]

Test of English as a Foreign Language

No other test in the world is as reliable a standard for measuring a nonnative speaker's ability to use English at the university level.

= TRUST

Join our Internet mailing list at
www.toefl.org/edindx.html

www.toefl.org
1-609-771-7100



Copyright © 2000 by Educational Testing Service. Educational Testing Service, ETS, the ETS logo, TOEFL, and the TOEFL logo are registered trademarks of Educational Testing Service. Test of English as a Foreign Language is a trademark of Educational Testing Service.



Test of English as a Foreign Language

Educational Testing Service

P.O. Box 6155

Princeton, NJ 08541-6155

USA

Phone: 1-609-771-7100

E-mail: toefl@ets.org

Web site: <http://www.toefl.org>



57332-15747 • U110M50 • Printed in U.S.A.
I.N. 989551