



TOEFL.

ISSN 1930-9317

TOEFL iBT Research Report

TOEFLiBT-06
June 2008

*Linking English-Language Test
Scores Onto the Common
European Framework of
Reference: An Application of
Standard-Setting Methodology*

Richard J. Tannenbaum

E. Caroline Wylie

Listening.

Learning.

Leading.®

**Linking English-Language Test Scores Onto the Common European Framework of Reference:
An Application of Standard-Setting Methodology**

Richard J. Tannenbaum and E. Caroline Wylie
ETS, Princeton, NJ



ETS is an Equal Opportunity/Affirmative Action Employer.

As part of its educational and social mission and in fulfilling the organization's non-profit Charter and Bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

Copyright © 2008 by ETS. All rights reserved.

No part of this report 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 U.S. and international copyright laws.

ETS, the ETS logos, GRADUATE RECORD EXAMINATIONS, GRE, LISTENING. LEARNING. LEADING., TOEFL, the TOEFL logo, and TOEIC are registered trademarks of Educational Testing Service (ETS). The TEST OF ENGLISH AS A FOREIGN LANGUAGE is a trademark of ETS.

TEST OF ENGLISH FOR INTERNATIONAL COMMUNICATION and TOEIC BRIDGE are trademarks of ETS.

College Board is a registered trademark of the College Entrance Examination Board.

Abstract

The Common European Framework of Reference (CEFR) describes language proficiency in reading, writing, speaking, and listening on a 6-level scale. In this study, English-language experts from across Europe linked CEFR levels to scores on three tests: the TOEFL[®] iBT test, the TOEIC[®] assessment, and the TOEIC *Bridge*[™] test. Standard-setting methodology (a modified Angoff approach and a modified examinee paper selection approach) was used to construct the linkages. Linkages were established for TOEFL iBT at levels B1, B2, and C1. Linkages were established for TOEIC at levels A1 through C1, with the exception of Reading at the C1 level. The TOEIC *Bridge* test was linked to its three targeted levels of the CEFR. The report details the methods, procedures, and results of the study.

Key words: English-language tests, TOEFL, TOEIC, TOEIC *Bridge* test, Common European Framework of Reference for Languages (CEFR), standard setting

The Test of English as a Foreign Language™ (TOEFL®) was developed in 1963 by the National Council on the Testing of English as a Foreign Language. The Council was formed through the cooperative effort of more than 30 public and private organizations concerned with testing the English proficiency of nonnative speakers of the language applying for admission to institutions in the United States. In 1965, Educational Testing Service (ETS) and the College Board® assumed joint responsibility for the program. In 1973, a cooperative arrangement for the operation of the program was entered into by ETS, the College Board, and the Graduate Record Examinations® (GRE®) Board. The membership of the College Board is composed of schools, colleges, school systems, and educational associations; GRE Board members are associated with graduate education. The test is now wholly owned and operated by ETS.

ETS administers the TOEFL program under the general direction of a policy board that was established by, and is affiliated with, the sponsoring organizations. Members of the TOEFL Board (previously the Policy Council) represent the College Board, the GRE Board, and such institutions and agencies as graduate schools of business, two-year colleges, and nonprofit educational exchange agencies.



Since its inception in 1963, the TOEFL has evolved from a paper-based test to a computer-based test and, in 2005, to an Internet-based test, TOEFL iBT. One constant throughout this evolution has been a continuing program of research related to the TOEFL test. From 1977 to 2005, nearly 100 research and technical reports on the early versions of TOEFL were published. In 1997, a monograph series that laid the groundwork for the development of TOEFL iBT was launched. With the release of TOEFL iBT, a TOEFL iBT report series has been introduced.

Currently this research is carried out in consultation with the TOEFL Committee of Examiners. Its members include representatives of the TOEFL Board and distinguished English as a second language specialists from the academic community. The Committee advises the TOEFL program about research needs and, through the research subcommittee, solicits, reviews, and approves proposals for funding and reports for publication. Members of the Committee of Examiners serve four-year terms at the invitation of the Board; the chair of the committee serves on the Board.

Current (2007-2008) members of the TOEFL Committee of Examiners are:

Alister Cumming (Chair)	University of Toronto
Geoffrey Brindley	Macquarie University
Frances A. Butler	Language Testing Consultant
Carol A. Chapelle	Iowa State University
Catherine Elder	University of Melbourne
April Ginther	Purdue University
John Hedgcock	Monterey Institute of International Studies
David Mendelsohn	York University
Pauline Rea-Dickins	University of Bristol
Mikyuki Sasaki	Nagoya Gakuin University
Steven Shaw	University of Buffalo

To obtain more information about the TOEFL programs and services, use one of the following:

E-mail: toefl@ets.org
Web site: www.ets.org/toefl

Acknowledgments

We extend our sincere appreciation to our colleagues who provided thoughtful reviews and offered sound advice on an earlier version of this report: Don Powers, Yasuyo Sawaki, Ari Huhta, and Glenn Fulcher. We also thank to our colleagues Jenny Dalalakis and Craig Stief, who provided invaluable contributions both before and during the study. We also thank the TOEFL and TOEIC business areas at ETS for their support of this work.

Table of Contents

	Page
Introduction.....	1
Standard Setting.....	2
English-Language Tests.....	5
Methodology.....	7
Panelist Familiarization.....	7
During the Study.....	8
Standard Setting.....	8
Panel 1 (the TOEFL iBT Test).....	12
Panelists.....	13
Standard-Setting Results for the TOEFL iBT Test.....	14
Final Evaluation of the TOEFL iBT Standard-Setting Process and Cutscores.....	18
Panel 2 (the TOEIC Test).....	19
Panelists.....	19
Standard-Setting Results for the TOEIC Test and the TOEIC <i>Bridge</i> Test.....	20
Final Evaluation of TOEIC Test and TOEIC <i>Bridge</i> Test Cutscores.....	25
Conclusions.....	28
References.....	32
List of Appendixes.....	35

List of Tables

	Page
Table 1. Panel 1 Demographics	13
Table 2. TOEFL Writing Judgments.....	14
Table 3. TOEFL Speaking Judgments	16
Table 4. TOEFL Listening Judgments.....	16
Table 5. TOEFL Reading Judgments.....	17
Table 6. Prompts on the Final TOEFL Evaluation Form.....	18
Table 7. Level of Influence of Information Sources for TOEFL iBT Standard Setting.....	19
Table 8. Panel 2 Demographics	20
Table 9. TOEIC Writing Judgments	21
Table 10. TOEIC Speaking Judgments.....	22
Table 11. TOEIC Listening Judgments.....	23
Table 12. TOEIC Reading Judgments	23
Table 13. TOEIC <i>Bridge</i> Test Reading Judgments.....	25
Table 14. TOEIC <i>Bridge</i> Test Listening Judgments.....	25
Table 15. Prompts on the Final TOEIC Evaluation Form	26
Table 16. Level of Influence of Information Sources for TOEIC Standard Setting.....	26
Table 17. Level of Influence of Information Sources for the TOEIC <i>Bridge</i> Test Standard Setting	27
Table 18. Scaled-Score Cutscore Results for the TOEFL iBT Test	29
Table 19. Scaled-Score Cutscore Results for the TOEIC Test	29
Table 20. Scaled-Score Cutscore Results for the TOEIC <i>Bridge</i> Test	29

Introduction

The Common European Framework Reference for Languages: Learning, Teaching, Assessment (CEFR) is intended to overcome the barriers to communication among language instructors, educators, curriculum designers, and agencies working in the field of language development by providing a common basis for describing and discussing stages of language development and the skills needed to reach different levels of language proficiency (The Common European Framework of Reference). The CEFR describes language proficiency in reading, writing, speaking, and listening on a 6-level scale, clustered in three bands: A1–A2 (Basic User), B1–B2 (Independent User), and C1–C2 (Proficient User).

The CEFR scales are becoming accepted in Europe as one means of reporting the practical meaning of test scores in ways that have a socially constructed meaning for teachers and other test-score users. That is to say, if a test score can be mapped (linked) to one of the levels of the CEFR, it becomes clearer what that score means—what candidates with at least that score are likely able to do. However, the usefulness of the CEFR should be considered with an understanding of some of its shortcomings. Weir (2005) cautioned that while the CEFR provides valuable information on language proficiency, the level descriptors do not offer sufficient information about how contextual factors affect performance across the levels, which he termed *context validity*. He also expressed concern that the CEFR does not adequately delineate how language develops across the levels in terms of cognitive or meta-cognitive processing. He concluded, “It is crucial that the CEFR is not seen as a prescriptive device but rather a heuristic, which can be refined and developed by language testers to better meet their needs. . . . It currently exhibits a number of serious limitations such that comparisons based entirely on the scales alone might prove to be misleading, given the insufficient attention paid in these scales to issues of validity” (p. 298). These criticisms notwithstanding, the CEFR is widely accepted as the benchmark against which language tests used across Europe should be compared.

The purpose of this study was to identify minimum scores (cutscores) on two English-language tests (the TOEFL® iBT test and TOEIC® assessment) that correspond to the A1 through C2 proficiency levels of the CEFR. Minimum scores were to be identified separately for the Speaking, Writing, Listening, and Reading sections of the two assessments. Minimum scores corresponding to the CEFR levels A1, A2, and B1 were also to be identified for a third test: the TOEIC *Bridge*™ Test (Listening and Reading sections.) By mapping test scores onto the CEFR,

an operational bridge is built between the descriptive levels of the CEFR and psychometrically sound, standardized assessments of English-language competencies, facilitating meaningful classification of CEFR-based communicative competence as well as tracking progress in English-language development. The study was not intended or designed, however, to establish a concordance between scores on the series of English-language tests, such that scores on one test could be used to identify comparable scores on the another test. Scores from each test were independently mapped to the CEFR levels; no attempt was made to link scores or score distributions across the tests.

Standard Setting

The process followed to map test scores onto the CEFR is known as *standard setting*. Standard setting is a general label for a number of approaches commonly used to identify test scores that support decisions about test takers' (candidates') level of knowledge, skill, proficiency, mastery, or readiness. Standard setting, according to Cizek (1993), is "the proper following of a prescribed, rational system of rules or procedures resulting in the assignment of a number to differentiate between two or more states or degrees of performance" (p. 100). There is no true cutscore; a cutscore is not the equivalent of a statistic that serves as an estimate for the true value in the population of interest (Zieky, 2001). Kane (2001) reinforced this point: "... standard setting still cannot be reduced to a problem of statistical estimation. Fundamentally, standard setting involves the development of a policy about what is required for each level of performance. This policy is stated in the performance standards and implemented through the cutscores" (p. 85). Cizek and Bunch (2007) summarized the judgmental nature of standard setting thusly: "To some degree, then, because standard setting necessarily involves human opinions and values, it can also be viewed as a nexus of technical, psychometric methods and policy making" (p. 18).

The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) offer guidelines for standard setting. The rationale and procedures for a standard-setting study should be clearly documented. This includes the method implemented, the selection and qualifications of the panelists, and the training provided. With respect to training, panelists should understand the purpose and goal of the standard-setting process (e.g., what decision or classification is being made on the basis of the test score), be

familiar with the test, have a clear understanding of the judgments they are being asked to make, and have an opportunity to practice making those judgments; and the standard-setting process should be designed so that the panelists may bring to bear their knowledge and experience in a reasonable way. The standard-setting procedures in this study were designed to comply with these guidelines.

Recent reviews of research on standard-setting approaches also reinforce a number of core principles for best practice: careful selection of panel members and a sufficient number to represent varying perspectives, sufficient time devoted to ensure development of a common understanding of the domain under consideration, use of an appropriate standard-setting methodology that allows for adequate training of judges, development of a description of each performance level, multiple rounds of judgments, and the inclusion of empirical data where appropriate to inform judgments (Brandon, 2004; Hambleton & Pitoniak, 2006). The approaches used in this study adhere to all of these guidelines.

In 2003 the Council of Europe published preliminary guidelines for linking language examinations to the CEFR (Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEF), Council of Europe, 2003). In those guidelines they described four interrelated sets of procedures: familiarization (of panelists with the CEFR), specification (evaluation of the alignment of test content with the CEFR), standardization of judgments, and empirical validation. This report of the standard-setting approach used to link the three English-language tests will focus primarily on the familiarization and standardization of judgments aspects of the linking procedure.

Our emphasis on standard setting, which is subsumed under Standardisation of Judgments (chapter 5 of the manual), was purposeful. The applied research objective for this study was constructing cutscores on the tests to differentiate among the levels of the CEFR. This study did not seek evidence of the extent to which the content of the tests, per se, matched the descriptive categories of the CEFR, which Figueras, North, Takala, Verhelst, and Van Avermaet (2005) associated with specification. It may be argued that not gathering direct judgments of this alignment brings into question the relevance of standard setting. In a broader context this raises the specter of potential construct incongruence between the defined proficiency levels (as reflected by the cutscores) and what the test actually measures (Mills & Jaeger, 1998). In this regard, it is important to acknowledge that the tests were not designed a priori to map to the

CEFR, and so the potential for not mapping to all the levels of the CEFR was recognized from the beginning. Nonetheless, because the CEFR aims to describe a robust framework for language that is “not irrevocably and exclusively attached to any one of a number of competing linguistic or educational theories or practices” (CEFR, p. 8) and because the tests were designed to measure the major language modalities (reading, writing, speaking, and listening), significant construct incongruence was not considered pervasive enough to preclude standard setting.

Evidence of procedural and internal validity was collected through panelist evaluation forms during the standard-setting process, and through analyses of the standard-setting judgments (discussed later in the report). *Procedural validity*, in part, addresses panelists’ understanding of the standard-setting process and their judged reasonableness of the cutscore outcomes; *internal validity*, in part, addresses documentation of the consistency of standard-setting judgments (see Hambleton & Pitoniak, 2006, for full details). *External validity* refers to comparing the cutscores values to other sources of information. The Council of Europe (2003) referred to this type of evidence as *empirical validity*, though we prefer, as did Kane (2001), to characterize this as *convergent validity evidence*. A pattern of convergent evidence would serve to fortify the reasonableness of the cutscores, though a divergence of outcomes between two or more sources of information would not necessarily mean that the panel-constructed cutscores were cause for concern. While there certainly is value in efforts to obtain convergent validity evidence, that was not part of this standard-setting study. But as Kane (1994) reassured in the context of credentialing assessment, a “well-designed and carefully conducted standard-setting study is likely to provide as good an indication of the most appropriate passing score as any other source of information” (p. 448).

Before moving on, it is important to point out that linking test scores to the CEFR is no mean feat, and doing so poses challenges from more than a methodological perspective. The concerns of Weir (2005) regarding the current state of the CEFR led him to report that “it is not surprising that a number of studies have experienced difficulty attempting to use the CEFR for test development or comparability purposes” (p. 283). But surely the imperfect nature of the CEFR is not the sole cause; any test is also an imperfect representation of its intended construct. Asking panelists to create an interpretative bridge between the CEFR and a test, particularly a test not designed a priori to measure the CEFR, should not be taken for granted, and should

appropriately be considered and treated as a research-based question. We now turn our attention to a brief description of the three tests that were considered in this study.

English-Language Tests

Three separate English-language tests were considered in this study: the TOEFL iBT test, TOEIC test, and TOEIC *Bridge* test. Each test will be described in turn.

The TOEFL iBT test. The TOEFL iBT test measures the ability of non-native speakers of English to communicate orally and in writing in English, to understand English as it is spoken in North American academic contexts (listening skill), and to understand short passages similar in topic and style to academic texts used in North American colleges and universities (reading skill).

Writing and speaking are tested in two constructed-response sections. The Speaking section consists of six tasks. Two are independent speaking tasks that require the candidate to respond to a question on a familiar topic, and four are integrated tasks that require candidates to combine both written and spoken information. Topics are either related to campus life or lecture situations. The Writing section consists of an opinion essay and an integrated essay, which requires candidates to listen to a brief lecture, read a short passage, and then respond in writing.

Listening and reading are assessed through selected-response (multiple-choice) items. In the Reading section, candidates are presented with three reading passages with approximately 14 questions per passage. These questions address basic comprehension, making inferences, and reading-to-learn skills. In the Listening section, there are six tasks: Two are dialogues that are based on campus-life situations and four are lecture scenarios with five or six questions per listening passage. The questions address basic understanding, pragmatic understanding, and connecting information. The questions are presented orally and in written form; answer choices are in written form only. Each separate test section (Listening, Reading, Writing, and Speaking) is reported on a scale that ranges from a low of 0 to a high of 30. The TOEFL iBT reliability (internal consistency) estimates are: .86, Reading; .87, Listening; .90, Speaking; and .78, Writing.

The TOEIC test. The TOEIC test measures the ability of non-native English communicators to communicate in English in the global workplace. The TOEIC test addresses listening comprehension skills and reading comprehension skills. The TOEIC test is a selected-response test and test items are developed from samples of spoken and written English from

countries around the world. Each section (Listening and Reading) is reported on a scale that ranges from a low of 5 to a high of 495.

Two optional modules have been added to the TOEIC test that address speaking and writing skills. The Writing section measures a test taker's ability to communicate clearly in written English (general communication skills to function in a workplace). There are three separate task types: sentence-level writing; responding to a written request; and writing an opinion essay. Scores on each task type are weighted, with the sentence-level tasks contributing the least weight and the opinion essay contributing the greatest weight.

The Speaking section measures a candidate's ability to speak clearly, to communicate, and to demonstrate understanding of material. There are six tasks of increasing complexity, and a weighting system is applied to scores, with the more basic tasks contributing less weight than the more complex tasks. Task 1 (considered the most basic task) consists of a read-aloud section that is scored for both pronunciation and intonation. In Task 2, candidates describe a picture. Task 3 requires candidates to respond to a series of questions with an audio lead-in that sets the context. Task 4 is similar, except that the questions are based on textual information. In Task 5, candidates are presented with a problem that is in audio form only and they are asked to propose a solution. Task 6 (the most heavily weighted task) requires them to give an opinion on a topic that is presented in the form of audio and text stimulus. Results for the Writing and Speaking sections are each reported on a scale that ranges from a low of 0 to a high of 200. The TOEIC reliability (internal consistency) estimates are: .93, Reading; .92, Listening; .89, Speaking; and .77, Writing.

The TOEIC Bridge test. The TOEIC *Bridge* test measures the listening-comprehension and reading-comprehension skills of non-native English communicators at a more basic level than the TOEIC test. The test consists of selected-response items, and results are reported on a scale that ranges from 10 to 90. The TOEIC *Bridge* test reliability (internal consistency) estimates are .86 for Reading and .86 for Listening.

Report structure. The remainder of this report is presented in three major sections. The first section describes the standard-setting methods that were implemented to establish the cutscores corresponding to the CEFR proficiency levels on each of the English-language tests. The second section focuses on the results specific to the TOEFL iBT test. The third section focuses on the TOEIC test and the TOEIC *Bridge* test. It is important that panelists are familiar with not only the test and its use but also with the test population. This background will help

inform their standard-setting judgments. To this end, two different panels of experts (with minimal overlap) were convened to participate in setting the cutscores on the tests—one panel for the TOEFL iBT test, and one panel for the TOEIC test and the TOEIC *Bridge* test. The two panels reflected the different contexts for which the tests are primarily used: Panel 1 (TOEFL iBT test; higher education) and Panel 2 (TOEIC test and the TOEIC *Bridge* test; business). A small number of panelists ($n = 5$) had experience with both sets of tests and test populations, and served on both panels. The composition of each panel is discussed in more detail at the start of the sections on the panels.

Methodology

This section will outline the procedures followed to help ensure that the committee members were familiar with, and had common, agreed-upon understandings of the CEFR-level descriptions for each language modality, and were prepared sufficiently to apply the standard-setting methodologies to both the constructed-response and selected-response test sections.

Panelist Familiarization

Familiarization relates to the panelists having a clear understanding of the CEFR. This was accomplished in two ways: first, before the panel members came to the study they were given an assignment to complete; and second, during the study itself, the members engaged in extensive discussions about each language modality as described by the CEFR. Both stages are described below.¹

Before the study. Prior to the study, the members on both panels were given an assignment (see Appendix A) to review selected tables from the CEFR (the Web site to the CEFR was provided) for each language modality and to write down key characteristics or indicators from the tables that described an English-language learner (candidate) with *just enough skills* to be performing at each CEFR level. The tables were selected to provide the panelists with a broad understanding of what learners were expected to be able to do for each of the language modalities. The selected CEFR tables for Speaking, for example, were Overall Oral Production (pages 58–59 of the CEFR), Overall Spoken Interaction, Understanding a Native Speaker Interlocutor, Conversation, and Informal Discussion With Friends (pages 74–77).

As they completed this pre-study assignment, they were asked to consider what distinguishes a candidate with just enough skills to be considered performing at a specific CEFR

level from a candidate with not enough skills to be performing at that level. For example, they were asked to consider what the least able C2 speaker can do that the highest performing C1 speaker cannot do, what the least able C1 speaker can do that the highest performing B2 speaker cannot do, and so on. The assignment was intended as part of a calibration of the members to a shared understanding of the minimum requirements for each of the CEFR levels.

During the Study

During the study, time was spent developing an agreed upon definition of the minimum skills needed to be considered performing at each CEFR level. The panelists were formed into three table groups and each group was asked to define and chart the skills of the least able candidate for A2, B2, and C2 levels; this was done separately for Writing, Speaking, Listening, and Reading. Panelists referred to their pre-study assignments and to the CEFR tables for each modality. Given that the focus for the standard setting was on the candidate who has *just enough* skills to be at a particular level, panelists were reminded that the CEFR describes the abilities of someone who is *typical* of a particular level. In particular, it was noted that some of the levels are divided into sublevels, so panelists were careful to pull phrases from the lower rather than the higher part of the level when thinking of the skills that would be possessed by a candidate who has *just* entered a particular level.

A whole-panel discussion of each group's charts followed, and a final agreed upon definition was established for three levels: A2, B2, and C2. Definitions of the least able candidate for A1, B1, and C1 levels were then accomplished through whole-panel discussion, using the A2, B2, and C2 descriptions as boundary markers. As before, the panelists also referred to their pre-study assignment and the relevant CEFR tables. These definitions served as the frame of reference for the standard-setting judgments; that is, panelists were asked to consider the test items in relation to these definitions. See Tables B1 through B8 in Appendix B for copies of each panel's agreed upon definitions. The definitions included short-hand notations and phrases that were based on the panelists' extended discussions; they were meant only to remind the panelists of the salient discussion points.

Standard Setting

Selected-response sections. A modified Angoff approach (Angoff, 1971; Brandon, 2004)—consistent with the standard-setting process outlined in the *Manual for Relating*

Language Examinations to the CEFR (Council of Europe, 2003)—was implemented for reading and listening modalities measured using selected-response items. Panelists were trained in the process and then given opportunity to practice making their judgments to ensure that they understood the procedure. This practice opportunity was also used to clarify any misunderstandings of the judgment process. At this point, panelists were formally asked to acknowledge if they understood what they were being asked to do and the overall judgment process. They did this by signing a training evaluation form confirming their understanding and readiness to proceed. In the event that a panelist was not yet prepared to proceed, he or she would have been given additional training by one of the ETS facilitators. All panelists signed off on their understanding and readiness to proceed.

Then they went through three rounds of operational judgments, with feedback and discussion between rounds. For each item, panelists were asked to consider the agreed upon definition of just-qualified (least able) candidates (for A2, B2, C2) and to judge the probability that a just-qualified (least able) candidate would have the skills needed to answer the item correctly. In order to facilitate setting six cutscores on each modality, panelists initially focused on A2, B2, and C2 levels; once established, these cutscores formed the boundaries for the A1, B1, and C1 cutscores. For the TOEIC *Bridge* test, only three CEFR levels were considered: A1, A2, and B1. These levels were identified a priori by ETS language experts as being most relevant for the TOEIC *Bridge* test. The cutscores for these levels were derived following the same procedure used to derive cutscores for A2, B2, and C2 levels on the TOEFL iBT test and the TOEIC test.

For the first round of judgments, panelists were asked to estimate the probability that a just-qualified candidate would answer the item correctly. They made these judgments using the following judgment scale (expressed as probabilities): 0, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 95, 100. The higher the probability, the easier a panelist believed an item was for the just-qualified A2, B2, or C2 candidate. This approach was used rather than asking panelists to judge whether a just-qualified candidate would answer the item correctly or incorrectly, which Impara and Plake (1997) referred to as a *yes/no variation of an Angoff approach*. Although perhaps a cognitively simpler task for panelists, the yes/no approach may increase the likelihood of positive or negative bias in the item judgments (Cizek & Bunch, 2007), so the more continuous judgment approach was implemented. (As will be presented later, panelists reported no problems understanding the Angoff process implemented or in making their judgments.) Panelists were

instructed to focus only on the alignment between the skill demanded by the item and the skill possessed by a just-qualified candidate, and not to factor guessing into their judgments. The focus on skill-to-skill alignment maintained the criterion-based nature of the standard-setting process; it is likely, however, that had adjustments for guessing been introduced, the largest positive effect would have been observed on the A1 and A2 levels.

The sum of each panelist's cross-item judgments represents his or her recommended cutscore. Each panelist's recommended cutscore was provided to the panelist at the end of the first round of judgments. The panel's average (panel's recommended cutscore), and the highest and lowest cutscores (individual panelists were unidentified) were compiled and presented to the panel to foster discussion. Panelists were then asked to share their judgment rationales.

As part of the feedback and discussion, item performance information (P+ values) was shared. Item P+ values are the average scores of the test-taking population on a particular form. For selected-response items the P+ value is reported as a value between 0 and 1. The higher a P+ value, the more candidates who identified the correct answer. In addition, P+ values were calculated for candidates scoring at or above the 75th percentile on that particular section (i.e., the top 25% of candidates) and for candidates at or below the 25th percentile (i.e., the bottom 25% of candidates). Examining item difficulty for the top 25% of candidates and the bottom 25% of candidates was intended to give panelists a better understanding of the relationship between overall language ability for that modality (total section score) and each of the items. The partitioning, for example, enabled panelists to see any instances where the same item was comparably difficult regardless of test takers' overall language ability, or where an item was found to be particularly challenging or easy for test takers at the different ability levels.

The item data for the TOEFL iBT test was based on more than 5,000 test takers; for the TOEIC test, it was based on more than 100,000 test takers; and for the TOEIC *Bridge* test, it was based on more than 4,000 test takers.

Before making their Round 2 judgments, panelists were asked to consider their peers' rationales and the normative information. For Round 2, judgments were made not at the item level but at the overall level of the modality; that is, panelists were asked to consider if they wanted to recommend a different section-level score for A2, B2, and C2. The transition to the section (modality) level introduced a shift from discrete items to the overall construct of interest. This holistic approach seemed more relevant and appropriate to the language construct than did

deconstructing the construct through another series of item-level judgments. Panelists had no difficulty with the holistic approach; this approach had also been used in a previous CEFR linking study (Tannenbaum & Wylie, 2005).

After making their second round of judgments, similar feedback was provided; in addition, the percentage of candidates who were classified into each of the three levels was presented. (These classifications were based on the same sets of data used to generate the item-level information.) The Round 2 average judgments for A2, B2, and C2 were applied to existing test-score distributions for the modality of focus, and the percentages of candidates classified into each level was presented and discussed. Following this level of feedback, the panelists had a final opportunity to change their section-level recommended cutscores.

The final judgments for A2, B2, and C2 were compiled and shared with the panelists; they were then asked to slot in the A1, B1, and C1 levels. Specifically, they were asked to review the A1, B1, and C1 descriptions of just-qualified candidates and to identify the minimum section-level scores for candidates just performing at these levels. Their judgments were constrained by the now-established A2, B2, and C2 cutscores. Panelists had an opportunity to discuss whether they considered any of the threshold proficiency levels to be located closer to one boundary than another. Once there had been a wide-ranging discussion, panelists then made their final individual judgments as to the minimum score associated with the A1, B1, and C1 levels.

Constructed-response sections. A modified examinee paper selection method (performance-sample approach) was implemented for the speaking and writing modalities measured using constructed-response items (Hambleton, Jaeger, Plake, & Mills, 2000). As with the modified Angoff approach, three rounds of judgments took place, with feedback and discussion, informed by data (average item scores within a section, and average items scores for candidates scoring at or below the 25th percentile and at or above the 75th percentile within a section, and classification information).

Panelists were asked to review the scoring rubrics and then to review (listen to or read) 11 samples of candidate performance at various points along the raw point scale for that modality. The samples (profiles of performance) were chosen to reflect a range of ability levels and were ordered from low-scoring to high-scoring performances. Panelists were presented with a table summarizing the item scores that formed each of the 11 performance profiles to facilitate their judgments. Panelists were then asked to identify the performance profile score for that

modality that would be expected of just-qualified candidates. Once the three rounds for A2, B2, and C2 were completed, the panelists were, as before, asked to slot the A1, B1, and C1 levels.

For the constructed-response standard-setting method, panelists were also formally asked to acknowledge if they understood the overall judgment process and specifically what they were being asked to do, by signing a training evaluation form confirming their understanding and readiness to proceed. All panelists signed off on their understanding and readiness to proceed.

For all rounds of judgment (except Round 1 for selected-response sections), panelists had the option of writing N/A (*not applicable*) for a cutscore if they deemed that the test section was not challenging enough to reach the upper levels of the CEFR, or if the test section was too challenging for candidates at the lower CEFR levels. In order for a cutscore to be reported, at least 67% of the panel had to make a cutscore recommendation. During the between-round discussions, the panel was informed of the number of panelists who had indicated N/A for any of the cutscore recommendations.

All cutscore decisions and subsequent discussions were based on raw scores, or the number of points expected to be earned by a just-qualified candidate on the form of the test reviewed. The results presented in the sections that follow are the scaled-score equivalents of the recommended raw cutscores.

In the end, each modality (reading, listening, speaking, and writing) may have up to six recommended cutscores, each identifying the minimum score for that modality believed necessary to be performing at A1, A2, B1, B2, C1, and C2 levels. Panelists had a final opportunity to review all cutscores and to comment on their level of comfort with the decisions.

The same process was followed for the TOEFL iBT test and the TOEIC test, while only the selected-response approach was needed for the Listening and Reading sections of the TOEIC *Bridge* test. The agendas in Appendix C show how time was allocated to the development of a common understanding of the CEFR by modality, training, and practice opportunities for the standard-setting approaches, panelists' judgments, and discussion of results after each round.

Panel 1 (the TOEFL iBT Test)

The committee members who formed Panel 1 conducted the linking study for the TOEFL iBT test. This section describes the makeup of the panel and presents the results of their work.

Panelists

Twenty-three experts representing 16 countries served on the panel that focused on mapping scores from the TOEFL iBT test onto the CEFR. The English-language specialist from ETS Europe, located in Brussels, organized the recruitment of the experts. The experts were selected for their experience with English-language instruction, learning, and testing, and their familiarity with the CEFR, the TOEFL assessment, and the test population. They were also selected to represent an array of European countries where the TOEFL test is used. Table 1 presents the demographic characteristics of the 23 panelists. Appendix D provides the panelists' affiliations.

Table 1

Panel 1 Demographics

		Number	Percent
Gender	Female	13	57%
	Male	10	43%
Selection Criteria ²	ESL teachers at language school (private or university)	18	
	Administrator of school/program where ELL classes are taught	9	
	Assessment expert or researcher	6	
	Educational consultant	5	
Country	Belgium	3	13%
	France	1	4%
	Germany	3	13%
	Greece	1	4%
	Hungary	1	4%
	Malta	1	4%
	The Netherlands	2	9%
	Norway	1	4%
	Poland	1	4%
	Russia	1	4%
	Slovakia	1	4%
	Spain	1	4%
	Sweden	1	4%
	Turkey	2	9%
	United Arab Emirates	1	4%
	United Kingdom	2	9%

Standard-Setting Results for the TOEFL iBT Test

The first-, second-, and third-round judgments for A2, B2, and C2 section-level judgments, along with the single round of judgments for A1, B1, and C1 (referred to as Round 4) are presented in Tables E1 through E4 in Appendix E. Each panelist's individual cutscores are presented for each round, as are the cross-panel summary statistics (mean, median, standard deviation, minimum, maximum, and standard error of judgment).

Table 2 presents the cross-panel mean judgments and the standard deviations by round for the TOEFL iBT Writing section. The second, third, and fourth columns present the results for the first three rounds where panelists focused on the A2, B2, and C2 cutscores. For the fourth round, the focus shifted to A1, B1, and C1, with panelists working within the constraints of the just-established A2, B2, and C2 cutscores. Raw scores on the Writing section range from 0 to 10, and scaled scores from 0 to 30. Scaled scores for A2, B2, and C2 are based on the Round 3 mean judgments; the scaled scores for A1, B1, and C1 are based on the Round 4 judgments.

Table 2
TOEFL Writing Judgments

CEF levels	Round 1 mean (SD)	Round 2 mean (SD)	Round 3 mean (SD)	Round 4 mean (SD)	Final scaled score
A1	-	-	-	N/A	N/A
A2	2.74 (1.60)	2.84 (1.48)	2.77 (1.14)	-	11
B1				5.17 (0.36)	17
B2	6.33 (1.08)	6.52 (0.79)	6.48 (0.68)		21
C1				9.00 (0.10)	28
C2	9.21 (0.79)	N/A	N/A		N/A

The standard deviations (SD) provide an indication of the variability among the panelist's ratings. It is typical that the greatest amount of variability occurs in the first round, before panel discussion and the presentation of item-level and classification information. After the feedback and discussion is provided, it is typical for convergence (a reduction in variability) to occur across the rounds. It is also possible to provide a rough approximation of the replicability of the

cutscores using the standard error of judgment (SEJ)³ (SD/\sqrt{P}) (Cizek & Bunch, 2007), where P is the number of panelists. If the standard-setting study were replicated (same processes and same types and numbers of panelists), the observed cutscores from the replications would be within 1 SEJ of the current cutscores about 68% of the time and within 2 SEJs about 95% of the time. The SEJs for the recommended A2, B1, B2, and C1 cutscores are .24, .07, .14, and .10, respectively. (The Level 1 judgments were the least independent and were restricted in potential range by the recommended Level 2 cutscores; therefore, the SEJ estimates are more likely to be lower than the estimates for Level 2 judgments.)

During the training for the constructed-response standard-setting approach, panelists had been informed that they could assign *not applicable* (N/A) if they felt that the test section was not challenging enough to reach the upper levels of the CEFR, or if the test section was too challenging for candidates at the lower CEFR levels. For a cutscore to be computed at any level, at least 67% of the group had to rate it as being applicable. As Table 2 shows, while in the first round at least 67% of the panelists indicated a possible C2 cutscore, after the discussion this percentage dropped, so no C2 cutscore was computed in Rounds 2 or 3. Furthermore, in the final round, an insufficient numbers of the panelists deemed the writing tasks to be accessible to a candidate at the A1 level, and so no A1 cutscore was computed.

Table 3 presents the cross-panel mean judgments and the standard deviations by round for the TOEFL iBT Speaking section. Raw scores on the Speaking section range from 0 to 24, and scaled scores from 0 to 30. As can be seen in the table, an insufficient number deemed the Speaking section to be able to differentiate C2-level performance. Consistent with expectations, convergence of judgments was observed across the rounds. The SEJs for the recommended cutscores for A1, A2, B1, B2, and C1 are .14, .30, .16, .31, and .16, respectively.

Table 4 presents the cross-panel mean judgments and the standard deviations by round for the TOEFL iBT Listening section. Raw scores on the Listening section range from 0 to 34, and scaled scores from 0 to 30. The variability in panelists' judgments remained more stable than was observed for the Writing and Speaking modalities; this may have been do to the difference in the judgment task, going from constructed-response items to selected-response items. The SEJs for the recommended cutscores for A2, B1, B2, and C1 are .50, .34, .64, and .22, respectively.

Table 3***TOEFL Speaking Judgments***

CEF levels	Round 1 mean (SD)	Round 2 mean (SD)	Round 3 mean (SD)	Round 4 mean (SD)	Final scaled score
A1	-	-	-	5.75 (0.68)	8
A2	9.55 (2.46)	9.86 (1.64)	9.95 (1.43)	-	13
B1	-	-	-	14.52 (0.79)	19
B2	17.55 (2.18)	17.65 (1.56)	17.78 (1.51)	-	23
C1	-	-	-	21.95 (0.79)	28
C2	N/A	N/A	N/A	-	N/A

Table 4***TOEFL Listening Judgments***

CEF levels	Round 1 mean (SD)	Round 2 mean (SD)	Round 3 mean (SD)	Round 4 mean (SD)	Final scaled score
A1	-	-	-	N/A	N/A
A2	2.54 (2.38)	3.19 (2.34)	3.27 (2.33)	-	N/S ⁴
B1	-	-	-	16.73 (1.58)	13
B2	25.10 (3.86)	25.05 (2.73)	26.41 (3.02)	-	21
C1	-	-	-	30.95 (1.02)	26
C2	33.31 (0.73)	33.13 (0.96)	N/A	-	N/A

For the Listening section, panelists initially awarded a C2 cutscore (using the Angoff approach for item-level judgments) in Rounds 1 and 2, although three panelists had cutscores of 34, meaning that they expected a just-qualified C2-level candidate to answer every item correctly. As discussions ensued, an increasing number decided that the test section did not sufficiently stretch a candidate in order to be sure that someone was at a C2 rather than C1 level. After the third round of judgments, fewer than 67% supported a cutscore for this level, so no cutscore was computed. Panelists conducted all discussions of each test section considering raw scores, and they considered that an A2 candidate would be able to answer correctly approximately three questions

on this section. However, there is no A2 scaled cutscore because the raw score was too low to provide a corresponding scaled score. Thus, there is no A1 cutscore either.

Table 5 presents the cross-panel mean judgments and the standard deviations by round for the TOEFL iBT Reading section. Raw scores on the Reading section range from 0 to 45, and scaled scores from 0 to 30. Similar to the Listening section, while panelists determined an A2 cutscore in the raw scale, it was too low to provide a corresponding scaled score; consequently, no A1 cutscore could be determined either. The variability in panelists' judgments remained relatively stable, consistent with what was observed for Listening. The SEJs for the recommended cutscores for B1, B2, C1, and C2 are .68, .81, .55, and .36, respectively.

Table 5
TOEFL Reading Judgments

CEF levels	Round 1 mean (SD)	Round 2 mean (SD)	Round 3 mean (SD)	Round 4 mean (SD)	Final scaled score
A1	-	-	-	N/A	N/A
A2	1.90 (3.06)	1.38 (1.41)	1.43 (1.91)	-	N/S
B1	-	-	-	13.59 (3.17)	8
B2	29.68 (6.28)	28.26 (3.95)	29.32 (3.91)	-	22
C1	-	-	-	39.91 (2.56)	28
C2	42.99 (1.65)	42.65 (1.62)	42.80 (1.74)	-	29

In summary, across the four modalities, the cutscore (A2, B2, and C2) means changed very little from Round 1 to Round 2 or from Round 2 to Round 3; for the constructed-response sections the variability (standard deviation) of the panelists' judgments tended to decrease from Round 1 to Round 2 and from Round 2 to Round 3, indicating a greater degree of panelist convergence. They tended to remain more stable for the selected-response sections. The third- and fourth-round mean scores may be accepted as the panel-recommended cutscores; that is, they reflect the minimum scores necessary to qualify for the A1 through C2 levels on the CEFR. As noted from the previous tables for the TOEFL iBT test, panelists did not believe that the Writing, Listening, and Reading sections of the test were accessible to just-qualified candidates at the A1 level. The panelists expressed that these sections were too demanding for such candidates. The

panel also believed that Listening and Reading sections were too demanding for just-qualified candidates at the A2 level. Conversely, the panel believed that the Writing, Speaking, and Listening sections were not challenging enough to recommend cutscores for just-qualified candidates at the C2 level. Overall, these results suggest that the TOEFL iBT test is considered to discriminate at the B1 through C1 levels of the CEFR.

Final Evaluation of the TOEFL iBT Standard-Setting Process and Cutscores

At the conclusion of the standard setting for all four sections of the TOEFL iBT test, panelists were asked to complete an evaluation form. This form served the purpose of collecting information about aspects of procedural validity. Panelists were asked to rate the clarity with which various aspects of the study were presented, and were asked to indicate overall their level of comfort with the full set of recommended cutscores. Table 6 lists a series of prompts about the standard-setting process that panelists were asked to respond to, using a 4-point scale that ranged from *strongly agree* to *strongly disagree*. Except for one of the statements listed in Table 6, all 23 of the panelists indicated that they agreed or strongly agreed, with at least half the panel selecting *strongly agreed* for each prompt. One panelist selected *disagree* for the statement “I understood the purpose of the study.”

Table 6

Prompts on the Final TOEFL Evaluation Form

The homework assignment was useful preparation for the study.
I understood the purpose of the study.
The instructions and explanations provided by the facilitators were clear.
The training in the standard-setting methods was adequate to give me the information I needed to complete my assignment.
The explanation of how the recommended cutscore were computed was clear.
The opportunity for feedback and discussion between rounds was helpful.
The process of making the standard-setting judgments was easy to follow.

In addition, the evaluation form had four questions about information sources that might have influenced panelists’ cutscore decisions. They were asked to respond to each one using a 3-point scale: *very influential*, *somewhat influential*, and *not influential*. Table 7 summarizes the results of the TOEFL iBT panel.

Table 7***Level of Influence of Information Sources for TOEFL iBT Standard Setting***

Information source	Very influential	Somewhat influential	Not influential
The definition of the <i>just-qualified candidate</i>	87%	13%	0%
The between-round discussions	39%	61%	0%
The cutscores of other committee members	17%	57%	26%
My own professional experience	83%	17%	0%

Finally, panelists were shown the complete set of cutscore recommendations that they had made, across all four modalities and for each CEFR level. They were asked to indicate their comfort level with the set of cutscores, using the following scale: *very comfortable*, *somewhat comfortable*, *somewhat uncomfortable*, and *very uncomfortable*. The modal response was *very comfortable*, with all but two panelists selecting *very comfortable* or *somewhat comfortable*. Both of these panelists indicated that they did not believe that a just-qualified A1 candidate would have sufficient skills to score any points on the TOEFL iBT Speaking section.

The results from the evaluation form support the procedural validity of the standard-setting study for the TOEFL iBT test. Panelists reported that the training provided was clear and prepared them for the judgment task, and that the standard-setting process was easy to follow. The just-qualified candidate descriptions, as appropriate, were influential in the panelists' judgments, and the panelists were able to rely on their expertise (professional experience) to inform their judgments. The panelists also reported that they were comfortable with the final recommendations.

Panel 2 (the TOEIC Test)

The language experts who formed Panel 2 conducted the linking study for both the TOEIC test and the TOEIC *Bridge* test. This section describes the makeup of the panel and presents the results of their work. Appendices G and H present the results by panelist and round.

Panelists

Twenty-two experts representing 10 countries served on the panel that focused on mapping the *Test of English for International Communication*[™] (TOEIC) assessments onto the CEFR. (Five panelists had also served on the TOEFL iBT panel.) The experts were selected for

their experience with English-language instruction, learning, and testing in the workplace, and their familiarity with the Common European Framework of Reference. They were also selected to represent an array of European countries where the TOEIC test is used. Table 8 presents the demographic characteristics of the 22 panelists. Appendix F provides the panelists' affiliations.

Table 8

Panel 2 Demographics

		Number	Percent
Gender	Female	13	59%
	Male	9	41%
Selection Criteria ⁵	ESL teacher at language school (private or university)	17	
	Administrator of school/program where ELL classes are taught	4	
	Assessment expert or researcher	5	
	Educational consultant	5	
Country	Belgium	2	9%
	France	6	27%
	Germany	2	9%
	Greece	3	14%
	Hungary	3	14%
	Italy	1	5%
	Malta	1	5%
	Poland	2	9%
	Russia	1	5%
	Slovakia	1	5%

Standard-Setting Results for the TOEIC Test and the TOEIC Bridge Test

TOEIC results. The first-, second-, and third-round judgments for A2, B2, and C2 section-level judgments, along with the single round of judgments for A1, B1, and C1 are presented in a series of tables (Tables G1 through G4 in Appendix G). Each panelist's individual cutscores are presented for each round, as are the cross-panel summary statistics (mean, median, standard deviation, minimum, maximum, and standard error of judgment).

Table 9 presents the cross-panel mean judgments and the standard deviations by round for the Writing section for the TOEIC test. The second, third, and fourth columns present the results for the first three rounds where panelists focused on the A2, B2, and C2 cutscores. For the fourth round, the focus shifted to A1, B1, and C1, with panelists working within the constraints

of the just-established A2, B2, and C2 cutscores. Raw scores on the Writing section range from 0 to 26, and scaled scores from 0 to 200.

Table 9

TOEIC Writing Judgments

CEF levels	Round 1 mean (SD)	Round 2 mean (SD)	Round 3 mean (SD)	Round 4 mean (SD)	Final scaled score
A1	-	-	-	5.45 (1.00)	30
A2	10.00 (1.90)	10.32 (1.64)	10.00 (1.63)	-	70
B1	-	-	-	14.68 (1.32)	120
B2	17.86 (3.00)	19.10 (2.36)	19.10 (2.02)	-	150
C1	-	-	-	23.63 (1.16)	200
C2	24.08 (1.66)	N/A	N/A	-	N/A

As for the TOEFL iBT test, panelists had the option of assigning *not applicable* (N/A) if they felt that the section could not assess a particular CEFR level, either because it was too difficult for candidates at a low level to access the tasks, or because it did not provide sufficient challenge to distinguish performances from candidates at the highest CEFR levels. For a cutscore to be computed at any level, at least 67% of the group had to rate it as applicable. As Table 9 shows, while in the first round at least 67% of the panelists indicated a possible C2 cutscore, after the discussion this percentage dropped, and so no C2 cutscore was computed in Rounds 2 or 3. The C1 raw cutscore (24, rounded) corresponds to the maximum scaled score of 200, indicating that a near perfect raw score in Writing is needed to reach the C1 level. This issue was discussed by the panelists, who nonetheless collectively believed that such a score was needed to be considered performing at the C1 level for the TOEIC Writing section. The variability in panelists' judgments decreased across the rounds. The SEJs for the recommended cutscores for A1, A2, B1, B2, and C1 are .21, .35, .28, .43, and .25, respectively.

Table 10 presents the cross-panel mean judgments and the standard deviations by round for the Speaking section for the TOEIC test. Raw scores on the Speaking section range from 0 to 24, and scaled scores from 0 to 200.

Table 10***TOEIC Speaking Judgments***

CEF levels	Round 1 mean (SD)	Round 2 mean (SD)	Round 3 mean (SD)	Round 4 mean (SD)	Final scaled score
A1	-	-	-	7.37 (1.50)	50
A2	11.09 (1.87)	10.91 (1.66)	10.95 (1.50)	-	90
B1	-	-	-	15.45 (0.96)	120
B2	18.00 (1.83)	18.45 (1.60)	18.59 (1.59)	-	160
C1	-	-	-	22.78 (0.81)	200
C2	N/A	N/A	N/A	-	N/A

As can be seen in the table, an insufficient number of panelists deemed the Speaking section to be able to distinguish C2-level performance across the rounds; and the C1 raw cutscore (23, rounded) corresponds to the maximum scaled score of 200, indicating that a near perfect raw score in Speaking is needed to reach the C1 level. As with the TOEIC Writing section, the panel collectively believed that such a score was needed to be considered performing at the C1 level for the TOEIC Speaking section. The variability in panelists' judgments decreased across the rounds. The SEJs for the recommended cutscores for A1, A2, B1, B2, and C1 are .32, .32, .21, .34, and .17, respectively.

Table 11 presents the cross-panel mean judgments and the standard deviations by round for the Listening section for the TOEIC test. Raw scores on the Listening section range from 0 to 100, and scaled scores from 5 to 495. For the Listening section, panelists initially awarded a C2 cutscore (using the Angoff approach) for item-level judgments in Round 1, although three panelists had cutscores of 100, meaning that they expected a just-qualified C2-level candidate to answer every item correctly. As discussions ensued, an increasing number decided that the test section did not sufficiently stretch a candidate in order to be sure that someone was at a C2 rather than C1 level. The variability in panelists' judgments fluctuated somewhat for the A2 level but decreased for the B2 level. The SEJs for the recommended cutscores for A1, A2, B1, B2, and C1 are 1.36, 2.41, 1.73, 2.16, and 1.23, respectively.

Table 11***TOEIC Listening Judgments***

CEF levels	Round 1 mean (SD)	Round 2 mean (SD)	Round 3 mean (SD)	Round 4 mean (SD)	Final scaled score
A1	-	-	-	17.30 (6.37)	60
A2	20.14 (12.8)	24.14 (9.34)	30.59 (11.31)	-	110
B1	-	-	-	57.68 (8.13)	275
B2	77.86 (15.4)	75.50 (13.35)	78.91 (10.15)	-	400
C1	-	-	-	93.59 (5.79)	490
C2	97.47 (3.0)	N/A	N/A	-	N/A

Table 12 presents the cross-panel mean judgments and the standard deviations by round for the Reading section for the TOEIC test. Raw scores on the Reading section range from 0 to 100, and scaled scores from 5 to 495.

Table 12***TOEIC Reading Judgments***

CEF levels	Round 1 mean (SD)	Round 2 mean (SD)	Round 3 mean (SD)	Round 4 mean (SD)	Final scaled score
A1	-	-	-	17.67 (5.97)	60
A2	20.20 (10.47)	28.36 (8.37)	33.27 (9.30)	-	115
B1	-	-	-	61.71 (5.11)	275
B2	78.75 (11.28)	79.32 (9.35)	80.55 (7.82)	-	385
C1	-	-	-	N/A	N/A
C2	98.30 (1.47)	N/A	N/A	-	N/A

Similar to the Listening section, panelists initially awarded a Reading C2 cutscore (with two panelists indicating a maximum score of 100); but again, as discussions ensued, an increasing number decided that the test section did not assess C2-level reading ability. Panelists

also concluded that the section did not access C1-level reading ability. Also consistent with the Listening section, there was some fluctuation in variability at the A2 level but a decrease in variability at the B2 level. The SEJs for the recommended cutscores for A1, A2, B1, and B2 are 1.27, 1.98, 1.09, and 1.67, respectively.

In summary, for the TOEIC Writing and Speaking sections, the cutscore (A2, B2, and C2) means changed very little from Round 1 to Round 2 or from Round 2 to Round 3. For the Writing and Speaking sections, the variability (standard deviation) of the panelists' judgments tended to decrease from Round 1 to Round 2 and from Round 2 to Round 3, indicating a greater degree of panelist convergence; there was less of a consistent decrease for the Listening and Reading sections. Unlike what was observed for the Writing and Speaking sections, the mean judgments, particularly at the A2 level, tended to shift between rounds for the Listening and Reading sections. The panelists expressed that the variability in their A2 recommendations was a reflection of divergent interpretations of the A2 just-qualified descriptors as applied to the test questions. Significant conversation followed the presentation of both the Round 1 and Round 2 results, as the panelists clarified their understanding of the A2 level. The third- and fourth-round mean scores may be accepted as the panel-recommended cutscores; that is, they reflect the minimum scores necessary to qualify for the A1 through C2 levels on the CEFR. As can be seen in the previous tables for the TOEIC test, the language experts of Panel 2 believed that Writing, Speaking, Listening, and Reading sections were not challenging enough to recommend cutscores for just-qualified candidates at the C2 level. The panel held the same view for Reading at the C1 level. Overall, these results suggest that the TOEIC test is considered to discriminate at the A1 through B2 levels of the CEFR.

TOEIC Bridge test results. The first-, second-, and third-round judgments for A1, A2, and B1 section-level judgments are presented in a series of Tables H1 and H2 in Appendix H. Each panelist's individual cutscores are presented for each round, as are the cross-panel summary statistics (mean, median, standard deviation, minimum, maximum, and standard error of judgment).

Table 13 and Table 14 present the cross-panel mean judgments and the standard deviations by round for the Listening and Reading sections. The three rounds of judgments all focused on the A1, A2, and B1 cutscores. For both test sections the raw scores ranged from 0 to 50, and the scaled scores from 10 to 90.

Table 13***TOEIC Bridge Test Reading Judgments***

CEF levels	Round 1 mean (SD)	Round 2 mean (SD)	Round 3 mean (SD)	Final scaled score
A1	17.78 (5.15)	17.45 (3.95)	17.19 (3.27)	46
A2	35.08 (5.77)	34.36 (5.07)	33.14 (4.29)	70
B1	46.73 (3.05)	45.72 (3.27)	45.72 (3.23)	86

The mean judgments remained relatively stable across the rounds, and there was a general decrease in the variability of panelists' judgments. The SEJs for the recommended cutscores for A1, A2, and B1 are .71, .94, and .71, respectively.

Table 14***TOEIC Bridge Test Listening Judgments***

CEF levels	Round 1 mean (SD)	Round 2 mean (SD)	Round 3 mean (SD)	Final scaled score
A1	17.64 (7.42)	20.14 (3.81)	20.24 (2.64)	46
A2	35.10 (4.75)	35.76 (3.48)	35.29 (4.66)	64
B1	47.46 (2.19)	47.50 (1.59)	47.47 (1.46)	84

The pattern of results for Listening was similar to that for Reading. The mean judgments remained stable across the rounds, and there was a decrease in the variability of panelists' judgments. The SEJs for the recommended cutscores for A1, A2, and B1 are .58, 1.02, and .32, respectively.

Final Evaluation of TOEIC Test and TOEIC Bridge Test Cutscores

At the conclusion of the standard setting for both the TOEIC test and the TOEIC *Bridge* test, panelists were asked to complete evaluation forms. The forms served the purpose of collecting information about aspects of procedural validity. Panelists were asked to rate the

clarity with which various aspects of the study were presented, and were asked to indicate overall their level of comfort with the full set of recommended cutscores. Table 15 lists a series prompts about the TOEIC standard-setting process that panelists were asked to respond to, using a 4-point scale that ranged from *strongly agree* to *strongly disagree*. For each of the statements listed in Table 15, 100% of the 22 panelists indicated that they agreed or strongly agreed, with at least two-thirds of the group selecting *strongly agree* for each prompt.

Table 15

Prompts on the Final TOEIC Evaluation Form

The homework assignment was useful preparation for the study.
I understood the purpose of the study.
The instructions and explanations provided by the facilitators were clear.
The training in the standard-setting methods was adequate to give me the information I needed to complete my assignment.
The explanation of how the recommended cutscore were computed was clear.
The opportunity for feedback and discussion between rounds was helpful.
The process of making the standard-setting judgments was easy to follow.

In addition, the TOEIC evaluation form had four questions about information sources that might have influenced panelists' cutscore decisions. They were asked to respond to each one, using a 3-point scale: *very influential*, *somewhat influential*, and *not influential*. Table 16 summarizes the results for the TOEIC panel.

Table 16

Level of Influence of Information Sources for TOEIC Standard Setting

Information source	Very influential	Somewhat influential	Not influential
The definition of the <i>just-qualified candidate</i>	82%	18%	0%
The between-round discussions	59%	41%	0%
The cutscores of other committee members	14%	73%	14%
My own professional experience	82%	18%	0%

Finally, panelists were shown the complete set of cutscore recommendations that they had made, across all four modalities, and for each CEFR level. They were asked to indicate their comfort level with the set of cutscores, using the following scale: *very comfortable*, *somewhat comfortable*, *somewhat uncomfortable*, and *very uncomfortable*.

For the evaluation of the TOEIC cutscores, the modal response was *very comfortable*, with all but three panelists selecting *very comfortable* or *somewhat comfortable*. Comments from these three panelists indicated they had concerns about setting cutscores at the A1 and C1 levels for the TOEIC test, since they did not think the test was sensitive enough to distinguish performance at the extremes of the CEFR.

The results from the evaluation for the TOEIC test provide evidence supporting the procedural validity of the standard-setting process. Panelists reported that the training provided was clear and prepared them for the judgment task, and that the standard-setting process was easy to follow. The just-qualified candidate descriptions, as appropriate, were influential in the panelists' judgments, and the panelists were able to rely on their expertise (professional experience) to inform their judgments. The majority of panelists also reported that they were comfortable with the final recommendations.

When the panel completed the standard setting for the TOEIC *Bridge* test, they completed a second evaluation form. They were not asked to respond a second time to the prompts listed in Table 15, only to indicate the influence level of various information sources and their comfort level with the final set of TOEIC *Bridge* test cutscores. Table 17 summarizes the results.

Table 17

Level of Influence of Information Sources for the TOEIC Bridge Test Standard Setting

Information source	Very influential	Somewhat influential	Not influential
The definition of the <i>just-qualified candidate</i>	81%	19%	0%
The between-round discussions	48%	52%	0%
The cutscores of other committee members	33%	57%	10%
My own professional experience	57%	43%	0%

The results of Table 17 differ somewhat from those seen previously in Table 7 and Table 16. Although the definition of the just-qualified candidate still had the most influence, the panel as a whole was less assured of their own professional experience as an influence. And while the cutscores of other panelists still had the least influence, their level of influence was higher for this test. These results parallel what was heard in informal conversations with panelists, many of whom noted having much less experience with students more likely to take the TOEIC *Bridge* test than the TOEIC test. In terms of panelists' comfort level with the TOEIC *Bridge* test cutscores, the modal response was *very comfortable*, with all but one panelist selecting *very comfortable* or *somewhat comfortable*. The one differing panelist had concerns that the TOEIC *Bridge* test was not able to distinguish B1 performance from A2 performance. Overall, these results support the procedural validity of the TOEIC *Bridge* test standard setting.

Conclusions

The purpose of this study was to construct cutscores on three English-language proficiency tests corresponding to different levels of the Common European Framework of Reference. For two of the tests, the TOEFL iBT test and the TOEIC test, the target CEFR levels were A1 through C2; for the TOEIC *Bridge* test, the target levels were A1, A2, and B1.

The cutscores were constructed following well-established standard-setting procedures. A modified Angoff approach was used for the selected-response section of the tests (Reading and Listening), and a modified examinee selection approach was used for the constructed-response sections of the tests (Writing and Speaking), the later being applicable only to the TOEFL iBT and TOEIC tests. Two panels of language experts, one for the TOEFL iBT test and one for the TOEIC test and the TOEIC *Bridge* test, were assembled and participated in the standard-setting studies. Table 18, Table 19, and Table 20 summarize the scaled-score cutscores for the three tests.

Linkages were successfully established between each section of the TOEFL iBT test and levels B1, B2, and C1 of the CEFR, and between each section of the TOEIC test and levels A1 through C1 of the CEFR, with the exception of Reading at the C1 level. Both the Reading and Listening sections of the TOEIC *Bridge* test were successfully linked to all three of the targeted levels of the CEFR.

Table 18

Scaled-Score Cutscore Results for the TOEFL iBT Test

	Writing (max. 30 points)	Speaking (max. 30 points)	Listening (max. 30 points)	Reading (max. 30 points)
A1	-	8	-	-
A2	11	13	-	-
B1	17	19	13	8
B2	21	23	21	22
C1	28	28	26	28
C2	-	-	-	29

Table 19

Scaled-Score Cutscore Results for the TOEIC Test

	Writing (max. 200 points)	Speaking (max. 200 points)	Listening (max. 495 points)	Reading (max. 495 points)
A1	30	50	60	60
A2	70	90	110	115
B1	120	120	275	275
B2	150	160	400	385
C1	200	200	490	-
C2	-	-	-	-

Table 20

Scaled-Score Cutscore Results for the TOEIC Bridge Test

	Listening (max. 90 points)	Reading (max. 90 points)
A1	46	46
A2	70	64
B1	86	84

The difficulty of linking test scores to the CEFR should not be underestimated. The CEFR, according to Weir (2005), does not provide sufficient information about how contextual factors affect performance across the levels, or adequately delineate how language develops across the levels in terms of cognitive or meta-cognitive processing. This may lead to difficulties in interpreting differences across the CEFR levels. Some of this was evident during the panelist discussions of the CEFR when developing the just-qualified descriptions; panelists noted that the descriptive language of the CEFR was not consistently applied across the levels, making it more difficult for them to differentiate among the levels. The difficulty, however, also is a function of the tests. It is more likely that tests developed specifically to map to the CEFR would pose less of a linking challenge than tests relying only on a post hoc approach, as was the present case.

Although the tests considered in this study measured the basic communicative modalities, all covered by the CEFR, the items on the tests were not specifically developed to operationalize these modalities necessarily as depicted by the CEFR. Although this did not preclude setting cutscores for some of the levels, it most likely was the reason why not all intended CEFR levels were mapped. The value of using level descriptors to inform test development, thereby increasing alignment and the potential meaningfulness of cutscores, was recently noted by Bejar, Braun, and Tannenbaum (2007) in the context of No Child Left Behind testing.

Although not all targeted CEFR levels were mapped, there was positive evidence of procedural validity. The majority of panelists for each test reported that they were adequately trained and prepared to conduct their standard-setting judgments, and that the standard-setting process was easy to follow. Panelists reported that the definition of the just-qualified candidate most influenced their judgments and that they were able to use their professional experience to inform their judgments. Furthermore, the majority of panelists reported that they were comfortable with the recommended cutscores. Procedural validity is an important criterion against which to evaluate the quality of the standard-setting process (Hambleton & Pitoniak, 2006; Kane, 2001).

External validity evidence is also desirable and most often takes the form of convergence with other sources of information (Hambleton & Pitoniak, 2006; Kane, 2001). In the present case, for example, convergent evidence could be obtained from teacher ratings of their students' English-language proficiency in terms of the CEFR (Council of Europe, 2003). Although a convergence of evidence would lend further support of the reasonableness of the panel-based cutscores, the meaning of a divergence of evidence is less clear, given that there is no true cutscore. "Differences in results from two different procedures would not be an indication that one was right and the other wrong; even if two methods did produce the same or similar cut scores, we could only be sure of precision, not accuracy" (Cizek & Bunch, 2007, p. 63). With this in mind, the cutscores from this study should be considered recommendations only; they are not absolutes. Potential users of these cutscores are advised to consider their specific needs and circumstances, and other relevant information that may be germane to determinations of the English-language proficiency of their test takers that was not part of this set of studies. It is reasonable for users to adjust these recommended cutscores to better accommodate their needs.

This set of standard-setting studies, we believe, represents a significant step forward in the evolution of research concerned with linking test scores to the CEFR. The use of the modified examinee paper selection method for constructed-response items, which enabled panelists to consider profiles of responses; the inclusion of item-data partitioned by test-taker ability levels; the shift from item-level judgments in the first round for the selected-response items to a more holistic judgment for the subsequent rounds; and the slotting of the Level 1 cutscores in relation to the Level 2 cutscores all reflect innovative and creative design elements in research studies whose primary objective is relating test scores to the CEFR. Continued advances in this area of applied research would seem warranted, given the increasing emphasis (and hence importance) of being able to interpret the meaning of test scores in terms of the proficiency levels of the CEFR.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.
- Bejar, I. I., Braun, H. I., & Tannenbaum, R. J. (2007). A prospective, progressive, and predictive approach to standard setting. In R. W. Lissitz (Ed.), *Assessing and modeling cognitive development in school: Intellectual growth and standard setting* (pp. 1-30). Maple Grove, MN: Journal of Applied Metrics Press.
- Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education, 17*, 59–88.
- Cizek, G. J. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement, 30*, 93–106.
- Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: SAGE Publications.
- Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge, UK: Cambridge University Press.
- Council of Europe. (2003). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). Manual: Preliminary pilot version*. Strasbourg, France: Language Policy Division.
- Figueras, N., North, B., Takala, S., Verhelst, N., & Van Avermaet, P. (2005). Relating examinations to the Common European Framework: A manual. *Language Testing, 22*, 261–279.
- Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. (2000). Setting performance standards on complex educational assessments. *Applied Psychological Measurement, 24*, 355–366.
- Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–470). Westport, CT: Praeger Publishers.

- Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement*, 34, 353–366.
- Kane, M. (1994). Validating performance standards associated with passing scores. *Review of Educational Research*, 64, 425–461.
- Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 53–88). Mahwah, NJ: Lawrence Erlbaum.
- Mills, C. N., & Jaeger, R. M. (1998). Creating description of desired student achievement when setting performance standards. In L. N. Hansche (Ed.), *Meeting the requirements of Title I: Handbook for the development of performance standards* (pp. 73–85). Washington, DC: U.S. Department of Education.
- Tannenbaum, R. J., & Wylie, E. C. (2005). *Mapping English language proficiency test scores onto the Common European Framework* (ETS Research Rep. No. RR-05-18; TOEFL Research Rep. No. RR–80). Princeton, NJ: ETS.
- Weir, C. J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing*, 22, 281–300.
- Zieky, M. J. (2001). So much has changed: How the setting of cutscores has evolved since the 1980s. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 19–52). Mahwah, NJ: Lawrence Erlbaum.

Notes

- ¹ The Council of Europe pilot manual for relating examinations to the CEFR says that samples of writing and speaking, standardized to the CEFR, may be available. We were not able to locate the samples or information about how the samples were standardized to the CEFR.
- ² Some members met more than one criterion, so percentages are not reported.
- ³ The standard error of judgment may be considered only a guideline, as it assumes independence of judgments and that panelists were randomly selected from all possible panelists. The former is true only for the first round of judgments, and the second assumption is not likely to be true.
- ⁴ The A2 raw cutscore was too low to scale.
- ⁵ Some members met more than one criterion, so percentages are not reported.
- ⁶ Panelist 17 absent for Day 3 of the TOEFL session.
- ⁷ Panelist 4 had to leave the meeting before the final judgments were made for A1, B1, and C1.
- ⁸ Panelist 6 missed a section of the questions during the Round 1 review and so had no estimated cutscore for the first round.

List of Appendixes

	Page
A – Panel 1 and Panel 2 Homework Tasks.....	36
B – Panel 1 and 2 Indicator Summaries of Language Skills Defined by the CEFR	46
C – Panel 1 and Panel 2 Agendas.....	55
D – Panelists’ Affiliations for Panel 1	64
E – Panelists’ Judgments for the TOEFL iBT Test.....	65
F – Panelists’ Affiliations for Panel 2.....	69
G – Panelists’ Judgments for TOEIC	70
H – Panelists’ Judgments for the TOEIC <i>Bridge</i> Test.....	74

Appendix A

Panel 1 and Panel 2 Homework Tasks

Study to Map the TOEFL iBT Test Onto the Common European Framework

The role of the Common European Framework of Reference (CEFR) is to foster mutual understanding across countries for users and language testers by providing a common language to describe the stages of language learning. ETS is seeking to benchmark several of its English-language proficiency tests onto this framework, using an expert-judgment standard-setting approach. At the study you will be familiarized with the TOEFL iBT test, receive training in the standard-setting process, and have an opportunity to practice making judgments.

During the study itself, the discussions will focus around all six levels of the CEFR. In order to facilitate discussions, it is very important that you become familiar with these levels. A PDF version of the framework can be found at the following address:

http://www.coe.int/t/dg4/linguistic/CADRE_EN.asp

The TOEFL iBT test addresses the four modalities of Speaking, Listening, Reading, and Writing, and we will be discussing the characteristics of a candidate who has just enough (the minimum) English-language skills to be considered performing at the A1, A2, B1, B2, C1, and C2 levels. This will be done for each of the four modalities. In other words, this candidate is the least able C2 performer in Speaking, Listening, Reading, and Writing; the least able C1 performer in each of the four modalities; the least able B2 in these modalities; and so on for B1, A2, and A1. This candidate is not the average performer or the highest performer in a level for a modality; this is the candidate who barely has the English-language skills to be classified at each of the six CEFR levels. You may think of it this way: If you lined up candidates within each of the six CEFR levels for a modality by their ability (lowest to highest) for, say, Speaking, the candidate with just enough English-speaking skills would be the very first candidate in the line.

In the section below, relevant tables from the CEFR have been identified by page number and title. Please review these CEFR tables. Highlight key words or phrases that help you to understand how the CEFR levels are defined.

Speaking: Pages 58–59: Overall Oral Production and Sustained Monologue (both tables).
Pages 74–77: Overall Spoken Interaction, Understanding a Native-Speaker Interlocutor, Conversation, Informal Discussion (with friends)

Writing: Pages 61–62: Overall Written Production, Creative Writing, Reports and Essays. Page 83: Overall Written Interaction, Correspondence.

Listening: Pages 66–68: Overall Listening Comprehension, Understanding Conversation Between Native Speakers, Listening as a Member of a Live Audience, Listening to Announcements and Instructions, Listening to Audio Media and Recordings. Page 75: Understanding a Native-Speaker Interlocutor.

Reading: Pages 69–71: Overall Reading Comprehension, Reading Correspondence, Reading for Orientation, Reading for Information and Argument, Reading Instructions

On the following sheets, at the top of the table, there is a global descriptor of levels of the CEFR. These were taken from Table 1 (p. 24), Common Reference Levels: Global Scale. Having reviewed the relevant CEFR tables, complete the attached sheets by briefly noting in your own words, in the space provided, the key characteristics or indicators from the CEFR tables (above) that describe an English-language learner (candidate) who has just enough skills to be performing at each of the CEFR levels. This is the least able C2 performer in Speaking, the least able C1 performer in Speaking, the least able B2 performer in Speaking, and so on through A1. Please complete this activity for all four modalities. You do not need to write very much. As you complete this activity, ask yourself: What can the least able C2 speaker, for example, do that the highest performing C1 speaker cannot do? What can the least able C1 speaker do that the highest performing B2 speaker cannot do? and so on.

Please bring your completed sheets to the October meeting. Your notes, along with those of your colleagues, will form the starting point for discussion during the study itself.

Key Characteristics by Language Modality of an English-Language Learner With Just Enough Skills to Be Performing at the Specified CEFR Level. This Is the Least Able Candidate in That CEFR Level.

C2 global descriptor: Can understand with ease virtually everything heard or read. Can summarise information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations

C1 global descriptor: Can understand a wide range of demanding, longer texts, and recognize implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organizational patterns, connectors and cohesive devices.

Speaking

Writing

Reading

Listening

Key Characteristics by Language Modality of an English-Language Learner With Just Enough Skills to Be Performing at the Specified CEFR Level. This Is the Least Able Candidate in That CEFR Level.

B2 global descriptor: Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussion in his/her field of specialization. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.

B1 global descriptor: Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst traveling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans.

Speaking

Writing

Reading

Listening

Key Characteristics by Language Modality of an English-Language Learner With Just Enough Skills to Be Performing at the Specified CEFR Level. This Is the Least Able Candidate in That CEFR Level.

A2 global descriptor: Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g., very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment, and matters in areas of immediate need.

A1 global descriptor: Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows, and things he/she has. Can interact in a simple way, provided the other person talks slowly and clearly and is prepared to help.

Speaking

Writing

Reading

Listening

Study to Map the *Test of English for International Communication*[™] (TOEIC) Assessment and the TOEIC *Bridge* Test Onto the Common European Framework

The role of the Common European Framework of Reference (CEFR) is to foster mutual understanding across countries for users and language testers by providing a common language to describe the stages of language learning. ETS is seeking to benchmark several of its English-language proficiency tests onto this framework, using an expert-judgment standard-setting approach. At the study you will be familiarized with the tests, receive training in the standard-setting process, and have an opportunity to practice making judgments.

During the study itself, the discussions will focus around all six levels of the CEFR. In order to facilitate discussions, it is very important that you become familiar with these levels. A PDF version of the framework can be found at the following address:

http://www.coe.int/t/dg4/linguistic/CADRE_EN.asp.

The [new] TOEIC test addresses Listening and Reading but now also includes modules in Speaking and Writing. We will be discussing the characteristics of a candidate who has just enough (the minimum) English-language skills to be considered performing at the A1, A2, B1, B2, C1, and C2 levels. This will be done for each of the four modalities. In other words, this candidate is the least able C2 performer in Speaking, Listening, Reading, and Writing; the least able C1 performer in each of the four modalities; the least able B2 in these modalities; and so on for B1, A2, and A1. This candidate is not the average performer or the highest performer in a level for a modality; this is the candidate who barely has the English-language skills to be classified at each of the six CEFR levels. You may think of it this way: If you lined up candidates within each of the six CEFR levels for a modality by their ability (lowest to highest) for, say, Speaking, the candidate with just enough English-speaking skills would be the very first candidate in the line.

The TOEIC *Bridge* test addresses Listening and Reading, and we will be focusing on A1, A2, and B1 levels of the CEFR. (The assignment below just needs to be completed once.)

In the section below, relevant tables from the CEFR have been identified by page number and title. Please review these CEFR tables. Highlight key words or phrases that help you to understand how the CEFR levels are defined.

Speaking. Pages 58–59: Overall Oral Production and Sustained Monologue (both tables).
Pages 74–77: Overall Spoken Interaction, Understanding a Native-Speaker Interlocutor,
Conversation, Informal Discussion (with friends)

Writing. Pages 61–62: Overall Written Production, Creative Writing, Reports and Essays.
Page 83: Overall Written Interaction, Correspondence

Listening. Pages 66–68: Overall Listening Comprehension, Understanding Conversation
Between Native Speakers, Listening as a Member of a Live Audience, Listening to
Announcements and Instructions, Listening to Audio Media and Recordings. Page 75:
Understanding a Native-Speaker Interlocutor

Reading. Pages 69–71: Overall Reading Comprehension, Reading Correspondence,
Reading for Orientation, Reading for Information and Argument, Reading Instructions

On the following three sheets, at the top of the table there is a global descriptor of two levels of the CEFR. These were taken from Table 1 (p. 24), Common Reference Levels: Global Scale. Having reviewed the relevant CEFR tables, complete the attached sheets by briefly noting in your own words, in the space provided, the key characteristics or indicators from the CEFR tables that describe an English-language learner (candidate) who has just enough skills to be performing at each of the CEFR levels. This is the least able C2 performer in Speaking, the least able C1 performer in Speaking, the least able B2 performer in Speaking, and so on through A1. Please complete this activity for all four modalities. You do not need to write very much. As you complete this activity, ask yourself: What can the least able C2 speaker , for example, do that the highest performing C1 speaker cannot do; What can the least able C1 speaker do that the highest performing B2 speaker cannot do, and so on.

Please bring your completed sheets to the October meeting. Your notes, along with those of your colleagues, will form the starting point for discussion during the study itself.

Key Characteristics by Language Modality of an English-Language Learner With Just Enough Skills to Be Performing at the Specified CEFR Level. This Is the Least Able Candidate in That CEFR Level.

C2 global descriptor: Can understand with ease virtually everything heard or read. Can summarise information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations

C1 global descriptor: Can understand a wide range of demanding, longer texts, and recognize implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organizational patterns, connectors and cohesive devices.

Speaking

Writing

Reading

Listening

Key Characteristics by Language Modality of an English-Language Learner With Just Enough Skills to Be Performing at the Specified CEFR Level. This Is the Least Able Candidate in That CEFR Level.

B2 global descriptor: Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussion in his/her field of specialization. Can interact with a degree of fluency and spontaneity that make regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topic issue giving the advantages and disadvantages of various options.

B1 global descriptor: Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst traveling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans.

Speaking

Writing

Reading

Listening

Key Characteristics by Language Modality of an English-Language Learner With Just Enough Skills to Be Performing at the Specified CEFR Level. This Is the Least Able Candidate in That CEFR Level.

A2 global descriptor: Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g., very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment, and matters in areas of immediate need.

A1 global descriptor: Can understand and use familiar everyday expression and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows, and things he/she has. Can interact in a simple way, provided the other person talks slowly and clearly and is prepared to help.

Speaking

Writing

Reading

Listening

Appendix B

Panel 1 and 2 Indicator Summaries of Language Skills Defined by the CEFR

Table B1

Panel 1 Indicators of CEFR Definitions of Proficiency in Writing

Writing skills of just-qualified A1
Can write simple, isolated phrases; may be incohesive. Can fill forms (not tax forms): name, address, simple phrases. Can express information on the here and now (immediacy). Can write on personal topics (“all about me”) with structured guidance.
Writing skills of just-qualified A2
Can write simple sentences linked with and, but, because. Can use basic word order. Can use high-frequency words. Can write about contexts in familiar situations. Can express self with limited cohesion. Can use intelligible spelling, sentence structure, syntax (in simple contexts).
Writing skills of just-qualified B1
Can create straightforward, connected, linear text. Can write generally intelligible text. Can write on wider scope of topics, familiar subjects in field of interest ... abstract feelings, emotions, and notions Can inform, describe, give gist.
Writing skills of just-qualified B2
Can write clear, detailed text. Can write in broader context (general interest), variety of topics in field of interest. Can synthesize arguments in known field of interest. Can provide supporting points. May make grammatical mistakes but these don't lead to misunderstanding.
Writing skills of just-qualified C1
Can take audience into account; is aware of audience. Can write to underlying salient issues. Can express with precision and accuracy plus high degree of grammatical accuracy. Can write with natural or personal style. Can write on complex subjects with subthemes plus evaluations. Can use broad lexical repertoire. Can write well-structured, complex text.
Writing skills of just-qualified C2
Has mastered a variety of writing styles, registers, and tones and can use them appropriately. Can use nuance and idiomatic language for stylistic effect and to demonstrate cultural awareness. Considers readers; emphasizes main points. Shows proficient use of language and meta-language.

Table B2

Panel 1 Indicators of CEFR Definitions of Proficiency in Speaking

Speaking skills of just-qualified A1
Simple, isolated statements/phrases restricted to very familiar/personal topics. Needs assistance from experienced, sympathetic interlocutor (struggles along with speaker to sustain interaction). Pronunciation has limited repertoire and strong interference from other language(s).
Speaking skills of just-qualified A2
Hesitation in speaking. Can handle short exchanges but cannot keep conversation going. Can give simple (series of) phrases about likes, dislikes, family, work. Can make simple requests for clarification. Intelligible to the listener with effort from listener. Can deliver basic rehearsed presentations (short) on familiar topics.
Speaking skills of just-qualified B1
Some fluency (linear sequence). Can cope with everyday situations. Can briefly give reasons and explanations. Can describe and briefly explain graphics/tables in fields of interest; with preparation. Can express self on familiar abstract thoughts, feelings, notions. Can maintain one-on-one/face-to-face conversation but may need assistance.
Speaking skills of just-qualified B2
Can give clear, detailed descriptions and prepared presentations attuned to the listener. Can develop clear arguments with relevant support and examples on wide range of topics related to fields of interest. Can sustain conversation with degree of fluency and spontaneity. Takes listener and cultural context into account. Monologue causes no undue stress to listener.
Speaking skills of just-qualified C1
No strain on listener. Expresses self fluently and spontaneously, <u>almost effortlessly</u> . Uses idiomatic speech. Uses precise and accurate grammar. Can vary intonation and place stress correctly. Can describe or present complex subjects (appropriately structured). Shows flexible/effective use of language (humor).
Speaking skills of just-qualified C2
Effective and flexible communication with audience. Can easily follow and contribute to complex discussion with all speakers. Can express fine shades of meaning. Can discuss abstract topics beyond own field. Uses multiple registers appropriately. Clear, well-constructed, smoothly flowing arguments. Demonstrates full confidence in speaking.

Table B3

Panel 1 Indicators of CEFR Definitions of Proficiency in Listening

Listening skills of just-qualified
Can understand very slow speech with familiar words and basic phrases on here and now. Can understand short and slow speech with pauses and repetition. Requires sympathetic speaker.
Listening skills of just-qualified A2
Can understand short, clearly, slowly, and directly articulated concrete speech on simple, everyday, familiar topics/matter. Can understand formulaic language (basic language and expressions). Can understand short directions, instructions, descriptions. Can extract relevant, important information from recorded messages.
Listening skills of just-qualified B1
Can understand main points. Can understand clear, standard speech on familiar matters and short narratives when presented relatively slowly Will sometimes need repetition and clarification in conversation. Can follow broadcast information carefully delivered. (Example: BBC World but not SkyNews) Can deduce sentence meaning.
Listening skills of just-qualified B2
Can understand standard speech on most topics. Can use macro-structural clues to check for overall understanding. Can grasp the main points of academic lectures. Can understand radio and television. Can understand speech from native speakers directed at him/her most of the time. Can understand extended speech and complex arguments; requires explicit markers. With some effort can catch most native-speaker discussion. Can understand standard dialect delivered at normal speed.
Listening skills of just-qualified C1
Can understand extended speech on abstract unfamiliar topics (e.g., lectures). Can understand <u>enough</u> but may need clarification. Can follow most speakers. Unfamiliar accents can cause difficulties in comprehension. Does not require explicit markers. Can recognize a wide range of idiomatic speech. Can listen between the lines; can infer implied meaning.
Listening skills of just-qualified C2
Has no difficulty understanding any kind of standard spoken language, even when delivered at fast native speed. Will need time to adjust to nonstandard or colloquial speech.

Table B4

Panel 1 Indicators of CEFR Definitions of Proficiency in Writing

Reading skills of just-qualified A1
Recognizes familiar names and words with visual or contextual support. Understands very short, simple texts, one phrase at a time. Needs time to re-read.
Reading skills of just-qualified A2
Can find specific information in simple, everyday material (e.g., advertising, brochures, menus, notices, directions, instructions, timetables, newspapers). Can understand simple and predictable material (e.g., job-related or private written communication). Can understand <u>short, simple texts</u> containing most commonly used vocabulary. Grasps the main point in text with predictable information or contexts, and/or texts with high-frequency vocabulary. Can infer at the vocabulary level.
Reading skills of just-qualified B1
Reads straightforward, factual text in field of interest. Reads personal letters. Reads material containing <u>some degree of abstraction</u> . Finds relevant information in everyday material. Can infer at sentence level.
Reading skills of just-qualified B2
Can read with a large degree of independence. Can read texts in a wide range of professional topics (may need dictionary). Has a broad, active vocabulary but has difficulty with low-frequency idioms. Understands articles written from a stance (opinions and attitudes). Can scan complex texts, locating relevant details. Shows inferencing ability at macro level (text level.)
Reading skills of just-qualified C1
Needs to re-read; more effort required than C2 for complex, extended text in all fields of interest. Identifies or infers opinion, intention, feelings of writer.
Reading skills of just-qualified C2
Reads practically all types of texts and styles, from most formal to highly colloquial. Can critically interpret both explicit and implicit meaning.

Table B5

Panel 2 Indicators of CEFR Definitions of Proficiency in Writing

Writing skills of just-qualified A1

Can produce simple, isolated phrases/sentences in familiar/personal, very concrete areas.
Writing is exclusively model-based
Writing is very formulaic, repetitive.

Writing skills of just-qualified A2

Can write short, simple sentences on concrete, familiar topics/events.
Can follow basic punctuation rules.
Can use simple connectors (*and, but, because*).
Can describe self and others.
Makes basic mistakes systematically.
Writing samples are characterized by intelligible spelling and basic punctuation.
Sentences are occasionally linked.

Writing skills of just-qualified B1

Can briefly give reasons/opinions.
Can produce straightforward text
Can produce narrative with greater use of logical connectors but still simple sentences
Can use wider range of text models
Can narrate simple story, conventional, factual, routine, linear
Writes on familiar, everyday topics.
Writing is awkward and shows interference from other language(s).

Writing skills of just-qualified B2

Can produce a clear, detailed text essay.
Can argue for/against a position.
Can describe advantages/disadvantages.
Variety of subjects related to field of interest.
Easy to follow the structure but cohesion may be lost at times.
Texts are based on standard patterns.
Writing achieves clear, effective communication.
Can synthesize.
Uses informal/formal register.
Writing includes vocabulary related to field and good terminology.
Can write compound and complex sentences that will not lead to misunderstanding and do not impede meaning.
Adapts standard format to personal needs.

Writing skills of just-qualified C1

Produces longer, well-structured and well-developed texts.
Uses language flexibly; mostly accurate
Elaborates to some degree.
Writes on complex subjects, with some degree of effort (time, dictionary, aids)
Can distinguish between formal and informal.
Uses efficient style (less wordy).

(Table continues)

Table B5 (continued)

Writing skills of just-qualified C2
<p>Produces clear, smoothly flowing, complex texts in an appropriate and effective style. Writing is more natural/spontaneous. Can write about all subjects. Writing includes finer shades of meaning and frequently includes idiomatic expressions. Produces smoothly flowing sentences/paragraphs; complex, extended texts. Writing is characterized by range-appropriate style/register. Uses cultural reference (e.g., politeness). Takes reader's needs into account. Can write complex, extended text. Maintains consistent, highly grammatical control of complex language. Makes few errors, if any.</p>

Table B6

Panel 2 Indicators of CEFR Definitions of Proficiency in Speaking

Speaking skills of just-qualified
<p>Provides simple, isolated phrases and sentences on very concrete, familiar topics/immediate needs. Must repeat, speak very slowly. Uses very limited vocabulary. Needs <u>very</u> sympathetic listener.</p>
Speaking skills of just-qualified A2
<p>Can present simple, short, rehearsed material/information on personal/familiar tasks. Can present short, simple, routine tasks Uses simple, fixed/formulaic phrases. Needs/asks for repetition. Listener makes effort. Speaks slowly (and needs sympathetic interlocutor). Talks about immediate needs. Intonation and stress not natural. Can understand what is said clearly, slowly, and directly; relies on sympathetic interlocutor (one speaker). Can participate in a simple and direct exchange of information.</p>
Speaking skills of just-qualified B1
<p>Can provide straightforward description in area of interest. <u>Briefly</u> provides arguments, reasons, and support of opinions. Enters conversation on familiar topics in standard language. <u>Begins</u> to sustain comprehensible speech but use circumlocutions—stays within area of interest. Speaks about everyday events, dreams, hopes, ambitions. Uses wide range of simple language.</p>

(Table continues)

Table B6 (continued)

Speaking skills of just-qualified B2
Can speak with a <u>degree</u> of fluency, spontaneity with native speaker
Can speak about familiar contexts, wider range in field of interest
Can sustain views by providing relevant explanations and arguments, discussions.
Strain on either party is minimal.
Speech is characterized by noticeably long pauses/hesitations when searching for patterns and expressions. Employs a limited number of cohesive devices. There is some jumpiness in longer contributions.
Errors do not impede message.
Uses some complex forms.
Uses minimal idiomatic language.
Uses good range of vocabulary in field of interest.
Produces/adapts to register of listener (formal/informal).
Can adjust to changes in discourse.
Shows good command of grammatical control.
Can present clear and detailed descriptions on a wide range of subjects related to field of interest.

Speaking skills of just-qualified C1
Can present almost effortlessly and can self-correct on wide range of discourse.
Can speak about complex, abstract topics.
Uses effective, precise, flexible language at length.
Can sustain one-to-many interaction.
Confirmation required when unfamiliar accent.
Can fill in gaps.
Can joke.

Speaking skills of just-qualified C2
Can present clear, standard, smoothly flowing descriptions or arguments on any subject/topic.
Can produce effortless speech using idiomatic expressions and colloquialisms, and can handle connotations.
Can interact without difficulty/hesitation and constraint with any native speaker, with appropriate register.
No awkward searching for words; uses wide range of cohesive devices and connectors.

Table B7

Panel 2 Indicators of CEFR Definitions of Proficiency in Listening

Listening skills of just-qualified A1

Needs very slow, carefully articulated speech with pauses and repetitions and helpful, sympathetic speaker.
Can understand basic, simple, formulaic phrases/words on self, family, and immediate surroundings in concrete context.

Listening skills of just-qualified A2

As long as speech production is short, simple, slow, and clear:
Can understand simple phrases and expressions that are related the most immediate needs.
Can generally catch the main point while listening to native speakers.
Can understand simple directions, instructions, and everyday conversations/exchanges related to field of interest.
Can understand slow, carefully articulated speech when given time to assimilate standard language/familiar variety on concrete topics.
Can derive meaning if accompanied by extra-linguistic/paralinguistic clues.

Listening skills of just-qualified B1

Understands main points in standard speech on familiar, regularly encountered, straightforward topics, simple technical information.
Can understand speech that is articulated relatively slowly or delivered at a relatively normal pace and with clarity.
May require some repetition.
Can guess some unknown words from context.

Listening skills of just-qualified B2

Can follow extended speech and lectures, provided the topic is reasonably familiar with clear signposts.
Can understand radio and recorded material in standard dialect at normal speed.
Can sometimes identify speaker mood and tone in obvious situations.
Can handle noisy environments.
Can understand main points on complex or abstract speech.
Can understand detail of everyday, concrete topics when talking to a native speaker.

Listening skills of just-qualified C1

Can follow relatively unstructured speech on complex (concrete and abstract) topics.
Can understand television with relative ease.
Can follow some slang and idioms.
Can distinguish between registers.
Can follow one-to-many conversations.
Can follow most lectures, etc., with relative ease.

Listening skills of just-qualified C2

Can understand any kind of natural, quickly spoken language on any topic (live or broadcast).
Can follow specific lectures/presentations.
Can summarize information from many spoken sources.
Can follow/cope with lectures with high degree of colloquialism, regional usage/variants, and unfamiliar terminology.
Can follow complex interaction face-to-face/third party,
Can recognize any kind of register and cultural references, styles, finer shades of meaning, inferences, connotations.
Can adjust to non-standard language (allowing time).

Table B8

Panel 2 Indicators of CEFR Definitions of Proficiency in Reading

Reading skills of just-qualified A1
Can read single phrases/isolated words at a time. Can read very short text with visual support (e.g., simple, written instructions, postcards). Can read familiar, basic words. Can read text related to personal, concrete experiences. Grasps basic idea of text (if short, simple, etc.). Not able to infer correctly meaning of unknown words.
Reading skills of just-qualified A2
As long as it is short, simply written in common, everyday language on concrete/personal topics or related to field of interest: Can find specific, predictable information in lists, signs, notices, instructions, menus, Can read and understand short personal letters, Can extract key information; can derive probable meaning of unknown words, Can follow specific, predictable information in simple, everyday material (e.g., tickets, calendar), Can identify main topic; unfamiliar text (especially when accompanied by visual support, logical structure), Derives probable meaning of unknown words, and Needs to reread.
Reading skills of just-qualified B1
Can read <u>straightforward</u> , factual texts/instructions on familiar topics/field of interest. Can find and understand information in everyday material (letters, brochures, and short official documents). Can recognize significant points, events, feelings, and wishes in personal or everyday texts that are clearly structured and signposted. Can deduce/extrapolate meaning of occasional unknown words in familiar context.
Reading skills of just-qualified B2
Can readily/easily understand broad range of texts; long, complex texts in field of interest. Can read with a high degree of independence. Can understand key points and detail from long texts in field of interest. Can discuss attitudes, viewpoints if clearly stated. Demonstrates developing inference skills. Can <u>recognize</u> registers/styles (formal/informal). Can understand and use references /sources. Can adjust speed of reading to task/purpose. Can understand some colloquial language; broad, active vocabulary but difficulty with low-frequency idioms. May need to <u>reread</u> difficult parts/sections. Can use different reading techniques and strategies.
Reading skills of just-qualified C1
Can read complex and demanding texts outside field of interest. May need to reread difficult sections with occasional use of dictionary Can identify fine points of detail, attitudes, and opinion (implied and stated) Can understand wide variety of idiomatic and colloquial expressions almost effortlessly
Reading skills of just-qualified C2
Can read with ease virtually all forms of written material, including abstract, technical, and literary works Can critically interpret and appreciate style and implicit meaning

Appendix C
Panel 1 and Panel 2 Agendas

**AGENDA: Mapping TOEFL iBT Test Onto the
Common European Framework**

Berlin

October 10, 2006

Day 1: TOEFL iBT Writing Section

8:30 – 9:00	Breakfast
9:00 – 9:30	Introductions/Welcome
9:30 – 10:00	Overview of ETS TOEFL iBT test, the CEF, and the purpose of the study
10:00 – 11:00	Table Groups: Define candidate focal groups for A2, B2, & C2 for Writing
11:00 – 11:15	Break
11:15 – 12:15	Room review of charts and creation of A1, B1, and C1 descriptions
12:15 – 13:15	Lunch
13:15 – 13:45	Introduction to TOEFL Writing section and rubrics
13:45 – 14:15	Overview of constructed-response standard-setting method
14:15 – 15:00	Individual review of essay exemplars and Round 1 judgments
15:00 – 15:15	Break (data entry)
15:15 – 15:45	Discussion of Round 1 results and score information
15:45 – 16:00	Round 2 individual judgments (A2, B2, C2)
16:00 – 16:15	Break (data entry)
16:15 – 16:45	Discussion of Round 2 results and impact data
16:45 – 17:00	Round 3 individual judgments (A2, B2, C2)
17:00 – 17:15	Break (data entry)
17:15 – 17:45	Discussion of slotting A1, B1, and C1 relative to final judgments for A2, B2, and C2
17:45 – 17:55	Final individual judgments for A1, B1, and C1
17:55 – 18:00	Wrap-up and adjourn

**AGENDA: Mapping TOEFL iBT Test Onto the
Common European Framework**

Berlin

October 11, 2006

Day 2: TOEFL iBT Speaking Section

8:30 – 9:00	Breakfast
9:00 – 9:15	Recap of process
9:15 – 10:15	Table Groups: Define candidate focal groups for A2, B2, and C2 for Speaking
10:15 – 11:15	Room review of charts and creation of A1, B1, and C1 descriptions
11:15 – 11:30	Break
11:30 – 12:00	Introduction to TOEFL Speaking section and rubrics
12:00 – 13:15	Individual review of speaking exemplars and Round 1 judgments
13:15 – 14:15	Lunch (data entry)
14:15 – 14:45	Discussion of Round 1 results and score information
14:45 – 15:00	Round 2 individual judgments (A2, B2, C2)
15:00 – 15:15	Break (data entry)
15:15 – 15:45	Discussion of Round 2 results and impact data
15:45 – 16:00	Round 3 individual judgments (A2, B2, C2)
16:00 – 16:15	Break (data entry)
16:15 – 16:45	Discussion of slotting A1, B1, and C1 relative to final judgments for A2, B2, and C2
16:45 – 16:55	Final individual judgments for A1, B1, and C1
16:55 – 17:00	Wrap-up and adjourn

**AGENDA: Mapping TOEFL iBT Test Onto the
Common European Framework**

Berlin

October 12, 2006

Day 3: TOEFL iBT Listening Section

8:30 – 9:00	Breakfast
9:00 – 9:45	Table Groups: Define candidate focal groups for A2, B2, and C2 for Listening
9:45 – 10:45	Room review of charts and creation of A1, B1, and C1 descriptions
10:45 – 11:00	Overview of selected-response standard-setting method
11:00 – 11:15	Break
11:15 – 12:00	Train/practice standard-setting approach
12:00 – 13:00	Individual Round 1 judgments (A2, B2, C2) on Listening items
13:00 – 14:00	Lunch (data entry)
14:00 – 14:45	Discussion of Round 1 results and score information
14:45 – 15:00	Round 2 individual judgments (A2, B2, C2)
15:00 – 15:15	Break (data entry)
15:15 – 15:45	Discussion of Round 2 results and impact data
15:45 – 16:00	Round 3 individual judgments (A2, B2, C2)
16:00 – 16:15	Break (data entry)
16:15 – 16:45	Discussion of slotting A1, B1, and C1 relative to final judgments for A2, B2, and C2
16:45 – 16:55	Final individual judgments for A1, B1, and C1
16:55 – 17:00	Wrap-up and adjourn

**AGENDA: Mapping TOEFL iBT Test Onto the
Common European Framework**

Berlin

October 13, 2006

Day 4: TOEFL iBT Reading Section

8:30 – 9:00	Breakfast
9:00 – 9:45	Table Groups: Define candidate focal groups for A2, B2, and C2 for Reading
9:45 – 10:45	Room review of charts and creation of A1, B1, and C1 descriptions
10:45 – 11:15	Train/practice standard-setting approach
11:15 – 11:30	Break
11:30 – 13:00	Individual Round 1 judgments (A2, B2, C2) on Reading items
13:00 – 14:00	Lunch (data entry)
14:00 – 14:45	Discussion of Round 1 results and score information
14:45 – 15:00	Round 2 individual judgments (A2, B2, C2)
15:00 – 15:15	Break (data entry)
15:15 – 15:45	Discussion of Round 2 results and impact data
15:45 – 16:00	Round 3 individual judgments (A2, B2, C2)
16:00 – 16:15	Break (data entry)
16:15 – 16:45	Discussion of slotting A1, B1, and C1 relative to final judgments for A2, B2, and C2
16:45 – 16:55	Final individual judgments for A1, B1, and C1
16:55 – 17:25	Final review of level scores for TOEFL assessment and evaluation forms
17:25 – 17:30	Wrap-up and adjourn

**AGENDA: Mapping TOEIC and TOEIC *Bridge* Test Onto the
Common European Framework**

Berlin

October 16, 2006

Day 1: TOEIC Writing Section

8:30 – 9:00	Breakfast
9:00 – 9:30	Introductions/Welcome
9:30 – 10:00	Overview of ETS’s new TOEIC test, the CEF, and the purpose of the study
10:00 – 11:00	Table Groups: Define candidate focal groups for A2, B2, and C2 for Writing
11:00 – 11:15	Break
11:15 – 12:15	Room review of charts and creation of A1, B1, and C1 descriptions
12:15 – 13:15	Lunch
13:15 – 13:45	Introduction to TOEIC Writing section and rubrics
13:45 – 14:15	Overview of constructed-response standard-setting method
14:15 – 15:00	Individual review of essay exemplars and Round 1 judgments
15:00 – 15:15	Break (data entry)
15:15 – 15:45	Discussion of Round 1 results and score information
15:45 – 16:00	Round 2 individual judgments (A2, B2, C2)
16:00 – 16:15	Break (data entry)
16:15 – 16:45	Discussion of Round 2 results and impact data
16:45 – 17:00	Round 3 individual judgments (A2, B2, C2)
17:00 – 17:15	Break (data entry)
17:15 – 17:45	Discussion of slotting A1, B1, and C1 relative to final judgments for A2, B2, and C2
17:45 – 17:55	Final individual judgments for A1, B1, and C1
17:55 – 18:00	Wrap-up and adjourn

**AGENDA: Mapping TOEIC and TOEIC *Bridge* Test Onto the
Common European Framework**

Berlin

October 17, 2006

Day 2: TOEIC Speaking Section

8:30 – 9:00	Breakfast
9:00 – 9:15	Recap of process
9:15 – 10:15	Table Groups: Define candidate focal groups for A2, B2, and C2 for Speaking
10:15 – 11:15	Room review of charts and creation of A1, B1, and C1 descriptions
11:15 – 11:30	Break
11:30 – 12:00	Introduction to TOEIC Speaking section and rubrics
12:00 – 13:15	Individual review of speaking exemplars and Round 1 judgments
13:15 – 14:15	Lunch (data entry)
14:15 – 14:45	Discussion of Round 1 results and score information
14:45 – 15:00	Round 2 individual judgments (A2, B2, C2)
15:00 – 15:15	Break (data entry)
15:15 – 15:45	Discussion of Round 2 results and impact data
15:45 – 16:00	Round 3 individual judgments (A2, B2, C2)
16:00 – 16:15	Break (data entry)
16:15 – 16:45	Discussion of slotting A1, B1, and C1 relative to final judgments for A2, B2, and C2
16:45 – 16:55	Final individual judgments for A1, B1, and C1
16:55 – 17:00	Wrap-up and adjourn

**AGENDA: Mapping TOEIC and TOEIC *Bridge* Test Onto the
Common European Framework**

Berlin

October 18, 2006

Day 3: TOEIC Listening Section

8:30 – 9:00	Breakfast
9:00 – 10:00	Table Groups: Define candidate focal groups for A2, B2, and C2 for Listening
10:00– 11:00	Room review of charts and creation of A1, B1, and C1 descriptions
11:00 – 11:15	Break
11:15 – 11:45	Overview of selected-response standard-setting method
11:45 – 12:30	Train/practice standard-setting approach
12:30 – 13:30	Lunch (data entry)
13:30 – 15:00	Individual Round 1 judgments (A2, B2, C2) on Listening items
15:00 – 15:30	Break (data entry)
15:30 – 16:15	Discussion of Round 1 results and score information
16:15 – 16:30	Round 2 individual judgments (A2, B2, C2)
16:30 – 16:45	Break (data entry)
16:45 – 17:15	Discussion of Round 2 results and impact data
17:15 – 17:25	Round 3 individual judgments (A2, B2, C2)
17:25 – 17:30	Wrap-up and adjourn

**AGENDA: Mapping TOEIC and TOEIC *Bridge* Test Onto the
Common European Framework**

Berlin

October 19, 2006

Day 4: TOEIC Reading Section

8:30 – 9:00	Breakfast
9:00 – 9:30	Discussion of slotting A1, B1, and C1 relative to final judgments for A2, B2, and C2
9:30 – 9:45	Final individual judgments for A1, B1, and C1
9:45 – 10:45	Table Groups: Define candidate focal groups for A2, B2, and C2 for Reading
10:45 – 11:00	Break
11:00 – 12:00	Room review of charts and creation of A1, B1, and C1 descriptions
12:00 – 12:30	Train/practice standard-setting approach
12:30 – 13:30	Lunch
13:30 – 14:30	Individual Round 1 judgments (A2, B2, C2) on Reading items
14:30 – 15:00	Break (data entry)
15:00 – 15:30	Discussion of Round 1 results and score information
15:30 – 15:40	Round 2 individual judgments (A2, B2, C2)
15:40 – 15:50	Break (data entry)
15:50 – 16:15	Discussion of Round 2 results and impact data
16:15 – 16:30	Round 3 individual judgments (A2, B2, C2)
16:30 – 16:45	Break (data entry)
16:45 – 15:15	Discussion of slotting A1, B1, and C1 relative to final judgments for A2, B2, and C2
15:15 – 15:25	Final individual judgments for A1, B1, and C1
15:25 – 15:30	Wrap-up and adjourn

**AGENDA: Mapping TOEIC and TOEIC *Bridge* Test Onto the
Common European Framework**

Berlin

October 20, 2006

Day 5: TOEIC Bridge Test Reading and Listening Sections

8:30 – 9:00	Breakfast
9:00 – 9:15	Overview of the TOEIC <i>Bridge</i> test
9:15 – 9:30	Review descriptions of focal groups for A1, A2, and B1 for Reading
9:30 – 10:30	Individual Round 1 judgments (A1, A2, B1) on Reading items
10:30 – 10:45	Break (data entry)
10:45 – 11:15	Discussion of Round 1 results and score information
11:15 – 11:45	Round 2 individual judgments (A1, A2, B1)
11:45 – 13:00	Lunch (data entry)
13:00 – 13:30	Discussion of Round 2 results and impact data
13:30 – 14:00	Round 3 individual judgments (A1, A2, B1)
14:00 – 14:15	Break
14:15 – 14:45	Review descriptions of focal groups for A1, A2, and B1 for Listening
14:45 – 15:30	Individual Round 1 judgments (A1, A2, B1) on Listening items
15:30 – 15:45	Break (data entry)
15:45 – 16:00	Discussion of Round 1 results and score information
16:00 – 16:10	Round 2 individual judgments (A1, A2, B1)
16:10– 16:20	Break (data entry)
16:20 – 16:40	Discussion of Round 2 results and impact data
16:40– 16:45	Round 3 individual judgments (A1, A2, B1)
16:45 – 17:00	Break
17:00 – 17:25	Final review of level scores for both the TOEIC test and the TOEIC <i>Bridge</i> test and evaluation forms
17:25 – 17:30	Wrap-up and adjourn

Appendix D
Panelists' Affiliations for Panel 1

Name	Affiliation
Bart Deygers	Ghent University
Tom Van Hout	Ghent University
Lut Baten	K.U. Leuven, Belgium
Mary Vigier	School of Management Clermont-Ferrand, France
Ekaterini Nikolarea	University of the Aegean
Felicitas Macgilchrist	European University Viadrina, Frankfurt
Jung Matthias	Institut für Internationale Kommunikation Duesseldorf
Mary Petersen	Logik Sprachtraining/Logik Academic Support
Orsolya Fulop	University College, Szekesfehervar, hungary
Martin Musumeci	University of Malta
Margreet de Hoop-Scherpenisse	Wagenmaker University, The Netherlands
Ingrid E. C. de Beer	James Boswell Institute, Utrecht University
Svein Magne Sirnes	Norsk Lektorlag
Slawomir Maskiewicz	Warsaw University, Poland
Konstantin Dibrova	St. Petersburg State University
Jana Beresova	Tnava University
Anna McCabe	St Louis University, Madrid
Christine Raisanen	Chalmers University of Technology, Göteborg, Sweden
Hakan Guven	Bilkent University, Ankara, Turkey
Carole Thomas	Bilkent University, Ankara, Turkey
Michael Fields	Higher Colleges of Technology, Abu Dhabi, UAE
Diane Schmitt	Nottingham Trent University, England
Spiros Papageorgiou	Lancaster University, England

Appendix E

Panelists' Judgments for the TOEFL iBT Test

Table E1

Judgments for the Writing Section of the TOEFL iBT Test

	Round 1			Round 2			Round 3 (final)			Round 4 (final)		
	A2	B2	C2	A2	B2	C2	A2	B2	C2	A1	B1	C1
P1	6	8	-	6	8	-	4	7	10	0	5.5	9.5
P2	4	8	10	4	8	10	3.5	7.5	9.5	0	5.5	9
P3	3	6	9	3	6	9	3	6.5	9.5		5	9
P4	4	8	10	4	8	-	4	8	-		5.5	9.5
P5	3	6	9	3	6	9	3	6.5	9	0	5.5	8.5
P6	1	6	9	2	6	9	2	6	9	0	5	9
P7	0	5	10	0.5	7	10	0.5	7	10	0	4	8
P8	4	8	-	4	7	-	4	7	-	0	5	9
P9	1.5	5	7.5	2.5	6	-	3	6	-	1	5	9
P10	3.5	6	8	3	6	9	3	6	-	0	5	8.5
P11	3.5	6.5	10	3.5	6.5	10	3.5	6.5	10	0	5	8.5
P12	0.5	7	-	0	7	-	0.5	6.5	-	0	5.5	9
P13	3	6	9	3	7	-	3	7	-	2	5.5	9
P14	3	6	9	3	6	9	3	6.5	9	1	5.5	8.5
P15	4	6.5	10	3.5	6.5	10	3.5	6.5	-	1	5	9
P16	4.5	7	10	4.5	6.5	-	4	6.5	-	0	5	9
P17	0	5	10	0	5	-	1	5	-	0	5.5	10
P18	4.5	6.5	8.5	4	6	8.5	4	6	8.5		5	8.5
P19	1	7.5	-	1	7.5	10	1	7.5	-	0	5.5	9
P20	3.5	5.5	9	3.5	6	9	2.5	5.5	9	1	5	9
P21	2	4	9.5	2	6	9	3	6	9	1.5	5	9
P22	1	6	8	-	6	8	-	6	8	0	5.5	10
P23	2.5	6	9.5	2.5	6	9.5	2	6	10	0	5	9.5
Mean	2.74	6.33	9.21	2.84	6.52	-	2.77	6.48	-	-	5.17	9.00
Median	3	6	9	3	6	-	3	6.5	-	-	5	9
Minimum	0.00	4.00	7.50	0.00	5.00	-	0.50	5.00	-	-	4.00	8.00
Maximum	6.00	8.00	10.00	6.00	8.00	-	4.00	8.00	-	-	5.50	10.00
SD	1.60	1.08	0.79	1.48	0.79	-	1.14	0.68	-	-	0.36	0.48
SEJ	0.33	0.23	0.16	0.31	0.16	-	0.24	0.14	-	-	0.07	0.10

Table E2***Judgments for the Speaking Section of the TOEFL iBT Test***

	Round 1			Round 2			Round 3 (final)			Round 4 (final)		
	A2	B2	C2	A2	B2	C2	A2	B2	C2	A1	B1	C1
P1	11	17	-	11	17	-	11	17	-	-	16	22
P2	11	19	-	10	18	-	10	18	-	6	16	22
P3	11	18	-	10	18	-	10	18	-	6	15	22
P4	14	19	-	13	18	-	13	18	-	-	16	22
P5	6	15	21	8	17	24	8	17	24	5	14	20
P6	6	15	22	6	18	24	6	18	24	6	14	21
P7	10	23	-	11	22	-	11	22	-	6	16	23
P8	10	18	-	10	18	-	10	18	-	6	14	22
P9	10	19	-	10	18	-	10	18	-	6	14	22
P10	12	17	23	14	17	24	12	17	-	6	14	21
P11	10	18	24	10	18	-	10	18	-	5	14	23
P12	6	18	-	8	18	-	-	18	-	-	14	-
P13	10	18	22	10	18	-	10	18	-	7	14	23
P14	14	-	-	10	18	-	10	18	-	6	14	22
P15	12	21	-	10	20	-	10	20	-	6	15	23
P16	8	17	-	9	17	-	-	17	-	-	15	22
P17	8	18	-	8	17	-	8	17	-	4	14	23
P18	10	14	-	10	14	-	10	14	-	6	15	21
P19	6	19	-	10	19	-	10	19	-	-	14	22
P20	10	18	22	10	17	-	10	18	-	5	14	21
P21	6	14	22	10	18	23	10	18	24	6	14	22
P22	-	15	24	-	15	24	-	15	24	-	14	22
P23	9	16	-	9	16	-	10	18	-	-	14	22
Mean	9.55	17.55		9.86	17.65		9.95	17.78		5.75	14.52	21.95
Median	14	18		10	18		10	18		6	14	22
Minimum	6	14		6	14		6	14		4	14	20
Maximum	14	23		14	22		13	22		7	16	23
SD	2.46	2.18		1.64	1.56		1.43	1.51		0.68	0.79	0.79
SEJ	0.51	0.45		0.34	0.32		0.30	0.31		0.14	0.16	0.16

Table E3***Judgments for the Listening Section of the TOEFL iBT Test***

	Round 1			Round 2			Round 3 (final)			Round 4 (final)		
	A2	B2	C2	A2	B2	C2	A2	B2	C2	A1	B1	C1
P1	6.95	22.40	33.35	7	25	34	10	28	34		20	32
P2	1.95	23.70	33.20	3	25	34	3	26	34		16	31
P3	1.15	23.80	33.65	3	25	-	4	27	-		17	31
P4	1.60	28.85	33.90	4	27	-	4	29	-		17	32
P5	2.95	29.00	33.15	4	26	33	4	27	33		15	30
P6	1.55	27.80	33.75	2	28	34	2	26	34		15	31
P7	1.60	32.10	33.60	2	30	34	2	30	34		17	29
P8	1.55	28.30	33.95	2	27	34	2	27	34		16	30
P9	9.70	23.70	32.55	8	24	33	3	29	-		15	30
P10	1.05	20.85	32.40	2	22	32	4	28	-		19	30
P11	3.25	20.60	33.40	3	25	-	3	29	-		18	33
P12	0.35	27.05	33.90	0	27	-	0	30	-		15	-
P13	0.60	22.50	31.15	1	23	31	2	25	33		17	31
P14	6.55	30.55	33.50	4	26	34	4	26	-		16	31
P15	4.35	29.10	33.60	4	29	33	4	29	-		20	33
P16	2.55	22.20	34.00	3	25	-	3	28	-		16	31
P17 ⁶	-	-	-	-	-	-	-	-	-		-	-
P18	2.00	22.20	31.95	2	22	32	2	22	32		16	31
P19	2.45	30.65	34.00	9	29	-	9	29	-		19	31
P20	0.85	19.70	33.30	-	22	33	3	22	33		17	30
P21	0.85	24.30	33.65	2	23	33	2	23	33		16	30
P22	0.40	19.85	34.00	0	20	34	0	20	34		16	31
P23	1.70	23.10	32.95	2	21	32	2	21	32		15	32
Mean	2.54	25.10	33.31	3.19	25.05	33.13	3.27	26.41			16.73	30.95
Median	1.65	23.75	33.55	3	25	33	3	27			16	31
Minimum	0.35	19.70	31.15	0.00	20.00	31.00	0.00	20.00			15.00	29.00
Maximum	9.70	32.10	34.00	9.00	30.00	34.00	10.00	30.00			20.00	33.00
SD	2.38	3.86	0.73	2.34	2.73	0.96	2.33	3.02			1.58	1.02
SEJ	0.51	0.82	0.16	0.50	0.58	0.20	0.50	0.64			0.34	0.22

Table E4***Judgments for the Reading Section of the TOEFL iBT Test***

	Round 1			Round 2			Round 3 (final)			Round 4 (final)		
	A2	B2	C2	A2	B2	C2	A2	B2	C2	A1	B1	C1
P1	0.85	17.15	44.10	-	22	-	-	30	-		12	43
P2	2.10	25.05	39.40	-	25	40	-	25	40		10	39
P3	0.45	26.30	43.45	-	26	43	-	26	43		9	39
P4	0.55	28.15	41.35	2	29	43					-	-
P5	2.00	35.25	44.00	-	25	44	-	25	44		10	37
P6	0.10	28.60	44.20	0	25	44	0	27	44		12	41
P7	1.65	30.90	41.20	2	33	43	2	30	43		12	34
P8	1.20	32.95	44.35	1	33	-	1	30	-		15	40
P9	4.70	28.10	41.45	5	30	42	7	29	43		19	42
P10	1.35	28.60	42.70	3	29	41	3	32	41		14	43
P11	0.10	24.85	42.85	0	30	45	0	32	45		16	40
P12	0.00	30.90	44.45	-	31	-	-	35	-		13	42
P13	0.40	34.30	42.25	0	30	43	0	29	-		15	38
P14	12.15	42.60	44.40	2	32	45	2	29	45		10	43
P15	2.25	34.20	42.10	2	28	42	2	28	42		10	41
P16	0.45	36.65	45.00	0	30	-	0	35	-		17	38
P17	0.10	36.00	44.90	0	30	-	0	36	-		17	43
P18	0.15	20.50	39.45	-	21	39	-	21	39		18	39
P19	9.70	38.35	45.00	-	37	-	-	36	-		19	43
P20	1.75	27.90	43.65	2	25	42	2	28	44		16	36
P21	1.30	27.70	43.50	2	27	43	1	27	43		11	37
P22	0.00	29.85	43.50	0	30	44	0	30	44		13	41
P23	0.50	17.85	41.60	1	22	42	-	25	42		11	39
Mean	1.90	29.68	42.99		28.26	42.65		29.32	42.80		13.59	39.91
Median	0.85	28.6	43.5		29	43		29	43		13	40
Minimum	0.00	17.15	39.40		21.00	39.00		21.00	39.00		9.00	34.00
Maximum	12.15	42.60	45.00		37.00	45.00		36.00	45.00		19.00	43.00
SD	3.06	6.28	1.65		3.95	1.62		3.91	1.74		3.17	2.56
SEJ	0.64	1.31	0.34		0.82	0.34		0.81	0.36		0.68	0.55

Appendix F
Panelists' Affiliations for Panel 2

Name	Affiliation
Kristien Van Hoegaerden	Group T Leuven University College
Jean Louis-Sauvage	University of Mons-Hainaut
Jean-Francois Jaouen	French Navy
Gretta Lachaise	ESIGELEE (Ecole Supererieuse d'Dngenieurs Générélistes)
Carl Storz	Institut National des Télécommunications
Abdi Kazeroni	Université de Technologie de Compiègne, Compiègne, France
Dawn Hallidy	University of Le Havre, France
Anne O'Mahoney	École Supérieure de Commerce de Toulouse, France
Charalambos Kollias	Hellenic American Union and Hellenic American University
Ekaterini Nikolarea	University of the Aegean
Melina Papaconstantinou	Technological Educational Institute of Kavala, Greece
Sue Luther	Georg-Simon-Ohm University of Applied Sciences, Nürnberg
Jung Matthias	<i>Institut für Internationale Kommunikation Duesseldorf</i>
Csaba Haidu	M-Prospect Nyelviskola, Hungry
Eva Lukacsi	Budapest School of Communication
Lucia Katona	Centre for Foreign Languages, Lorand Eötvös University, Budapest
Brunella Belluomini	Language Data Bank, Italy
Martin Musumeci	University of Malta
Zofia Prele	Agricultural University, Wroclaw, Poland
Jadwiga Bolechowska	Agricultural University of Wroclaw
Konstantin Dibrova	St. Petersburg State University
Jana Beresova	Tnava University

Appendix G
Panelists' Judgments for TOEIC

Table G1

Judgments for the Writing Section of the TOEIC Test

	Round 1			Round 2			Round 3 (final)			Round 4 (final)		
	A2	B2	C2	A2	B2	C2	A2	B2	C2	A1	B1	C1
P1	10	18	-	10	20	-	10	20	-	6	16	25
P2	7	22	26	8	22	-	7	22	-	6	14	24
P3	8	15	23	10	18	-	10	18	-	-	15	23
P4	9	20	-	9	20	-	9	20	-	-	15	-
P5	10	20	25	10	20	26	10	20	26	4	15	24
P6	10	20	-	10	20	-	10	20	-	3	12	25
P7	11	22	-	11	22	-	11	21	-	6	16	25
P8	10	13	26	13	16	-	13	16	-	6	15	22
P9	10	22	-	10	22	-	10	22	-	3	13	23
P10	11	21	23	11	21	-	11	21	-	6	16	-
P11	11	18	25	11	18	-	10	18	-	6	14	25
P12	13	21	-	13	21	-	13	21	-	5	16	25
P13	12	19	-	12	18	-	12	18	-	6	16	24
P14	11	14	21	12	15	22	12	15	22	5	14	23
P15	14	17	25	12	18	25	10	17	25	6	16	22
P16	11	16	25	11	21	-	9	20	-	5	16	25
P17	10	17	-	10	20	-	10	20	-	6	15	24
P18	7	-	-	7	-	-	7	-	-	6	14	-
P19	7	14	24	9	14	24	8	16	25	6	12	22
P20	11	16	21	11	18	23	10	19	24	6	13	22
P21	10	17	24	10	21	-	10	20	-	6	16	23
P22	7	13	25	7	16	25	8	17	25	6	14	23
Mean	10.00	17.86	24.08	10.32	19.10		10.00	19.10		5.45	14.68	23.63
Median	10	18	25	10	20		10	20		6	15	24
Minimum	7.00	13.00	21.00	7.00	14.00		7.00	15.00		3.00	12.00	22.00
Maximum	14.00	22.00	26.00	13.00	22.00		13.00	22.00		6.00	16.00	25.00
SD	1.90	3.00	1.66	1.64	2.36		1.63	2.02		1.00	1.32	1.16
SEJ	0.41	0.64	0.35	0.35	0.50		0.35	0.43		0.21	0.28	0.25

Table G2***Judgments for the Speaking Section of the TOEIC Test***

	Round 1			Round 2			Round 3 (final)			Round 4 (final)		
	A2	B2	C2	A2	B2	C2	A2	B2	C2	A1	B1	C1
P1	9	18	-	10	19	-	10	19	-	8	16	24
P2	11	22	-	11	22	-	11	22	-	7	17	24
P3	12	19	-	11	19	-	11	19	-	-	16	-
P4	6	18	-	6	18	-	6	18	-	-	14	-
P5	12	19	23	12	20	24	12	20	24	4	15	23
P6	12	17	20	12	17	20	12	20	24	6	15	22
P7	12	19	-	12	20	-	12	20	-	8	16	24
P8	11	15	24	11	15	24	11	15	24	8	15	22
P9	11	15	22	11	18	-	11	18	-	8	15	23
P10	11	17	23	11	18	23	11	18	-	8	15	22
P11	12	19	24	12	19	-	12	19	-	9	15	23
P12	13	19	-	13	19	-	13	19	-	10	16	23
P13	11	18	23	11	18	-	11	18	-	7	14	23
P14	13	17	22	12	17	-	11	17	-	8	16	22
P15	12	17	21	12	18	22	12	18	22	9	15	22
P16	12	17	24	12	18	24	12	18	-	-	16	-
P17	13	19	-	12	19	-	12	19	-	9	16	22
P18	11	22	-	10	22	-	10	22	-	6	18	-
P19	11	18	24	8	17	24	11	17	-	6	16	22
P20	12	18	24	12	18	-	11	18	-	7	15	24
P21	11	18	24	11	18	24	11	18	-	7	15	23
P22	6	15	24	8	17	24	8	17	24	5	14	22
Mean	11.09	18.00		10.91	18.45		10.95	18.59		7.37	15.45	22.78
Median	11.5	18		11	18		11	18		8	15	23
Minimum	6.00	15.00		6.00	15.00		6.00	15.00		4.00	14.00	22.00
Maximum	13.00	22.00		13.00	22.00		13.00	22.00		10.00	18.00	24.00
SD	1.87	1.83		1.66	1.60		1.50	1.59		1.50	0.96	0.81
SEJ	0.40	0.39		0.35	0.34		0.32	0.34		0.32	0.21	0.17

Table G3***Judgments for the Listening Section of the TOEIC Test***

	Round 1			Round 2			Round 3 (final)			Round 4 (final)		
	A2	B2	C2	A2	B2	C2	A2	B2	C2	A1	B1	C1
P1	21.40	96.60	100	25	93	-	40	93	-	20	65	-
P2	18.50	94.75	99.90	19	95	-	19	95	-	15	75	100
P3	22.25	94.60	99.65	22	85	-	35	85	-	-	59	-
P4	33.10	77.60	100	35	80	-	40	83	-	-	60	-
P5	30.20	80.20	97.65	30	80	100	30	80	100	20	55	100
P6	6.85	79.65	97.95	30	80	95	30	80	100	20	55	95
P7	35.60	96.30	99.55	40	85	-	45	85	-	30	65	100
P8	11.20	67.40	97.10	20	67	-	30	75	-	26	60	95
P9	42.90	87.90	99.95	38	72	-	38	90	-	22	55	-
P10	12.30	61.30	96.70	12	61	94	30	70	94	20	45	85
P11	26.45	54.70	93.55	26	55	90	35	60	90	20	45	80
P12	4.65	49.00	90.30	5	49	90	10	60	95	5	45	85
P13	6.70	43.60	89.55	12	44	97	20	60	95	15	55	95
P14	3.05	77.30	97.95	12	78	97	18	76	96	10	55	90
P15	22.15	78.80	94.55	20	78	97	25	75	100	10	55	90
P16	28.35	73.35	98.35	25	75	-	30	80	-	16	56	99
P17	13.15	88.40	97.45	15	80	-	30	80	-	15	60	95
P18	3.85	94.20	100	20	94	-	5	94	-	4	78	-
P19	6.30	92.05	99.45	30	80	-	35	75	-	20	55	95
P20	27.55	70.95	96.95	35	75	95	35	75	-	20	60	95
P21	25.25	74.35	98.40	30	77	-	55	85	-	18	56	97
P22	42.50	79.95	99.35	30	78	99	38	80	-	20	55	95
Mean	20.19	77.86	97.47	24.14	75.50		30.59	78.91		17.30	57.68	93.59
Median	21.78	79.23	98.15	25	78		30	80		20	55.5	95
Minimum	3.05	43.60	89.55	5	44		5	60		4	45	80
Maximum	42.90	96.60	100	40	95		55	95		30	78	100
SD	12.47	15.37	3.00	9.34	13.35		11.31	10.15		6.37	8.13	5.79
SEJ	2.66	3.28	0.64	1.99	2.85		2.41	2.16		1.36	1.73	1.23

Table G4***Judgments for the Reading Section of the TOEIC Test***

	Round 1			Round 2			Round 3 (final)			Round 4 (final)		
	A2	B2	C2	A2	B2	C2	A2	B2	C2	A1	B1	C1
P1	27.60	92.05	99.70	35	92	-	45	92	-	-	65	-
P2	2.45	88.95	99.55	20	89	-	28	89	-	10	70	-
P3	16.45	86.10	98.00	30	85	-	38	88	-	-	61	-
P4	28.25	73.15	99.50	30	85	-	40	90	-	-	65	-
P5	31.00	89.70	98.00	31	90	100	30	88	100	10	65	95
P6 ⁸	-	-	-	40	80	95	40	80	95	20	70	94
P7	27.45	95.50	100	40	90	-	42	90	-	25	65	-
P8	16.30	68.60	98.35	30	80	-	30	80	-	20	-	-
P9	22.55	83.35	99.50	28	82	-	34	80	-	-	60	-
P10	17.70	65.20	97.35	35	65	97	35	65	97	25	60	90
P11	44.90	65.90	96.95	40	60	97	35	70	-	25	55	90
P12	12.70	72.65	97.20	20	80	97	20	80	97	10	60	90
P13	4.40	54.70	94.10	15	60	95	20	70	95	20	55	97
P14	9.00	83.20	98.45	16	80	97	20	80	97	10	60	90
P15	21.10	81.25	96.40	21	81	96	30	80	97	20	70	90
P16	27.65	82.35	98.30	36	85	-	40	85	-	20	62	-
P17	19.40	84.40	98.15	25	80	98	30	80	-	20	60	90
P18	5.45	88.70	99.65	11	89	-	11	89	-	5	50	-
P19	11.10	92.60	99.65	25	75	-	35	75	-	20	58	100
P20	24.40	71.05	96.85	30	70	-	40	70	-	20	65	-
P21	23.10	68.15	98.60	35	77	-	40	80	-	20	60	-
P22	31.25	66.30	100	31	70	-	49	71	-	18	60	-
Mean	20.20	78.75	98.30	28.36	79.32		33.27	80.55		17.67	61.71	
Median	21.10	82.35	98.35	30	80		35	80		20	60	
Minimum	2.45	54.70	94.10	11	60		11	65		5	50	
Maximum	44.90	95.50	100	40	92		49	92		25	70	
SD	10.47	11.28	1.47	8.37	9.35		9.30	7.82		5.97	5.11	
SEJ	2.23	2.40	0.31	1.78	1.99		1.98	1.67		1.27	1.09	

Appendix H

Panelists' Judgments for the TOEIC *Bridge* Test

Table H1

Judgments for the Reading Section of the TOEIC Bridge Test

	Round 1			Round 2			Round 3 (final)		
	A1	A2	B1	A1	A2	B1	A1	A2	B1
P1	17.70	30.80	45.80	18	31	46	18	31	46
P2	27.90	35.20	48.60	23	35	49	23	35	49
P3	12.15	33.10	46.55	18	35	-	18	35	-
P4	17.75	37.60	46.90	18	38	-	18	38	-
P5	11.40	34.55	46.95	11	35	47	15	35	50
P6	22.30	35.95	48.95	20	36	47	17	35	47
P7	20.20	42.80	49.75	21	43	-	-	-	-
P8	10.80	40.55	48.80	11	41	-	15	38	-
P9	18.20	40.15	49.30	20	38	48	18	37	46
P10	12.10	27.20	48.75	20	35	50	20	35	50
P11	19.30	31.10	46.85	19	31	47	17	30	44
P12	16.05	34.10	45.05	16	30	42	17	32	44
P13	19.65	32.65	43.85	18	33	44	18	30	44
P14	21.40	42.75	49.20	19	39	48	18	34	47
P15	18.50	36.35	45.00	19	36	45	17	34	46
P16	22.55	39.75	47.75	18	35	45	18	34	45
P17	20.70	37.35	46.50	20	37	45	20	35	45
P18	4.70	18.10	35.10	5	18	35	5	18	35
P19	16.20	36.45	45.95	16	35	46	17	34	46
P20	20.10	36.30	46.35	20	35	47	17	34	46
P21	24.20	41.10	48.65	17	32	45	17	34	46
P22	17.40	27.80	47.35	17	28	47	18	28	47
Mean	17.78	35.08	46.73	17.45	34.36	45.72	17.19	33.14	45.72
Median	18.35	36.13	46.93	18	35	46.5	18	34	46
Minimum	4.70	18.10	35.10	5	18	35	5	18	35
Maximum	27.90	42.80	49.75	23	43	50	23	38	50
SD	5.15	5.77	3.05	3.95	5.07	3.27	3.27	4.29	3.23
SEJ	1.10	1.23	0.65	0.84	1.08	0.70	0.71	0.94	0.71

Table H2***Judgments for the Listening Section of the TOEIC Bridge Test***

	Round 1			Round 2			Round 3 (final)		
	A1	A2	B1	A1	A2	B1	A1	A2	B1
P1	22.90	36.70	49.50	23	37	48	23	36	48
P2	22.15	36.45	50.00	22	36	50	22	36	50
P3	15.70	38.40	49.45	20	38	-	20	38	-
P4	23.25	40.00	49.65	25	43	-	25	43	-
P5	15.70	41.20	48.35	16	41	48	16	40	50
P6	18.45	34.75	49.60	15	35	50	22	35	47
P7 ^a	-	-	-						
P8	9.35	39.55	47.75	20	36	48	20	36	48
P9	13.70	41.40	49.45	20	40	-	20	38	-
P10	13.50	28.90	48.40	20	29	48	20	29	48
P11	18.90	31.20	45.80	19	34	46	20	35	46
P12	31.60	39.90	46.50	25	35	47	22	35	47
P13	28.05	38.35	47.50	28	38	48	22	38	48
P14	33.00	42.35	48.80	25	38	48	23	37	48
P15	12.00	30.95	43.60	15	35	47	20	36	47
P16	9.95	30.30	46.95	17	34	47	17	34	47
P17	14.40	29.80	46.35	17	33	45	18	33	46
P18	12.55	28.90	43.85	13	29	44	13	19	44
P19	4.70	31.25	42.70	21	38	47	20	38	47
P20	18.30	35.95	46.55	20	35	-	20	36	47
P21	10.20	28.45	46.00	20	35	49	20	36	49
P22	22.15	32.25	49.85	22	32	-	22	33	-
Mean	17.64	35.10	47.46	20.14	35.76	47.50	20.24	35.29	47.47
Median	15.70	35.95	47.75	20	35	48	20	36	47
Minimum	4.70	28.45	42.70	13	29	44	13	19	44
Maximum	33.00	42.35	50.00	28	43	50	25	43	50
SD	7.42	4.75	2.19	3.81	3.48	1.59	2.64	4.66	1.46
SEJ	1.62	1.04	0.48	0.83	0.76	0.35	0.58	1.02	0.32

^a Panelist 7 had to leave after the TOEIC *Bridge* test Reading judgments were made and did not participate in the TOEIC *Bridge* test Listening judgments.



Test of English as a Foreign Language
PO Box 6155
Princeton, NJ 08541-6155
USA

To obtain more information about TOEFL programs and services, use one of the following:

Phone: 1-877-863-3546
(US, US Territories*, and Canada)

1-609-771-7100
(all other locations)

E-mail: toefl@ets.org
Web site: www.ets.org/toefl

*America Samoa, Guam, Puerto Rico, and US Virgin Islands