




**TOEFL**

ISSN 1552-8219

# Research Reports

*RR - 80*

*November 2005*



Mapping English Language  
Proficiency Test Scores  
Onto the Common  
European Framework

Richard J. Tannenbaum

E. Caroline Wylie

**Mapping English Language Proficiency Test Scores  
Onto the Common European Framework**

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

RR-05-18



*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 © 2005 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.

EDUCATIONAL TESTING SERVICE, ETS, the ETS logos, Graduate Record Examinations, GRE, TOEFL, the TOEFL logo, TOEIC, TSE, and TWE are registered trademarks of Educational Testing Service. Test of English as a Foreign Language, Test of Spoken English, Test of Written English, and Test of English for International Communication are trademarks of Educational Testing Service.

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

Graduate Management Admission Test and GMAT are registered trademarks of the Graduate Management Admission Council.

## **Abstract**

The Common European Framework describes language proficiency in reading, writing, speaking, and listening on a six-level scale. The Framework provides a common language with which to discuss students' progress. This report describes a study conducted with two panels of English language experts to map scores from four tests that collectively assess Reading, Writing, Speaking, and Listening on to two levels of the Framework. Panel 1 recommended cut scores for the Test of English as a Foreign Language™ (TOEFL®), the Test of Spoken English™ (TSE®), and the Test of Written English™ (TWE®). Panel 2 recommended cut scores for The Test of English for International Communication™ (TOEIC®). A modification of the Angoff (1971) standard-setting approach was used for multiple-choice questions, and a benchmark method (Faggen, 1994) or examinee paper selection method (Hambleton, Jaeger, Plake, & Mills, 2000) was used for constructed-response questions.

Key words: English language tests, standard setting, Angoff, benchmark method

---

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.

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, junior and community colleges, nonprofit educational exchange agencies, and agencies of the United States government.



A continuing program of research related to the TOEFL test is carried out 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, 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.

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

Current (2004-2005) members of the TOEFL Committee of Examiners are:

Catherine Elder (Chair)	Monash University
Deena Boraie	The American University in Cairo
Micheline Chalhoub-Deville	University of Iowa
Glenn Fulcher	University of Dundee
Marysia Johnson Gerson	Arizona State University
April Ginther	Purdue University
Bill Grabe	Northern Arizona University
Keiko Koda	Carnegie Mellon University
David Mendelsohn	York University
Tim McNamara	The University of Melbourne
Terry Santos	Humboldt State University

---

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

**E-mail: [toefl@ets.org](mailto:toefl@ets.org)**  
**Web site: [www.ets.org/toefl](http://www.ets.org/toefl)**

## Table of Contents

Introduction.....	1
Standard Setting.....	2
Section I: Methods .....	3
Panelist Orientation.....	3
Panelist Training .....	4
Standard-Setting Process for Constructed-Response Tests .....	5
Standard-Setting Process for Selected-Response (Multiple-Choice) Tests .....	7
Section II: Panel 1—TSE, TWE, TOEFL.....	10
Panelists .....	10
Standard Setting Results for TSE, TWE, TOEFL .....	10
Section III: Panel 2—TOEIC.....	13
Panelists .....	13
Standard Setting Results for TOEIC.....	13
Conclusions.....	16
References.....	18
Notes .....	19
List of Appendixes.....	20

## **List of Tables**

Table 1. TSE, TWE, TOEFL Panel Demographics .....	11
Table 2. First- and Second-Round B1-Level Judgments: TSE, TWE, TOEFL.....	12
Table 3. First- and Second-Round C1-Level Judgments: TSE, TWE, TOEFL.....	12
Table 4. TOEIC Panel Demographics .....	14
Table 5 . First- and Second-Round B1-Level Judgments: TOEIC .....	15
Table 6. First- and Second-Round C1-Level Judgments: TOEIC .....	15
Table 7. Summary of B1 and C1 Recommended Cut Scores .....	16

## Introduction

Descriptive taxonomies of language ability levels, such as the Common European Framework (CEF), serve to articulate well-constructed expectations of language proficiency for language learners. The CEF targets all modern European languages, including English, and “describes in a comprehensive way what language learners have to learn to do in order to use a language for communication and what knowledge and skills they have to develop so as to be able to use it effectively” (*The Common European Framework*, n.d., p. 1). The CEF defines six levels of proficiency clustered in three bands: A1 – A2 (“Basic User”), B1 – B2 (“Independent User”), and C1 – C2 (“Proficient User”). These levels provide guidance to language educators and instructors to identify existing levels of language competency of language learners and to develop curriculum and courses to advance communicative competence. “The Framework also defines levels of proficiency, which allow learners’ progress to be measured at each stage of learning and on a life-long basis” (*The Common European Framework*, p. 1). Furthermore, and most relevant to this study, the Framework provides a means of “describing the levels of proficiency in existing tests and examinations, thus enabling comparisons to be made across different systems of examinations” (*The Common European Framework*, p. 19).

The purpose of this study was to identify scores on a series of English language tests—The Test of English as a Foreign Language™ (TOEFL®), The Test of Spoken English™ (TSE®), The Test of Written English™ (TWE®), and The Test of English for International Communication™ (TOEIC®)—that correspond to the B1 and C1 proficiency levels of the CEF. The B1 level reflects the entering point of the “Independent User” band and the C1 level the entering point of the “Proficient User” band. By mapping test scores onto the CEF, an operational bridge is built between the descriptive levels of the CEF and psychometrically sound, standardized assessments of English language competencies, facilitating meaningful classification of CEF-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. Scores from each test were independently mapped to the two CEF levels.

The CEF levels are not based on a particular test, but band descriptors. Thus, the approach of asking a relevant group of examinees to take both the CEF “test” and one of the tests of interest to this study and correlating the results was not feasible. Since scores and score bands



cannot be easily related via a concordance table, a standard setting approach provided a suitable alternative as described below.

### **Standard Setting**

The process followed to map test scores onto the CEF is known as standard setting. Standard setting is a general label for a number of approaches used to identify test scores that support decisions about test takers' (candidates') level of knowledge, skill, proficiency, mastery, or readiness. For example, in order for an international student to gain admittance into a university where the language of instruction is English (such as a North American university or a European university), typically he or she must achieve a certain score (standard) on the TOEFL. This score, set by each institution, reflects the minimum level of English language competence the particular institution believes necessary in order for that prospective student to function successfully at the institution. The score reflects a standard of "readiness to learn" subject matter taught in English at that institution. Students with TOEFL test scores at or above the threshold score have demonstrated a sufficient level of English proficiency to study at the institution; those with test scores below the threshold have not yet demonstrated a sufficient level of English language proficiency to study at the institution. In this example, one threshold score classifies students into two levels of proficiency; more than one threshold score may be established on the same test to classify candidates into multiple levels of proficiency.

It is important to recognize that a cut score, a threshold test score, is a function of informed expert judgment. There is no absolute, unequivocal cut score. There is no single "correct" or "true" score. A cut score reflects the values, beliefs, and expectations of those experts who participate in its definition and adoption, and different experts may hold different sets of values, beliefs, and expectations. Its determination may be informed by empirical information or data, but ultimately, a threshold score is a judgment-based decision.

As noted by the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education [AERA, APA, & NCME], 1999), 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. The standard-setting procedures in this study were designed to comply with these guidelines; the methods and results of the study are described below.

This report is presented in three major sections. The first section describes the standard-setting methods that were implemented to establish the cut scores corresponding to a B1 CEF proficiency level and a C1 CEF proficiency level on each of the English language tests. The second section focuses on the results specific to the TSE, TWE, and TOEFL. The third section focuses on the TOEIC. Two different panels of experts (with minimal overlap) were convened to participate in setting the B1 and C1 cut scores on the tests—one panel for the TSE, TWE, TOEFL, and one panel for the TOEIC. The two panels reflected the different contexts for which the tests are primarily used—Panel 1: higher education and Panel 2: business. The composition of each panel is discussed in more detail at that start of Section II and Section III.

## **Section I: Methods**

### ***Panelist Orientation***

Panelists were provided with an overview of the purpose of the study and a definition of threshold scores (or cut scores), as applied to the current purpose. Appendix A provides the agendas for both panels. Cut scores were defined as the level of performance on each of the tests that reflected the English language proficiency of a candidate who was just at the B1 level on the CEF and of a candidate who was just at the C1 level on the CEF.<sup>1</sup> Each cut score was defined as the minimum score believed necessary to qualify a candidate at the B1 level and at the C1 level. The panelists were also provided with brief overviews of each of the tests for which they would be mapping scores onto the CEF (setting cut scores).

- *TSE*. The TSE measures the ability of nonnative speakers of English to communicate orally in English. It consists of nine items for which a candidate must generate a verbal response involving, for example, narration, persuading, recommending, and giving and supporting an opinion. Responses to each item are scored using a rubric ranging from a low of 20 to a high of 60 in 10-point intervals. Twelve independent assessors contribute to a candidate's overall TSE score.<sup>2</sup> Items scores are averaged to

- arrive at the overall score, which is reported in intervals of five: 20, 25, 30, 35, 40, 45, 50, 55, 60.
- *TWE*. The TWE measures the ability of non-native writers of English to produce an essay in response to a given topic, demonstrating their ability to generate and organize ideas, to support those ideas with examples, and to use conventions of standard written English. The response is scored using a features-defined rubric ranging from a low of 1 to a high of 6 in one-point intervals. Two independent assessors score the response and a mean score is computed; the overall TWE score, therefore, is reported in half-point intervals: 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0, 5.5, 6.0.
  - *[Paper-based] TOEFL*. The TOEFL measures the ability of nonnative communicators of English ability to understand English as it is spoken in North American academic contexts (Listening skill), to recognize language appropriate for standard written English (Structure skill), this being part of the larger writing construct complemented by the TWE, and to understand short passages similar in topic and style to academic texts used in North American colleges and universities (Reading skill). The [paper-based] TOEFL is a selected-response (multiple-choice) test that is reported on a scale that ranges from a low of 310 to a high of 677.
  - *TOEIC*. The TOEIC measures the ability of nonnative English communicators to communicate in English in the global workplace. The TOEIC addresses listening comprehension skills and reading comprehension skills. The test items are developed from samples of spoken and written English from countries around the world. The TOEIC is a selected-response test that is reported on a scale that ranges from a low of 10 to a high of 990.<sup>3</sup>

### ***Panelist Training***

The first major event of the training process had panelists summarizing the key descriptors of a B1 level of proficiency on the CEF and a C1 level of proficiency. This was done in two small groups, one focusing on the B1 level and the other on the C1 level. Each level was defined in terms of the English-language skill(s) being measured by the particular test that was the immediate focus. For example, the first test to be addressed by Panel 1 was the TSE;

therefore, the B1 and C1 levels of the CEF corresponding to speaking were summarized. Each small group was asked to record on chart paper the main points that defined their assigned CEF level. This exercise was designed to bring the groups to an agreed upon, shared understanding of the CEF levels. Each group's charted summary was posted and discussed so that the whole panel had an opportunity to comment and, as appropriate, suggest modifications. The whole-panel agreed-upon summaries remained posted to guide the standard-setting judgment process for the TSE. (Prior to the meeting, each panelist was given a homework assignment to review the CEF and selected tables of level descriptors and to write down key indicators. Panelists were encouraged to refer to their homework notes to facilitate the summarization exercise. See Appendix B for a copy of the homework assignment, and Tables C1 through C6 in Appendix C for copies of each panel's agreed-upon key indicator summaries for each language skill.

The exercise of summarizing the B1 and C1 levels was repeated in turn for each of the language skills addressed by the test of under consideration. Once the standard-setting judgments were completed for the TSE, the TWE was presented, so the summary process was repeated for writing. After standard-setting judgments were completed for the TWE, the TOEFL became the focus. The exercise was then repeated for reading and listening. The CEF Writing descriptors included aspects of writing relevant to the TOEFL structure section, although the CEF writing descriptors focus on grammar in the context of productive skills while the TOEFL structure section assesses receptive skills. The panel that worked on the TOEIC test completed the same exercise for listening and reading, the two language skills measured by the TOEIC.

### ***Standard-Setting Process for Constructed-Response Tests***

The TSE and the TWE are both constructed-response tests that require candidates to produce original responses, rather than select from a set of given options, as in the case of multiple-choice tests. The standard-setting process as applied to the TSE will be described in some detail. An abbreviated presentation of the process will follow for the TWE because the same process was used in both cases.

The standard-setting process applied to the TSE is variously known as the benchmark method (Faggen, 1994) or the examinee paper selection method (Hambleton, Jaeger, Plake, & Mills, 2000). As applied to the TSE, the process included the panelists first reviewing the nine items of the TSE and the scoring rubric. Operationally, the panelists were asked to read a TSE item and to listen to sample spoken responses to the item that served to illustrate each whole-

number score point on the rubric (20, 30, 40, 50, 60). The panelists were asked to consider the difficulty of the English language skill addressed by the item, the language features valued by the rubric, and the skill set of a B1-level candidate (as previously defined). Panelists, independently, were asked to pick the lowest scoring sample response that, in their expert judgment, most appropriately reflected the response of a candidate who was just at the B1-level of proficiency. Because, as noted previously, TSE responses are averaged, panelists were able to pick from among the range of reported scores (20, 25, 30, 35, 40, 45, 50, 55, 60). So for example, if a panelist believed that a B1-level candidate would score higher than a 30 on an item, but not quite as high as a 40, the panelist would be able to pick a score of 35. They were then asked to repeat the judgment process for a candidate at the C1-level of proficiency. This basic process was followed for each of the nine TSE items.

Panelists independently completed their B1 judgment and C1 judgment for the first TSE item and were asked to stop. Panelists were then asked to share their judgments for the first item—what scores did they give for the B1-level candidate and the C1-level candidate? The purpose of the facilitated discussion was for panelists to hear the judgment rationales of their peers. The goal was to make more explicit the diversity of relevant perspectives reflected by the panel and to give panelists an opportunity to consider a viewpoint that they had not previously considered; the goal was not to have panelists conform to single expectation of B1 or C1 levels of performance on TSE items. 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. Panelists were given the chance to change their B1 and C1 judgments for the first item before proceeding, independently, on to the remaining eight items of the TSE. The completion of the B1 and C1 judgments for all nine of the TSE items was considered to be first-round judgments.

The ETS facilitators computed each panelist's B1 and C1 standard-setting judgments for the TSE, taking the mean score across the nine items for each panelist. The mean cut score across all panelists was computed, as was the median, standard deviation, minimum cut score,

and maximum cut score. The cross-panelist summary information was posted and used to facilitate a discussion. Each panelist also had his or her own B1 and C1 TSE cut scores. In general, the panelists with the minimum cut score and maximum cut score were asked to begin the discussion, with other panelists encouraged to share their cut scores and decision rationales. At the conclusion of the group discussion, the panelists were given an opportunity to change their overall B1 and C1 TSE cut scores if they felt that they wished to reflect some aspect of the discussion in their final judgment. Having considered each item separately for the first-round judgment, and so become familiar with the demands of the test, panelists were asked to consider overall performance for their second round judgments. The discussion had begun with a presentation of the mean raw total scores, and panelists had discussed their decision rationales in relation to the total score. Thus, making their second-round judgments at the overall level was in keeping with nature of the discussion, and panelists were easily able to make the transition. Panelists were reminded that they could keep their first-round cut scores; they were not obligated or expected to change their cut scores. Panelists then recorded their second-round (final) judgments. (See the Appendix D for a copy of the judgment recording form—for first-round and second-round decisions—completed by each panelist.)

This basic process was also applied to the TWE, which is also a constructed-response assessment for which candidates produce an essay in response to a given topic. There is only one topic, so, in essence, the TWE is a single-item test. As with the TSE, panelists reviewed the essay topic and scoring rubric. They then reviewed sample essays illustrative of each of the rubric score points. Panelists were asked to independently select the sample response that, in their expert judgment, reflected most appropriately the response of a candidate just at the B1-level of proficiency, and then select another response just at the C1-level of proficiency. Panelists were able to pick half-point scores, as with the TSE. So, for example, if a panelist believed that a B1-level candidate would score higher than a 4, but not quite as high as a 5, the panelist would be able to pick a cut score of 4.5. The first-round of independent judgments was followed by a whole-group discussion. Panelists were then given the opportunity to change their B1- and C1-level judgments.

### ***Standard-Setting Process for Selected-Response (Multiple-Choice) Tests***

The [paper-based] TOEFL and TOEIC tests are selected-response tests whereby candidates chose or select a response to an item from a given set of options. The standard-setting

process as applied to the TOEFL will be described in some detail. The same process was subsequently applied to the TOEIC test reviewed by Panel 2.

The general standard-setting process applied to the TOEFL is known as the Angoff method (Angoff, 1971). The general approach remains the most widely used standard-setting method for selected-response tests (Mehrens, 1995; Cizek, 1993; Hartz & Auerbach, 2003). The first section of the TOEFL test addressed was Structure. This section measures the ability of a non-native communicator to recognize language appropriate for standard written English. There are 40 items in the Structure section. As applied to the Structure section, panelists were asked to read an item, consider the difficulty of the English-language skill addressed by the item, and to judge the probability that a B1-level candidate would know the correct response. Panelists recorded their item-level judgments on a form (see the Appendix D for an example of a judgment form used for a selected-response section), with the following probability scale: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9. A judgment of 0.1, for example, corresponds to a 10 percent probability of knowing the correct answer. As a rule-of-thumb, panelists were informed that a difficult item (for the general TOEFL test-taking population)—that is, one that requires a relatively high-level of English proficiency—might fall into the range of 0.1 to 0.3: a 10- to 30-percent probability of knowing the correct answer. A relative easy item might fall into the 0.7 to 0.9 range: 70- to 90-percent probability of knowing the correct answer; and a moderately difficult item might fall into the range of 0.4 to 0.6: 40- to 60-percent probability of knowing the correct answer.

Prior to making their “live” first-round standard-setting judgments for the Structure items, panelists were given an opportunity to practice making judgments on five sample Structure items from a previously administered [paper-based] edition of the TOEFL. The edition chosen had been administered to more than 750,000 examinees in 1997–1998. For each sample item, each panelist was asked to record the probability that a B1-level candidate would know the correct answer and then the probability that a C1-level candidate would know the correct answer (practice recoding forms were provided.) Once each panelist noted his or her response, a whole-group discussion occurred in which panelists were asked to share their item-level decision rationales. After the discussion of each item, the correct answer was revealed, as was the proportion of 1997–1998 examinees that chose the correct answer, and whether the item would be classified as being easy, of medium difficulty, or difficult, based on our rule-of-thumb guidelines. (The facilitators clarified that these percent correct values were based on the general

population of TOEFL examinees and that the panels' task was to consider how an examinee at the B1 level and an examinee at the C1 level would perform.) Panelists were encouraged to discuss the practice items in terms of their difficulty level, and what might make each one more or less challenging to both a B1- and a C1-level English language learner. The practice session helped to calibrate the panelists and helped to make explicit the range of relevant professional perspectives reflected by the panel. The practice session also helped to clarify any misunderstanding of the judgment process. Panelists were then asked to complete their "live" B1- and C1-level judgments for the first four items of the Structure section and then to stop. This provided an opportunity to answer panelists' questions. The panelists confirmed that they understood the process and were then asked to complete their round-one judgments for the Structure section.

The ETS facilitators computed each panelist's B1 and C1 standard-setting judgments for the TOEFL Structure section, summing the probabilities across the 40 items, first for the B1 judgments and then for the C1 judgments. For example, if a panelist recorded a 0.5 for each of the 40 items for a B1-level candidate, that panelist's B1 score would be 20; so according to that panelist, 20 items would need to be answered correctly for a candidate to be considered at the B1 proficiency level of the CEF. If a panelist recorded 0.8 for each of the 40 items for a C1-level candidate, that panelist's C1 score would be 32; so according to that panelist 32 items would need to be answered correctly for a candidate to be considered at the C1 proficiency level of the CEF. The mean score across all panelists was computed, as was the median, standard deviation, minimum score, and maximum score. The cross-panelist summary information was posted and used to facilitate a discussion. Each panelist also had his or her own B1 and C1 TOEFL Structure scores. In general, the panelists with the minimum score and maximum score were asked to begin the discussion, with other panelists encouraged to share their judgments. At the conclusion of the group discussion, the panelists were given an opportunity to change their overall B1 and C1 TOEFL Structure scores. Similar to the constructed response sections, panelists adjusted overall scores rather than item level decisions since the discussion phase was conducted at the section level. Panelists were reminded that they could keep their first-round scores; they were not obligated or expected to change their scores. Panelists then recorded their second-round (final) judgments.



This same process of practice and discussion followed by “live” round-one judgments, discussion, and a final (round-two) judgment, was followed for the 44 reading items and the 50 listening items of the TOEFL. (For the Listening section, panelists listened to the tape-recorded speaking stimulus for each item.) The same process was also implemented for the TOEIC test sections, 100 reading items and 100 listening items addressed by Panel 2.

## **Section II: Panel 1—TSE, TWE, TOEFL**

### ***Panelists***

Twenty-one experts served on the panel that focused on mapping scores from the TSE, the TWE, and the TOEFL onto the CEF. The English language specialist from ETS Europe, located in Utrecht, The Netherlands, 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 CEF. They were also selected to represent an array of European countries where TOEFL is used. Table 1 presents the demographic characteristics of the 21 panelists. Appendix E provides the panelists’ affiliations.

### ***Standard Setting Results for TSE, TWE, TOEFL***

The first-round and second-round section-level judgments for each of the tests are presented in a series of tables (Tables F1 through F5 in Appendix F). Each panelist’s individual B1 and C1 cut scores are presented for each round, as are the cross-panel summary statistics (mean, median, standard deviation, minimum, and maximum).

Table 2 presents B1-level cross-panel statistics for each of the three tests, and Table 3 presents the statistics for the C1-level judgments. Note that for the TOEFL test, first the raw-score statistics are presented by section, and then the overall total TOEFL scaled score. Scaled scores reflect the [paper-based] TOEFL reporting scale. The mean and median raw section-level scores were translated to scaled scores using a conversion table. The total TOEFL scaled scores were computed from the sum of the section scaled scores multiplied by ten thirds. The scaled scores represent the panel-recommended B1 and C1 cut scores for TOEFL. The presented TSE and TWE scores reflect the reporting scale for these tests.

**Table 1*****TSE, TWE, TOEFL Panel Demographics***

	Number	Percent
Gender		
Female	11	52%
Male	10	48%
Panelist selection criteria <sup>a</sup>		
Teacher of English as a second language within a language school or within a language center of a university	18	
Administrator of School/Program where TOEFL classes/equivalent taught	11	
Assessment expert in field of English as a second/foreign language	8	
Member of assessment policy group for assessing second/foreign languages within the CEF	8	
Country <sup>b</sup>		
Belgium	1	5%
Germany	2	9%
Greece	1	5%
Hungary	3	14%
Italy	3	14%
Malta	1	5%
Netherlands	1	5%
Norway	1	5%
Poland	1	5%
Scotland	1	5%
Slovakia	1	5%
Slovenia	1	5%
Spain	2	9%
Sweden	1	5%
Turkey	1	5%

<sup>a</sup>Some members met more than one criterion, so percentages are not reported. <sup>b</sup>Some members represented a region that they were not originally from.

**Table 2*****First- and Second-Round B1-Level Judgments: TSE, TWE, TOEFL***

B-1 judgments Test	Round 1			Round 2		
	Mean	Median	SD	Mean	Median	SD
TSE	46	46	2.9	45	45	1.1
TWE	4.3	4.3	0.3	4.4	4.5	0.3
TOEFL Structure section (raw scores)	19	19	4.5	19	19	3.2
TOEFL Reading section (raw scores)	23	23	6	23	23	4.4
TOEFL Listening section (raw scores)	23	23	4.8	23	23	3.7
[Paper-based] TOEFL (scaled scores)	457	457		457	457	

*Note.* Mean and median values are truncated.

**Table 3*****First- and Second-Round C1-Level Judgments: TSE, TWE, TOEFL***

C-1 judgments Test	Round 1			Round 2		
	Mean	Median	SD	Mean	Median	SD
TSE	57	57	2.6	56	55	2.2
TWE	5.7	5.5	0.2	5.7	5.5	0.3
TOEFL Structure section (raw scores)	31	31	1.5	31	31	1.3
TOEFL Reading section (raw scores)	36	37	1.4	36	36	1.3
TOEFL Listening section (raw scores)	41	41	2.0	41	41	1.5
[Paper-based] TOEFL (scaled scores)	560	563		560	560	

*Note.* Mean and median values are truncated.

The B1 and C1 cut score means (and medians) changed very little from round one to round two as can be seen in Table 2 and Table 3. The variability (standard deviation) of the panelists' judgments for both the B1 and C1 levels tended to decrease from round one to round two, indicating a greater degree of panelist consensus; although there was a nominal increase in the TSE variability.

The second-round mean scores may be accepted as the panel-recommended cut scores, that is, the minimum scores necessary to qualify for the B1- and C1-levels on the CEF. Thus the TSE B1 and C1 cut scores are 45 and 55<sup>4</sup> respectively, for the TWE they are 4.5<sup>5</sup> and 5.5<sup>6</sup>, respectively, and for the [paper-based] TOEFL the scaled scores are 457 and 560, respectively.

### **Section III: Panel 2—TOEIC**

#### ***Panelists***

Twenty-one experts served on the panel that focused on mapping the TOEIC onto the CEF. (Three panelists had also served on the TSE, TWE, and TOEFL panel.) The Director of Development, TOEIC Programme, France, and the language specialist from ETS Europe, organized the recruitment of the experts. The experts were selected for their experience with English language instruction, learning, and testing in the workplace, and their familiarity with the CEF. They were also selected to represent an array of European countries where the TOEIC is used.

#### ***Standard Setting Results for TOEIC***

The first-round and second-round section-level judgments for the TOEIC test are presented in Table H1 (Listening) and Table H2 (Reading) in Appendix H. Each panelist's individual B1 and C1 cut scores are presented for each round, as are the cross-panel summary statistics (mean, median, standard deviation, minimum, and maximum).

Table 4 presents the demographic characteristics of the 21 panelists. Appendix G provides the panelists' affiliations.

Table 5 presents B1-level cross-panel statistics for the Reading and Listening TOEIC sections; and Table 6 presents the same for the C1-level statistics. The TOEIC total scaled score means and medians were obtained from the sum of the section level scaled scores (Listening and Reading), which came from a raw-to-scaled score TOEIC conversion table. The scaled scores represent the panel-recommended B1 and C1 cut scores for TOEIC.

**Table 4*****TOEIC Panel Demographics***

	Number	Percent
Gender		
Female	10	48%
Male	11	52%
Panelist selection criteria <sup>a</sup>		
Teacher of English as a Second Language within a Language School or Language Center of a University	13	
Administrator of School/Program offering TOEIC classes/equivalent	16	
Assessment expert in field of English as a Second/Foreign Language	8	
Member of Assessment Policy group for assessing Second/Foreign languages within the CEF	6	
Human Resources administrator responsible for language training	2	
Country <sup>b</sup>		
Belgium	1	5%
England	1	5%
France	3	14%
Germany	5	24%
Greece	1	5%
Hungary	4	19%
Ireland	1	5%
Italy	2	10%
Malta	1	5%
Poland	1	5%
Switzerland	1	5%

<sup>a</sup>Some members met more than one criterion, so percentages are not reported. <sup>b</sup>Some members represented a region that they were not originally from.

**Table 5*****First- and Second-Round B1-Level Judgments: TOEIC***

B-1 judgments	Round 1			Round 2		
	Mean	Median	SD	Mean	Median	SD
TOEIC Listening (raw scores)	58	60	8.7	58	59	6.1
TOEIC Reading (raw scores)	56	55	7.1	55	55	4.2
TOEIC (scaled scores)	555	565		550	560	

*Note.* Mean and median raw values are truncated.

**Table 6*****First- and Second-Round C1-Level Judgments: TOEIC***

C-1 judgments	Round 1			Round 2		
	Mean	Median	SD	Mean	Median	SD
TOEIC Listening (raw scores)	83	84	3.9	84	84	4.6
TOEIC Reading (raw scores)	83	83	3.5	83	82	3.3
TOEIC (scaled scores)	875	880		880	875	

*Note.* Mean and median raw values are truncated.

The B1 and C1 cut score means (and medians) changed slightly from round one to round two as can be seen in Tables 5 and 6. The variability (standard deviation) of the panelists' judgments for the B1 level decreased from round one to round two, indicating a greater degree of panelist consensus; the variability increased somewhat for the C1 level between the two rounds.

The second-round mean scores may be accepted as the panel-recommended cut scores, that is, the minimum scores necessary to qualify for the B1- and C1-levels on the CEF. Thus, the TOEIC B1 and C1 scaled cut scores are 550 and 880, respectively.

## Conclusions

The purpose of this study was to arrive at CEF B1-level and C1-level recommended cut scores on a series of language proficiency tests, thus creating an operational bridge between the descriptive levels of the CEF and standardized tests of English language proficiencies. The study was not intended or designed, however, to establish a concordance between scores on the series of English language tests. The mapping of TOEFL scores to the two CEF proficiency levels, for example, was independent of the mapping of the TOEIC scores. It is, therefore, inappropriate to infer a concordance between scores on any of the tests in the current context.

Two panels, each of 21 experts, were invited to participate in the standard-setting studies. The benchmark method (Faggen, 1994)—also referred to as the examinee paper selection method (Hambleton, Jaeger, Plake, & Mills, 2000)—and a modification of the Angoff method (1971) were applied to the constructed-response questions and selected-response questions respectively. Table 7 below summarizes the B1 and C1 cut scores for the four tests.

**Table 7**

*Summary of B1 and C1 Recommended Cut Scores*

Test	B1 cut score	C1 cut score
TSE	45	55
TWE	4.5	5.5
[Paper-based] TOEFL	457	560
TOEIC	550	880

It is common practice to set standards or cut scores on assessments used to classify test takers into one or more classifications of proficiency, competence, or readiness; and the Angoff method—with its many modifications—continues to be the most frequently implemented and supported approach for multiple-choice assessments (Hurtz & Hertz, 1999).

One common modification is the inclusion of more than one round of item-level judgments, with discussion between rounds (Busch & Jaeger, 1990). The rationale for such discussion—which may or may not be accompanied by normative data, such as item P-values—is that panelists have the opportunity to hear and consider other relevant perspectives, which they

can then incorporate into their next round of item-level judgments. The inclusion of discussion tends to result in higher cut scores and reduced variability (Hurtz & Auerbach, 2003).

The present study did include two rounds of judgments, with between-round discussion, but introduced a somewhat unique feature: the focus of the judgments between rounds shifted from the item-level (round one) to the domain or construct level (round two). In the case of TOEIC, for example, the shift was from judgments of each of the 100 discrete items within the Reading Comprehension section to judgments about the overall Reading Comprehension section. The first-round item-level judgments were important and necessary to engage the panelists in considerations of language skill difficulty levels posed by each of the items defining each skill domain. But given the holistic nature of language skills, it was believed to be more meaningful and appropriate for the second round of judgments to be framed in terms of each overall domain, Reading Comprehension and Listening Comprehension, again, using TOEIC as the example. Once panelists understood, through their engagement with the items, item content, and received feedback about their domain cut scores and the panels' cut scores (computed from the item judgments) for the B1 and C1 levels, the stage was set for meaningful discussion at the domain or construct level; hence, it was believed more meaningful and relevant to make post-discussion judgments at that same level, rather than deconstructing the domain, in essence, by repeating item-level judgments during the second round.

Whether comparable results would have been obtained by repeating the item-level judgments for round two is an empirical question, not answerable by this study. Nonetheless, the shift in emphasis or the lens of judgment between rounds merits additional exploration, particularly in the context of English language tests with well-delineated construct domains.



## References

- American Educational Research Association, American Psychological Association, and 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.
- Busch, J. C., & Jaeger, R. M. (1990). Influence of type of judge, normative information, and discussion on standards recommended for the National Teacher Examinations. *Journal of Educational Measurement*, 27, 145–163.
- Cizek, G. J. (1993). *Reactions to National Academy of Education report: Setting performance standards for student achievement*. Washington, DC: National Assessment Governing Board.
- Faggen, J. (1994). *Setting standards for constructed response tests: An overview* (ETS RM-94-19). Princeton, NJ: ETS.
- 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.
- Hurtz, G. M., & Auerbach, M. A. (2003). A meta-analysis of the effects of modifications to the Angoff method on cutoff scores and judgment consensus. *Educational and Psychological Measurement*, 63, 584–601.
- Hurtz, G.M., & Hertz, N.R. (1999). How many raters should be used for establishing cutoff scores with the Angoff Method? A generalizability theory study. *Educational and Psychological Measurement*, 59, 885–897.
- Mehrens, W.A. (1995). Methodological issues in standard setting for educational exams. In *Proceedings of Joint Conference on Standard Setting for Large-Scale Assessments* (pp.221–263). Washington, DC: National Assessment Governing Board and National Center for Education Statistics.
- The Common European Framework in its political and educational context*. (n.d.). Retrieved March 2, 2004, from the Council of Europe Web site:  
<http://culture2.coe.int/portfolio/documents/0521803136txt.pdf>

## Notes

- <sup>1</sup> These cut scores also may be considered to define the boundary or borderline between A2 and B1 proficiency and B2 and C1 proficiency.
- <sup>2</sup> For each candidate, three of the nine items are scored by two assessors (and the mean score used) in order to provide data regarding scorer consistency, and the remaining six items are each separately scored by a single assessor.
- <sup>3</sup> Different reporting scales are used across the tests (TSE, TWE, TOEFL, TOEIC) to avoid confusion and to help ensure that one score is not substituted for a score on another test that has a different meaning.
- <sup>4</sup> The TSE Round 2 mean C1 judgment was 56 points, but the reporting scale is in increments of 5. Thus, the C1 cut score is 55.
- <sup>5</sup> The TWE Round 2 mean B1 judgment was 4.4, but the reporting scale is in increments of .5. Thus the B1 cut score is 4.5.
- <sup>6</sup> The TWE Round 2 mean C1 judgment was 5.7, but the reporting scale is in increments of .5. Thus the B1 cut score is 5.5.

## List of Appendixes

	Page
A - Panel 1 and Panel 2 Agendas .....	21
B - Panel 1 and Panel 2 Homework Tasks.....	23
C - Panel 1 and 2 Indicator Summaries of Language Skills Defined by the CEF .....	29
D - Judgment Forms Used by Panelists .....	35
E - Panelists' Affiliations for Panel 1 .....	38
F - First- and Second-Round Judgments for TSE, TWE, and TOEFL.....	39
G - Panelists' Affiliations for Panel 2.....	44
H - First- and Second-Round Judgments for TOEIC.....	45

## Appendix A

### Panel 1 and Panel 2 Agendas

#### *Panel 1 Agenda: Mapping TOEFL, TWE, and TSE onto the CEF*

February 2<sup>nd</sup> to 4<sup>th</sup>, 2004 — Utrecht, The Netherlands.

#### *Day 1*

8:30 – 9:00	Breakfast
9:00 – 9:30	Introductions
9:30 – 10:00	Overview of ETS language tests, the CEF and the purpose of the study
10:00 – 10:30	Standard-setting training: constructed response items
10:30 – 11:15	Define candidate focal groups for Levels B1 and C1 on the CEF (Speaking)
11:15 – 11:30	Break
11:30 – 13:00	Training and standard-setting judgments on TSE
13:00 – 14:00	Lunch
14:00 – 14:15	Panelists review individual recommended TSE cut scores
15:15 – 15:00	Discussion and final standard-setting judgments
15:00 – 15:15	Break
15:15 – 15:30	Overview of the TOEFL
15:30 – 16:15	Define candidate focal groups for Levels B1 and C1 on the CEF (Writing)
16:15 – 16:30	Wrap up for the day and adjourn

#### *Day 2*

8:30 – 9:00	Breakfast
9:00 – 9:30	Recap of previous day's focal group definitions
9:30- 10:45	Training and standard-setting judgments on Writing component
10:45 – 11:00	Break
11:00 – 11:30	Standard-setting training: selected-response items
11.30 – 12:15	Standard-setting training: Practice judgments (Structure items)
12:15 – 13:00	Standard-setting judgments on Structure items
13:00 – 14:00	Lunch
14:00 – 14:30	Define candidate focal groups for Levels B1 and C1 on the CEF (Reading)
14:30 – 15:15	Standard-setting training: Practice judgments (Reading items)
15:15 – 15:30	Break
15:30 – 16:30	Standard-setting judgments on Reading items
16:30 – 16:45	Wrap up for the day and adjourn

### ***Day 3***

8:30 – 9:00	Breakfast
9:00 – 9:30	Define candidate focal groups for Levels B1 and C1 on the CEF (Listening)
9:30 – 10:15	Standard-setting training: Practice judgments (Listening items)
10:15 – 10:30	Break
10:30 – 12:00	Standard-setting judgments on Listening items
12:00 – 13:30	Extended lunch
13:30 – 14:30	Discussion and final standard-setting judgments
14:00 – 14:30	Wrap up and adjourn
14:30 – 15:00	Break

### ***Panel 2 Agenda: Mapping TOEIC Onto the CEF***

February 5<sup>th</sup> to 6<sup>th</sup>, 2004 — Utrecht, The Netherlands.

### ***Day 1***

8:30 – 9:00	Breakfast
9:00 – 9:30	Introductions
9:30 – 10:00	Overview of TOEIC, the CEF and the purpose of the study
10:00 – 10:30	Standard-setting training: Selected response items
10:30 – 11:15	Define candidate focal groups for Levels B1 and C1 on the CEF (Listening)
11:15 – 11:30	Break
11:30 – 12:30	Standard-setting training: Practice judgments (Listening items)
12:30 – 13:30	Lunch
13:30 – 15:30	Standard-setting judgments on Listening items
15:30 – 15:45	Break
15:45 – 16:30	Define candidate focal groups for Levels B1 and C1 on the CEF (Reading)
16:30 – 16:45	Wrap up for the day and adjourn

### ***Day 2***

8:30 – 9:00	Breakfast
9:00 – 9:30	Recap of previous day's focal group definitions
9:30 – 10:15	Standard-setting training: Practice judgments (Reading items)
10:15 – 10:30	Break
10:30 – 12:30	Standard-setting judgments on Reading items
12:00 – 13:30	Extended lunch
13:30 – 14:30	Discussion and final standard-setting judgments (Reading & Listening sections)
14:30 – 15:00	Wrap up and adjourn

## Appendix B

### Panel 1 and Panel 2 Homework Tasks

#### ***Homework Task for Panel 1: Study to Map the Test of English as a Foreign Language, the Test of Spoken English, and the Test of Written English Onto the Common European Framework***

The role of the Common European Framework (CEF) is to foster mutual understanding across countries for users and language testers by providing a common language to describe the stages of language learning. Educational Testing Service 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 the B1 and C1 levels of the CEF. In order to facilitate discussions, it is very important that you become familiar with the CEF in general and these two levels in particular. A PDF version of the Framework can be found at the following address: <http://www.culture2.coe.int/portfolio/documents/0521803136txt.pdf>

The ETS tests that we will be benchmarking at the study address the four modalities of Speaking, Listening, Reading, and Writing, and we will be discussing the characteristics of a B1 and C1 candidate by language modality. In the section below, relevant tables from the CEF have been identified by page number and title. Please review these CEF tables, paying close attention to the B1 and C1 levels. Highlight key words or phrases.

*Speaking*: page 58, Overall Oral Production; page 59, Sustained Monologue (both tables); page 74, Overall Spoken Interaction; page 75, Understanding a Native Speaker Interlocutor; page 112, Vocabulary Range; page 112, Vocabulary Control; page 117, Phonological Control

*Writing*: page 61, Overall Written Production; page 62, Reports and Essays; page 83, Overall Written Interaction; page 112 Vocabulary Range; page 112, Vocabulary Control; page 114, Grammatical Accuracy; page 118, Orthographic Control

*Listening*: page 66, Overall Listening Comprehension; page 66, Understanding Conversation between Native Speakers; page 67, Listening as a Member of a Live Audience

*Reading*: page 69, Overall Reading Comprehension; page 70, Reading for Orientation; page 70, Reading for Information and Argument

On the following sheet, at the top of the table, there is a global descriptor of a candidate with B1-level proficiencies and C1-level proficiencies. Having reviewed the relevant CEF tables, complete the attached sheet by briefly noting in your own words, in the space provided, the key characteristics or indicators from the CEF tables that describe an English Language learner who:

1. you believe is at the B1-level of proficiency
2. you believe is at the C1-level of proficiency

For example, considering first the tables that define proficiencies related to Speaking, review each of the CEF tables listed above, and identify critical descriptors in the tables that help you distinguish between the B1 and C1 levels of proficiency. For example, you might note among other things that a B1 learner can provide a “straightforward description as a linear sequence of points” while a C1 learner can provide a “clear, detailed description.”

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 a B1 and C1 English Language Learner***

---

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 that 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.

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

---

Please bring this completed sheet with you to the meeting February 2nd. Thank you.



## ***Homework Task for Panel 2: Study to Map the Test of English for International Communication Onto the Common European Framework***

The role of the Common European Framework (CEF) is to foster mutual understanding across countries for users and language testers by providing a common language to describe the stages of language learning. Educational Testing Service is seeking to benchmark one 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 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 the B1 and C1 levels of the CEF. In order to facilitate discussions, it is very important that you become familiar with the CEF in general and these two levels in particular. A PDF version of the Framework can be found at <http://www.culture2.coe.int/portfolio/documents/0521803136txt.pdf>.

The ETS test that we will be benchmarking at the study addresses the modalities of Listening and Reading, and we will be discussing the characteristics of a B1 and C1 candidate by language modality. In the section below, relevant tables from the CEF have been identified by page number and title. Please review these CEF tables, paying close attention to the B1 and C1 levels. Highlight key words or phrases.

*Listening:* page 66, Overall Listening Comprehension; page 66, Understanding Conversation between Native Speakers; page 67, Listening as a Member of a Live Audience

*Reading:* page 69, Overall Reading Comprehension; page 70, Reading for Orientation; page 70, Reading for Information and Argument

On the following sheet, at the top of the table, there is a global descriptor of a candidate with B1-level proficiencies and C1-level proficiencies. Having reviewed the relevant CEF tables, complete the attached sheet by briefly noting in your own words, in the space provided, the key characteristics or indicators from the CEF tables that describe an English language learner who:

1. you believe is at the B1-level of proficiency
2. you believe is at the C1-level of proficiency

For example, considering first the tables that define proficiencies related to Listening, review each of the CEF tables listed above, and identify critical descriptors in the tables that help you distinguish between the B1 and C1 levels of proficiency. For example, you might note among other things that a B1 learner can understand “straightforward factual information” while a C1 learner can understand “extended speech on abstract and complex topics.”

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 a B1 and C1 English Language Learner***

---

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 that 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.

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.

---

Listening

---

Reading

---

Please bring this completed sheet with you to the meeting February 5th. Thank you

## Appendix C

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

**Table C1**

***Panel 1 Indicators of B1 and C1 Proficiency in Speaking***

---

B1 Speaking
<ul style="list-style-type: none"><li>• Can speak about familiar topics</li><li>• Can convey an opinion</li><li>• Sufficient structures/templates to express oneself</li><li>• Clear and intelligible speech</li><li>• Coping strategies for filling in gaps in language knowledge when speaking</li><li>• Can extract/understand the major points from conversations, presentations</li><li>• Understands speech when it is clear, standard language, well articulated</li></ul>
C1 Speaking
<ul style="list-style-type: none"><li>• Detailed and complex range of subjects and language</li><li>• Extensive range of lexis and idioms</li><li>• Generally accurate, no significant errors</li><li>• Speech is fluent and effortless</li><li>• Can tailor language to be appropriate in a wide range of contexts and situations</li><li>• Can vary intonation and stress in order to express shades of meaning</li></ul>

---

**Table C2**

***Panel 1 Indicators of B1 and C1 Proficiency in Writing***

---

<b>B1 Writing</b>
<ul style="list-style-type: none"><li>• Can write straightforward connected texts</li><li>• Can respond to topics of personal interest and give simply formulated opinions about factual (and abstract) information</li><li>• Errors in everyday vocabulary, but they do not hinder understanding</li><li>• Errors in grammar, but they do not hinder understanding</li><li>• Intelligible spelling, punctuation, and layout</li></ul>
<b>C1 Writing</b>
<ul style="list-style-type: none"><li>• Genre awareness – proper use of one genre</li><li>• Register awareness – proper use of register</li><li>• Appropriate style</li><li>• Consistent point of view</li><li>• Can manage abstractions</li><li>• Clarity and precision of expression – coherence/cohesion, lexicon-grammar, idiomatic structures/collocation</li><li>• Ability to support, expand and conclude complex arguments</li><li>• Layout, paragraphing and punctuation consistent and appropriate</li><li>• Flexibility, effectiveness, and efficiency of language choices (allusions, etc.)</li></ul>

---

**Table C3**

***Panel 1 Indicators of B1 and C1 Proficiency in Listening***

---

B1 Listening
<ul style="list-style-type: none"><li>• Can process simple conversations</li><li>• Understand standard familiar accent, clear speech</li><li>• Comprehend straightforward information</li><li>• Can pick out main points of message</li><li>• Understands details on familiar topics</li></ul>
C1 Listening
<ul style="list-style-type: none"><li>• Can follow most lectures, broadcasts, debates</li><li>• Can decide what's relevant</li><li>• Can identify attitudes and implied information</li><li>• Can understand nonstandard accents and speaking with some sound interference</li><li>• Can understand idioms and colloquial speech</li><li>• May not be able to understand everything in culturally or context-related situations (e.g., sitcoms)</li></ul>

---

**Table C4**

***Panel 1 Indicators of B1 and C1 Proficiency in Reading***

---

B1 Reading
<ul style="list-style-type: none"><li>• Can understand straightforward factual texts on familiar subjects</li><li>• Can scan for facts and details</li><li>• Can skim for gist</li><li>• Can follow lines of argument and identify main conclusions in argumentative texts</li><li>• Can understand and identify clearly signaled, explicitly stated information</li><li>• Can recognize and comprehend basic genres</li></ul>
C1 Reading
<ul style="list-style-type: none"><li>• Can comprehend detailed, lengthy texts outside one's field</li><li>• Can understand the main points, finer points and details in a wide variety of professional and academic texts</li><li>• Can scan newspapers, articles and reports on a wide range of topics and decide what is relevant for further detailed reading</li><li>• Can identify attitudes and implied information</li></ul>

---

**Table C5**

***Panel 2 Indicators of B1 and C1 Proficiency in Listening***

---

B1 Listening
<ul style="list-style-type: none"><li>• Can understand:<ul style="list-style-type: none"><li>• Straightforward, factual information</li><li>• Standard speech (everyday conversation) in standard dialect/ accent, clearly articulated</li></ul></li><li>• Can follow main points of discussion between native speakers</li><li>• Can follow / understand clear, concrete instructions presented in a structured manner</li><li>• Can follow a speech/ lecture on a familiar subject</li></ul>
C1 Listening
<ul style="list-style-type: none"><li>• Can understand abstract and complex and technical topics – beyond own field 85%</li><li>• Can understand:<ul style="list-style-type: none"><li>• Non-standard usage</li><li>• Unpredictable situations/ context</li></ul></li><li>• Can understand fast native speaker speech in films/ newscasts/ lectures</li><li>• Can understand<ul style="list-style-type: none"><li>• Cultural contexts</li><li>• Implicit meanings</li><li>• Implied relationships</li></ul></li><li>• Can identify levels of formality with relative ease</li><li>• Can understand extended / lengthy speech in most professional contexts</li></ul>

---



**Table C6**

***Panel 2 Indicators of B1 and C1 Proficiency in Reading***

---

B1 Reading
<ul style="list-style-type: none"><li>• Can understand straightforward factual texts on familiar subjects</li><li>• Scans long texts on everyday material and extracts main points</li><li>• Can recognize explicit line of argument in topic/issue, but not in finer detail</li><li>• Can identify main conclusions in logically organized texts</li></ul>
C1 Reading
<ul style="list-style-type: none"><li>• Can understand in detail lengthy/ complex, related/ unrelated texts</li><li>• Can differentiate between key and subtle points (weigh information)</li><li>• Requires occasional support/re-reading in lengthier, more technical texts</li><li>• Can scan quickly for specific/implicit information</li></ul>

---

## Appendix D

### Judgment Forms Used by Panelists

#### *Test of Spoken English*

##### *Round 1 Judgments*

Item	Circle the score that a Level <u>B1</u> candidate would achieve on each item										Circle the score that a Level <u>C1</u> candidate would achieve on each item									
1	20	25	30	35	40	45	50	55	60	20	25	30	35	40	45	50	55	60		
2	20	25	30	35	40	45	50	55	60	20	25	30	35	40	45	50	55	60		
3	20	25	30	35	40	45	50	55	60	20	25	30	35	40	45	50	55	60		
4	20	25	30	35	40	45	50	55	60	20	25	30	35	40	45	50	55	60		
5	20	25	30	35	40	45	50	55	60	20	25	30	35	40	45	50	55	60		
6	20	25	30	35	40	45	50	55	60	20	25	30	35	40	45	50	55	60		
7	20	25	30	35	40	45	50	55	60	20	25	30	35	40	45	50	55	60		
8	20	25	30	35	40	45	50	55	60	20	25	30	35	40	45	50	55	60		
9	20	25	30	35	40	45	50	55	60	20	25	30	35	40	45	50	55	60		

---

**Do Not Write in this Space.**

	End of round 1 B1 cut score	End of round 1 C1 cut score
My initial recommended cut score (range 20 – 60)		
Group average		

##### *Round 2 Judgments*

	Write the overall score that a Level <u>B1</u> candidate would achieve this test	Write the overall score that a Level <u>C1</u> candidate would achieve this test
My final recommended cut score (range 20 – 60)		

---

**(Signature)**



---

---

**Do Not Write in this Space.**

	End of round 1, B1 cut score	End of round 1, C1 cut score
My initial recommended cut score (range 0 to 44)		
Group average		

---

*Round 2 Judgments*

	Write the overall score that <u>Level B1</u> candidate would achieve on this test	Write the overall score that <u>Level C1</u> candidate would achieve on this test
My final recommended cut score (range 0 to 44)		

---

\_\_\_\_\_  
**(Signature)**

## Appendix E

### Panelists' Affiliations for Panel 1

#### *Panel 1 Standard-Setting Participants*

Name	Affiliation
Charles van Leeuwen	Universiteit Maastricht, the Netherlands
Christine Räisänen	Chalmers University of Technology, Göteborg, Sweden
Craig Dicker	Bureau of Education and Cultural Affairs, US Department of State, Hungary
Dede Teeler	Communications Officer of Computer SIG, IATEFL, Italy
Eberhard Fugmann	Oberschulamts Freiburg, Germany
Ekaterini Nikolarea	School of Social Sciences, University of the Aegean, Greece
Ewa Osiecka	CODN – National Inservice Teacher Training Center, Poland
Gabor Rebek-Nagy	Pecs University Medical School, Hungary
Glen Fulcher	Centre for Applied Language Studies, University of Dundee, UK
Herbert Doebler	Oberschulamts Stuttgart, Germany
Jana Beresova	Trnava University, Slovakia
Lucia Katona	Institute for Foreign Languages, Hungary
Lut Baten	Institute for Modern Languages, University of Leuven, Belgium
Martin Musumeci	Academic Division, MABEC Support Unit, University of Malta
Michael Fields	Isik University, Istanbul, Turkey
Mick Sumbling	Universitat Autònoma de Barcelona, Spain
Roberta Farber	The British School of Pisa, Italy
Sabine Krauss	International Language School, Italy
Sonja Sentocnik	The National Education Institute, Slovenia
Svein Sirnes	Norsk Lektorlag, Norway
Teresa Nandin	Pompeu Fabra University Language Teaching Programme, Barcelona, Spain

## Appendix F

### First- and Second-Round Judgments for TSE, TWE, and TOEFL

**Table F1**

*Judgments for the Test of Spoken English*

	Round 1 judgments		Round 2 (final) judgments	
	B1	C1	B1	C1
P1	45	60	45	60
P2	46	59	45	60
P3	51	59	45	60
P4	42	52	45	55
P5	46	57	45	55
P6	46	53	45	55
P7	40	56	45	55
P8	46	55	45	55
P9	43	53	45	55
P10	44	53	45	55
P11	47	59	45	60
P12	49	59	45	55
P13	51	58	45	55
P14	47	57	45	55
P15	44	56	45	55
P16	42	57	45	55
P17	45	55	45	55
P18	50	60	45	55
P19	45	55	45	55
P20	49	59	50	60
Mean (truncated)	46	57	45	56
Median (truncated)	46	57	45	55
Standard deviation	2.94	2.58	1.12	2.22
Minimum	40	52	45	55
Maximum	41	60	50	60

*Note.* Panelist 21 was not present for the judgments on this test.

**Table F2*****Judgments for the Test of Written English***

	Round 1 judgments		Round 2 (final) judgments	
	B1	C1	B1	C1
P1	4.5	5.5	4.5	5.5
P2	4.5		4.5	
P3	4.5	6	4.5	6
P4	4	5.5	4	5.5
P5	4	6	4	5
P6	4	5.5	4.5	5.5
P7	4	6	4	6
P8	4.5	5.5	4.5	5.5
P9	4	5.5	4.5	5.5
P10	4.5	5.5	4.5	5.5
P11	4	6	4	6
P12	4.5	5.5	4.5	6
P13	4.5	5.5	4.5	5.5
P14	4	6	4	6
P15	5	6	5	6
P16	4	6	4	6
P17	4.5	5.5	4.5	5.5
P18	4	5.5	4	5.5
P19	4.5	5.5	4.5	5.5
P20	4	5.5	4.5	5.5
Mean (truncated)	4.3	5.7	4.4	5.7
Median (truncated)	4.3	5.5	4.5	5.5
Standard deviation	0.3	0.2	0.3	0.3
Minimum	4	5.5	4	5
Maximum	5	6	5	6

*Note.* Panelist 21 was not present for the judgments on this test. Panelist 2 did not make Level C1 judgments, as she did not believe that the test measured C1 proficiency.

**Table F3*****Judgments for TOEFL—Structure Section***

	Round 1 Judgments		Round 2 (final) judgments	
	B1	C1	B1	C1
P1	12	31	14	30
P2	13	32	14	31
P3	17	29	19	28
P4	25	32	19	31
P5	17	30	19	31
P6	22	31	21	31
P7	22	33	19	31
P8	18	32	18	31
P9	18	30	19	31
P10	23	32	22	31
P11	8	31	9	30
P12	19	33	19	31
P13	15	29	17	28
P14	22	31	22	30
P15	24	31	21	31
P16	22	33	21	33
P17	21	29	19	30
P18	25	33	19	31
P19	17	32	17	32
P20	19	28	19	28
P21	24	31	23	31
Mean raw score (truncated)	19	31	19	31
Median raw score (truncated)	19	31	19	31
Standard deviation	4.5	1.5	3.2	1.3
Minimum raw score	8	28	9	28
Maximum raw score	25	33	23	33
Mean scaled score	45	56	45	56
Median scaled score	45	56	45	56

*Note.* Two of the Structure items were deleted from the analyses, as they are not used in operational administrations. The panelists' judgments were adjusted appropriately to maintain the relationship between their initial (unadjusted) round-one and round-two judgments. The adjustment resulted in no differences between the unadjusted and adjusted round-two mean B1 and C1 cut scores, 19 and 31, respectively.



**Table F4*****Judgments for TOEFL—Reading Section***

	Round 1 judgments		Round 2 (final) judgments	
	B1	C1	B1	C1
P1	19	36	20	36
P2	25	38	22	38
P3	27	37	25	36
P4	16	37	20	37
P5	27	38	27	38
P6	28	35	23	35
P7	21	38	22	37
P8	30	38	30	38
P9	20	37	21	36
P10	23	36	22	36
P11	9	35	10	35
P12	9	36	20	35
P13	22	36	22	36
P14	31	38	31	38
P15	29	37	29	37
P16	22	37	23	37
P17	26	35	25	35
P18	22	36	23	36
P19	20	37	20	37
P20	23	33	23	33
P21	26	34	26	35
Mean raw score (truncated)	23	36	23	36
Median raw score (truncated)	23	37	23	36
Standard deviation	6	1.4	4.4	1.3
Minimum raw score	9	33	10	33
Maximum raw score	31	38	31	38
Mean scaled score	46	56	46	56
Median scaled score	46	57	46	56

**Table F5*****Judgments for TOEFL—Listening Section***

	Round 1 judgments		Round 2 (final) judgments	
	B1	C1	B1	C1
P1	18	40	18	40
P2	19	42	20	42
P3	20	40	23	41
P4	22	42	23	41
P5	24	42	23	41
P6	24	38	23	40
P7	20	42	22	42
P8	21	42	21	42
P9	27	42	27	42
P10	25	40	23	40
P11	10	39	10	39
P12	18	42	20	40
P13	26	41	23	41
P14	26	44	26	44
P15	28	41	28	41
P16	22	39	22	41
P17	23	36	23	38
P18	32	42	25	41
P19	23	43	23	43
P20	28	38	25	38
P21	27	38	25	40
Mean raw score (truncated)	23	41	23	41
Median raw score (truncated)	23	41	23	41
Standard deviation	4.8	2.0	3.7	1.5
Minimum raw score	10	36	10	38
Maximum raw score	32	44	28	44
Mean scaled score	46	56	46	56
Median scaled score	46	56	46	56

## Appendix G

### Panelists' Affiliations for Panel 2

#### *Panel 2 Standard-Setting Participants*

Name	Affiliation
Abdi Kazeroni	Université de Technologie de Compiègne, France
Brunella Casucci Belluomini	Language Data Bank, Italy
Hajdu Csaba	M-Prospect Language School, Hungary
Douglas Stevenson	Memori-X Language Lab, Budapest, Hungary
Volker Gehmlich	University of Applied Sciences at Osnabrueck, Germany
Gina Noonan	Institute of Technology, Carlow, Ireland
Charalambos Kollias	Business English Instructor, Greece
Isabelle Mangini-Nennot	Language Training Supervisor of EADS, France
Jan van Maele	Group T, Belgium
Hegedus Judit	International Business School, Hungary
Klaus Oelschlegel	Georg-Simon-Ohm Fachhochschule, Nürnberg, Germany
Lucia Katona	Institute for Foreign Languages, Hungary
Lynn Strebel	AKAD Language and Culture, Switzerland
Mary Petersen	Logik Sprachtraining, Germany
Maurice Cassidy	International House, UK
Roberta Farber	The British School of Pisa, Italy
Sue Luther	Georg-Simon-Ohm University of Applied Sciences, Nürnberg, Germany
Vera Dickman	Ecole Nationale Supérieure des Telecommunications, Paris, France
Wolfgang Rothfritz	University of Applied Sciences at Suedwestfallen, Germany
Zbigniew Szczepanczyk	Global Village, Kielce, Poland
Martin Musumeci	Academic Division, MABEC Support Unit, University of Malta

## Appendix H

### First- and Second-Round Judgments for TOEIC

**Table H1**

*Judgments for TOEIC—Listening Section*

	Round 1 judgments		Round 2 (final) judgments	
	B1	C1	B1	C1
P1	72	86	60	84
P2	46	79	60	80
P3	65	88	65	90
P4	60	87	63	87
P5	70	83	66	83
P6	60	85	61	85
P7	56	87	56	87
P8	53	76	53	76
P9	50	74	50	74
P10	47	81	50	81
P11	66	85	59	83
P12	68	87	68	87
P13	65	82	65	85
P14	69	83	59	83
P15	66	83	59	83
P16	53	82	50	80
P17	51	84	51	84
P18	53	81	54	81
P19	60	88	60	95
P20	42	88	46	88
P21	56	84	58	84
Mean raw score (truncated)	58	83	58	84
Median raw score (truncated)	60	84	59	84
Standard deviation	8.7	3.9	6.1	4.6
Minimum raw score	42	74	46	74
Maximum raw score	72	88	68	95
Mean scaled score	310	475	310	480
Median scaled score	325	480	320	480

**Table H2*****Judgments for TOEIC—Reading Section***

	Round 1 judgments		Round 2 (final) judgments	
	B1	C1	B1	C1
P1	64	82	56	82
P2	43	83	50	83
P3	55	88	55	90
P4	54	85	54	85
P5	62	83	56	83
P6	55	85	55	85
P7	60	86	58	85
P8	59	82	59	82
P9	52	78	52	78
P10	46	79	48	80
P11	55	78	55	78
P12	63	89	63	89
P13	64	81	60	80
P14	57	76	57	80
P15	66	82	55	80
P16	57	84	50	80
P17	48	85	50	85
P18	46	82	50	82
P19	65	87	60	85
P20	43	87	48	85
P21	54	79	56	80
Mean raw score (truncated)	56	83	55	83
Median raw score (truncated)	55	83	55	82
Standard deviation	7.1	3.5	4.2	3.3
Minimum raw score	43	76	48	78
Maximum raw score	66	89	63	90
Mean scaled score	245	400	240	400
Median scaled score	240	400	240	395