Research Memorandum
ETS RM–14-05

Mapping the *TOEFL® Primary™* Test Onto the Common European Framework of Reference

Patricia A. Baron

Spiros Papageorgiou

July 2014

ETS Research Memorandum Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Memorandum series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Memorandum series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

Mapping the *TOEFL*® *Primary*™ Test
Onto the Common European Framework of Reference

Patricia A. Baron and Spiros Papageorgiou

Educational Testing Service, Princeton, New Jersey

July 2014

Abstract

A standard-setting study was conducted to link scores on the *TOEFL*® *Primary*™ test to the Common European Framework of Reference (CEFR). The CEFR describes six levels of language proficiency organized into three bands: A1 and A2 (basic user), B1 and B2 (independent user), and C1 and C2 (proficient user). According to the Council of Europe, the CEFR is frequently used to provide a common basis for teaching languages across Europe, and the focus is on the knowledge and skills language learners need for language communication to act effectively. The TOEFL Primary test measures the English-language skills of learners aged 8 years or older who are developing communicative competence in English. The test consists of multiple-choice questions for the Reading and Listening section and constructed-response tasks for the Speaking section. Three levels of the CEFR, A1, A2, and B1, were the focus of this study. A variation of a yes–no Angoff standard-setting approach was applied to the Reading and Listening sections, and a variation of a performance profile approach was applied to the Speaking section. Eighteen educators from 15 countries served on the standard-setting panel.

Key words: CEFR, TOEFL Primary, standard setting, cut scores

Acknowledgments

A standard-setting study was conducted to link Reading, Listening, and Speaking scores on the *TOEFL® Primary™* test to the Common European Framework of Reference (CEFR). The CEFR describes six main levels of language proficiency organized into three bands: A1 and A2 (basic user), B1 and B2 (independent user), and C1 and C2 (proficient user). According to the Council of Europe (2001),

> The [CEFR] provides a common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, textbooks, etc. across Europe. It describes in a comprehensive way what language learners have to learn in order to use a language for communication and what knowledge and skills they have to develop so as to be able to act effectively. (p. 1)

Language test providers typically refer to the CEFR levels to help language test score users interpret the meaning of test scores because "the six main levels of the CEFR have become a common currency in language education" (Alderson, 2007, p. 660) and reference to its scales can "add meaning to the scores" (Kane, 2012, p. 8). The work reported herein is part of the ongoing research to add meaning to the scores of English-language tests designed by ETS following a standard-setting methodology (Baron & Tannenbaum, 2011; Tannenbaum & Baron, 2010; Tannenbaum & Wylie, 2008).

The outcomes of a standard-setting study are minimum test scores (cut scores) needed to reach defined performance levels. In the present case, the performance levels are specified by the CEFR descriptors. As noted in Tannenbaum and Baron (2013), one value of linking test scores to performance level descriptors is that the meaning of the scores becomes apparent—a test taker who achieves the minimum score to enter a performance level is expected to have the skills that define the level. This information may be used, for example, to support decisions about instruction, learning, or placement. In this regard, the outcomes of standard setting are directly associated with score interpretation and use; standard setting becomes an integral part of the overall validity argument (Bejar, Braun, & Tannenbaum, 2007; Kane, 2006). This memorandum describes the standard-setting study conducted to link TOEFL Primary test scores to the CEFR.

The TOEFL Primary Test

The TOEFL Primary Test is an English as a foreign language (EFL) assessment intended for young learners. The purpose of TOEFL Primary is to accurately measure the English-language skills of young learners aged 8 years and older who are developing communicative

competence in English. TOEFL Primary provides test users with a snapshot of young EFL learners' English skills in three areas: reading, listening, and speaking. TOEFL Primary is designed to support teaching and learning by providing meaningful feedback to teachers who use this information to guide their instruction. The test consists of multiple-choice questions for Reading and Listening and constructed-response tasks for Speaking. Results in the form of both numeric scaled scores and band levels are reported for each skill area. In this report, the three parts of the test assessing a specific skill area (Reading, Listening, or Speaking) will be referred to as *test sections*.

Reading Section

This paper-and-pencil section includes 36 items that are associated with a variety of reading stimuli (pictures or passages or both). Because TOEFL Primary is a test for lower proficiency young learners, the options of the Reading items may include one word only or longer sentences. Reading text types include correspondence, instructional, narrative, and short, age-appropriate academic passages. Test takers are given 30 minutes to complete the section. For the purpose of standard setting, an expanded 57-item version of the Reading section was used, based on field-tested items (see further discussion in the Method section of this report).

Listening Section

This paper-and-pencil section includes 36 questions associated with audio-recorded stimuli such as short utterances, social conversations, stories, and short, age-appropriate academic speeches and conversations. Each audio recording is played only once, and multiple-choice options are read to the student along with the question stems. Students' completion of the Listening section is paced by the recorded stimuli and is approximately 30 minutes in duration. As with Reading, an expanded 57-item form was used for standard setting (see further discussion in the Method section of this report).

Speaking Section

This section is delivered via computer or tablet and includes seven scored tasks and one unscored warm-up task. Tasks that elicit student expression of opinions and preferences, description of a scene, statement of a request, or questions are scored on a 3-point rubric. Tasks that require students to give a sequence of directions or narrate a sequence of events are scored on a 5-point rubric. Students view the task prompts in the context of an overarching narrative

scenario; the prompts require students to address characters in the scenario, thus giving students a clear purpose and audience for speaking. The Speaking section is paced by the scenario, and different task types have different response times, with the longest being 30 seconds.

The CEFR has been found to have some limitations as a basis for developing language tests (Weir, 2005), in particular tests for young learners, as we discuss later. For this reason, the TOEFL Primary was specifically designed neither to measure the range of proficiency levels addressed by the CEFR nor to describe language skills in the same way as expressed by the CEFR. However, because standard-setting panelists cannot develop cut scores on a test for levels or skills that are not included in the test design (Tannenbaum & Baron, 2013), ETS experts in assessment development and English-language learning (hereinafter referred to as the TOEFL Primary assessment team) evaluated the TOEFL Primary to identify the specific CEFR levels that were considered most clearly aligned with the TOEFL Primary Listening, Reading, and Speaking sections before conducting the standard-setting study. The experts consulted the *ELP Checklists for Young Learners* (cf. Council of Europe, n.d.) and identified three CEFR levels for Reading and Listening sections: A1, A2, and B1; four levels were identified for the Speaking section: CEFR levels A1, A2, B1, and B2. The process of standard setting focused only on those levels, which seem to be more relevant to the assessment of lower and mid-proficiency learners in primary education.

Method

The standard-setting task for the panelists was to recommend the minimum scores on each section of the test to reach each of the targeted CEFR levels. A variation of a yes–no Angoff standard-setting approach was applied to the Listening and Reading sections, which include multiple-choice items. The yes–no approach likely reduces the cognitive load placed on panelists by the more traditional modified Angoff approach (Impara & Plake, 1997) and has been applied in a variety of contexts, including to English-language testing in general (O'Neill, Buckendahl, Plake, & Taylor, 2007; Wendt & Woo, 2009) and, more specifically, to linking English-language test scores to the CEFR (Kantarcioğlu, Thomas, O'Dwyer, & O'Sullivan, 2010). A variation of a performance profile approach was applied to the Speaking section (Tannenbaum, 2010; Tannenbaum & Baron, 2010; Tannenbaum & Wylie, 2008; Zieky, Perie, & Livingston, 2008). The performance profile approach includes actual samples of test-taker responses, which provide direct evidence of their English-language skills. Details of how these

two standard-setting approaches were implemented are described in the procedures section. (Appendix A includes the agenda for the standard-setting study.)

Panelists

Eighteen educators from 15 countries served on the standard-setting panel. Ten educators reported being either full-time or part-time teachers of English as a second or foreign language; eight reported holding administrative positions. Two reported having less than 5 years of experience teaching students similar to the target age for TOEFL Primary (8+); seven panelists had between 5 and 10 years of experience; and nine panelists had more than 10 years of experience. One panelist reported having less than 5 years of experience teaching English as a second or foreign language; 5 panelists had between 5 and 10 years of experience; and 12 panelists had more than 10 years of experience. (See Appendix B for a list of the panelists and the countries they represent.)

Application of the CEFR to Tests for Young Learners: The Case of the TOEFL Primary Test

The extent to which the CEFR descriptors are appropriate for younger learners has been questioned in the literature, in particular regarding the assessment of higher language ability levels, as well as competences other than linguistic (Hasselgreen, 2005; Papageorgiou, 2010). Schneider and Lenz (2000, p. 42) advocated for reformulations of the descriptors to make them more comprehensible and better adapted to the experiences of young learners (see also Hasselgreen, 2012, p. 420). With these considerations in mind, the TOEFL Primary development team set out to adapt CEFR descriptors for the purposes of this study to make them more age appropriate and supportive of panelists' work during the standard-setting meeting, while retaining the core meaning of the text that distinguishes CEFR levels. The adaptation of the descriptors is also deemed to be aligned with the philosophy of the CEFR as a reference source. Users of the CEFR are encouraged to reflect on the content of the book and apply the relevant information to their own educational contexts (Council of Europe, 2001).

The TOEFL Primary assessment team initially located CEFR descriptors that seemed relevant to the performance of young learners and then revised the language of the descriptors to make it more age appropriate for young learners. Feedback on the adapted descriptors was collected from a group of external academic consultants with experience in language teaching and assessment of young learners. The finalized modified descriptors (see Appendix C) were

included in the material distributed to the standard-setting panelists on the first day of the meeting.

Premeeting Assignment: Familiarization With the CEFR and the TOEFL Primary Test

Prior to the standard-setting study, the panelists were asked to complete two preparatory activities. All panelists were asked to complete an assignment related to the CEFR and to take the TOEFL Primary test. The assignment was intended as part of a calibration of the panelists to reach a shared understanding of the minimum requirements for each of the targeted CEFR levels for Listening and Reading (A1, A2, and B1) and for Speaking (A1, A2, B1, and B2). They were also provided with tables of the modified CEFR descriptors, developed to address the young learners' tasks and skills that related to those measured by the test (see previous section).

The panelists were asked to review the modified CEFR descriptors and to write down what they believed students in the targeted age range should be able to do if they are at the *beginning* of the A1, A2, and B1 levels. This was done for Reading and Listening and at the beginning of A1, A2, B1, and B2 levels for Speaking. The panelists brought their completed assignments to the standard-setting meeting.

The day before the meeting, each panelist took all three sections of the TOEFL Primary test at a secure location on the ETS campus. The experience of taking the test is necessary for the panelists to understand the scope of what the test measures and the difficulty of the test questions and tasks. To better show panelists what actual test takers see when they take the operational test, Reading and Listening were delivered on paper, whereas Speaking was delivered on the computer. All panelists signed a nondisclosure and confidentiality form before accessing the test.

Standard-Setting Test Materials

The test sections used in the standard-setting meeting contained more test items than will be administered operationally to student test takers. The items used in the standard-setting meeting were TOEFL Primary test items previously administered in Listening and Reading field test sections. There were three reasons for collecting judgments on items in excess of the number in an operational form. The first reason relates to the fact that TOEFL Primary is operationally administered as two assessments, called Step 1 and Step 2, targeting two levels of language skills. Step 1 is targeted to assess students with lower level language skills than those assessed on Step 2. A longer standard-setting test form permitted items from both Step 1 and Step 2 to be included in the judgment process. The second reason for collecting judgments on items from

both TOEFL Primary Steps was to allow a scale representing performance across Step 1 and Step 2. This can be done by employing item response theory (IRT) scaling, which shows where test takers stand on the language ability continuum. Once the standard-setting process is complete, panel-recommended cut scores, based on the item pool, can be mapped to the reporting scale for each TOEFL Primary step through the IRT scaling process. The third reason relates to the expected consistency and reliability of cut scores. By including more items in the standard-setting process, panelists' judgments on items may be more consistent and the cut scores more reliable. This is similar to having more items contributing to a test score, which, as Gulliksen (1987) has shown, produces a more reliable estimate of knowledge or ability. However, it should be pointed out that for this to hold in standard setting, additional items would be needed near the location of the cut scores. In the current study, the additional items mostly targeted difficulty at the A2 level, which is anticipated to be an important cut score for score users.

Procedures

Standard Setting

Research in standard setting reinforces a number of core principles for best practice: careful selection of panel members and a sufficient number of panel members to represent varying perspectives; sufficient time devoted to developing a common understanding of the domain under consideration; adequate training of panel members; development of a description of each performance level; multiple rounds of judgments; and the inclusion of data, where appropriate, to inform judgments (Brandon, 2004; Council of Europe, 2009; Hambleton & Pitoniak, 2006, Tannenbaum & Katz, 2013). The approach used in this study adhered to all these guidelines.

Yes–no Angoff approach. A variation of a yes–no Angoff approach was applied to the Listening and Reading sections in the same manner as Tannenbaum and Baron (2013). The panelists judged 57 questions for each section, owing to the addition of items covering both TOEFL Primary Step 1 and Step 2 (see the Standard-Setting Test Materials section in this report). Each question was worth 1 point. As noted in the Method section of this report, the yes–no alternative is more straightforward and cognitively less challenging for panelists than the more traditional modified Angoff approach, which relies on estimating probabilities or percentages. The decision-making logic offered by the yes–no approach also supported a direct

estimation for each of the three targeted levels of the CEFR, A1, A2, and B1, with less panelist fatigue. For each multiple-choice question, a panelist judged if a student at the beginning of each CEFR level would know the correct answer—a yes-or-no decision. Because each level of the CEFR represents an increasing performance standard (a higher expectation), if a yes decision was made for the A1 level, then logic dictates that a yes must be made for the A2 and B1 levels. If a yes decision was made for the A2 level, then, again, a yes must be made for the B1 level. The same procedures were followed for Listening and Reading. Standard setting occurred first for the Listening section, and then was repeated for the Reading section. We describe the details of the yes–no approach next.

One of the first activities for the panelists was to discuss the test section. The goal was to have the panelists begin to think about and articulate their perceptions of the general difficulty of the tested content for students. The panelists were asked to identify and discuss content that most students aged 8–12 years (that is, the typical range for TOEFL Primary test takers) learning English as a foreign or second language (a) would find particularly challenging and (b) would not necessarily find challenging. Following this discussion, the panelists defined the minimum skills needed to reach each of the targeted CEFR levels (A1, A2, and B1). This was a continuation of the premeeting assignment. A student who has these minimally acceptable skills is referred to as a *just qualified candidate* (JQC). These JQC descriptions served as the frame of reference for the standard-setting judgments; that is, panelists were asked to consider the test questions in relation to these definitions. The first JQC to be defined was for the A2 level. We chose to start with this level because this level is anticipated to be important to score users; it is at this level that most of the test questions are targeted. The A2 JQC then served as an anchor point for defining the A1 and B1 levels.

Panelists worked in three small groups, with each independently defining the A2 level for the test section. Copies of the modified CEFR descriptors for young learners were provided to each panelist, and panelists were encouraged to refer to their premeeting assignment notes. A whole-panel discussion of the small-group definitions was facilitated and concluded with a consensus definition for the A2 JQC. Definitions of the JQCs for A1 and B1 were accomplished through whole-panel discussion, using the agreed-on A2 description as a starting point. (See Appendix D for the Listening and Reading JQC Descriptions.)

Panelists were trained in the variation of the yes–no Angoff standard-setting process and given an opportunity to practice making their judgments. At this point, panelists were asked to

sign a training evaluation form confirming their understanding and readiness to proceed, which all panelists did. Then they went through three rounds of operational judgments, with feedback and discussion between rounds. In Round 1, for each question, each panelist judged if a JQC would know the correct answer (yes or no). A judgment was made first for the A1 JQC, then for the A2 JQC, and then for the B1 JQC, before moving to the next test question. To calculate the cut scores, a yes was coded as 1 and a no as 0. The sum of each panelist's cross-question judgments represents his or her recommended cut score for that CEFR level. The results from Round 1 were projected to facilitate discussion.

Each panelist's recommended cut scores (A1, A2, and B1) were displayed by panelist ID number. Summary statistics of the panel's cut scores were also displayed: the mean, minimum, maximum, and standard deviation of judgments. The panelists engaged in a discussion of their reactions to these Round 1 results. Next, for each question, the number of panelists providing a yes judgment was displayed. Panelists were encouraged to identify questions for discussion, and questions for which *less than* 67% of the panelists reached agreement were flagged for panel discussion. In other words, if fewer than 12 panelists judged either yes or no for a question (at any of the three CEFR levels), it was flagged. Panelists were asked to share their decision rationales for the flagged questions. Panelists were encouraged during the discussion to refer to the JQC descriptions to support their rationales. After discussing the Round 1 judgments, panelists made Round 2 (final) judgments. Panelists were informed that they were not required to make changes to their Round 1 judgments.

One source of information for panelists to consider that is often included in standard-setting studies is the proportion of test takers who correctly answered the test questions (i.e., *p*-values). These data are typically presented after Round 1 judgments. We did not include *p*-values in this study. Research indicates that standard-setting panelists may be unduly influenced by *p*-value information, even when the *p*-value information is not accurate (see Clauser, Mee, Baldwin, Margolis, & Dillion, 2009; Tannenbaum & Baron, 2013). The question-level data available for the TOEFL Primary test were from a pilot-test administration ($N = 3{,}766$) of a set of items from the item pool, not an operational administration of a test form. Item difficulty information (IRT *b*-parameters) was calculated during item calibration based on data from this administration. Although we are confident that the item difficulty information accurately reflects how these pilot participants performed on these items, and would provide relative item difficulty, we had some reservations about the disconnect between the item difficulty values for the items in

the pool and the values that will exist when these items are administered in operational forms for two tests (Step 1 and Step 2). There are additional challenges when communicating the translation between the IRT-based statistics and the proportion-correct metric, as well as the translation from yes–no statistics to the item difficulty scale. Given that panelists may overcorrect their judgments to be more in line with item difficulty data, and with time limitations, we elected not to include the item-level data.

Performance profile approach. A variation of a performance profile approach was applied to the Speaking section. This section requires test takers to produce responses to offer direct evidence of their English-language proficiency. We describe the details of the approach next.

Consistent with the standard-setting process for the Listening and Reading sections, panelists first engaged in a discussion of the test section, articulating the content that would and would not likely challenge most students (aged 8–12 years) learning English as a foreign or second language. Following this discussion, the panelists defined the minimum skills needed to reach each of the targeted CEFR levels—the JQC descriptions for A1, A2, B1, and B2. The first JQC to be defined was for the A2 level. Panelists worked in three small groups, with each independently defining the A2 level for the Speaking test section. Copies of the modified CEFR descriptors for Speaking were provided to each panelist (see Appendix C), and panelists were encouraged to refer to their premeeting assignment notes. A whole-panel discussion of the small-group definitions was facilitated and concluded with a consensus definition for the A2 JQC. Definitions of the JQCs for A1, B1, and B2 were accomplished through whole-panel discussion, using the agreed-on A2 description as a starting point. (See Appendix D for the Speaking JQC descriptions.)

Panelists were trained in the variation of the performance profile standard-setting process and were given an opportunity to practice making their judgments. At this point, panelists were asked to sign a training evaluation form confirming their understanding and readiness to proceed, which all panelists did.

In this approach, panelists reviewed the tasks and corresponding scoring rubrics. They then listened to samples of student (test taker) responses to the tasks. A student's set of responses to the tasks formed a profile; the sum of the task scores is that student's total (section) score. The Speaking section includes seven tasks; four tasks were scored using a 3-point rubric, whereas three tasks were scored using a 5-point rubric. Profiles were sampled to represent the most

frequently occurring score patterns across the range of total scores. Forty-five Speaking section profiles were selected, ranging from a score of 3 to 27 total raw points. The profiles were presented in increasing score order, and a subset of the Speaking samples was broadcasted. Initially, three samples were presented to train the panelists with the standard-setting method. Then 12 more samples were broadcasted to illustrate performances that received a range of scores. Panelists asked for 11 additional Speaking samples to clarify performances that received adjacent scores, and these were broadcasted until all panelists were satisfied to make their cut score judgments. Each panelist was also provided with a printed sheet that showed the task scores and the total score for each profile to facilitate the judgment process.

Each panelist was asked to review the JQC descriptions for A1, A2, B1, and B2 for the Speaking section and then to review the performance profiles (evidence of students' proficiency). The standard-setting judgment was for each panelist to decide on the total score that an A1 JQC would most likely earn; that an A2 JQC would most likely earn; that a B1 JQC would most likely earn; and that a B2 JQC would most likely earn. Because each successive JQC represents a higher performance expectation, cut scores (total section scores) should increase as one advances from the A1 to A2 to B1 to B2 level.

Two rounds of judgments took place, with feedback and discussion between rounds, similar to that which occurred for the variation of the yes–no Angoff method applied to the Listening and Reading sections. After Round 1, each panelist's individual cut score recommendations were displayed, as was a summary of the panel's recommendations (mean, minimums and maximums, and standard deviations). Panelists shared their judgment rationales. Panelists had the opportunity to adjust their Round 1 judgments in Round 2.

Results

The first set of results summarizes the panel's standard-setting judgments by round for each of the test sections. The Round 2 mean cut scores are considered the panel's final recommendations. The operational cut scores are based on the Round 2 means, rounded to the next highest raw score. These results are followed by a summary of the panel's responses to a final evaluation survey.

Listening

Table 1 summarizes the results for the Listening section. The maximum raw score for Listening is 57 points. The panel's mean (average) cut score recommendations for levels A1 and

A2 increased, and for level B1 decreased slightly, at Round 2. The standard deviations tended to decrease across the rounds as expected, given the opportunity for discussion between the rounds. The standard error of judgment (SEJ) estimates the uncertainty in the panelists' judgments and is computed by dividing the standard deviation of the judgments by the square root of the number of panelists (Cizek & Bunch, 2007). Cohen, Kane, and Crooks (1999) suggested that an SEJ should be no more than half the value of the standard error of measurement (SEM) to reduce the impact on misclassification rates (false positives and false negatives). The SEM for the Listening section used in the standard-setting meeting was 2.8 raw points. The Round 2 recommended cut scores are are exactly half of the SEM (1.4) for level A2 and less than half for A1 and B1 for the Listening section.

The A1 cut scores for Listening ranged from 0 to 8; the minimum recommended cut score for A1 (0) was based on one panelist. This means that one panelist judged that the least able (just qualified) A1 learner will not be able to get any item correct of the 57 items in the Learning section and that most panelists found the test to be inaccessible to the student at the beginning of the A1 level. The B1 cut score ranged from 48 to 57. This means that one panelist judged that the least able B1 learner would get all presented questions correct. The pool of 57 items included items from both Step 1 and Step 2 TOEFL Primary tests.

Table 1

Listening: Standard-Setting Results

Statistic	Round 1			Round 2		
	A1	A2	B1	A1	A2	B1
Mean	3.5	31.5	53.4	3.6	31.8	53.3
Minimum	0.0	19.0	44.0	0.0	23.0	48.0
Maximum	8.0	47.0	57.0	8.0	44.0	57.0
SD	2.1	7.6	4.2	2.0	6.1	3.4
SEJ	0.5	1.8	1.0	0.5	1.4	0.8

Reading

Table 2 summarizes the results for the Reading section. The maximum raw score for Reading is 57 points. The panel's mean cut score recommendations for all three levels increased from Round 1 to Round 2. The standard deviations tended to decrease across the rounds. The

SEM for the Reading section used in the standard-setting study was 2.7 raw points. The Round 2 recommended cut scores were less than half of the SEM (1.35) for all three levels of the Reading section.

Table 2

Reading: Standard-Setting Results

Statistic	Round 1			Round 2		
	A1	A2	B1	A1	A2	B1
Mean	14.1	38.6	55.2	14.3	38.7	55.4
Minimum	6.0	23.0	52.0	7.0	25.0	52.0
Maximum	24.0	46.0	57.0	20.0	47.0	57.0
SD	4.9	5.1	1.9	3.7	4.5	1.7
SEJ	1.2	1.2	0.4	0.9	1.1	0.4

Speaking

Table 3 summarizes the results for the Speaking section. The maximum raw score for Speaking is 27 points. The panel's mean cut score recommendations for A1, A2, and B1 increased at Round 2 and remained the same for CEFR level B2. The standard deviations tended to decrease across the rounds. The SEM for the Speaking section is 2.2 raw points. All SEJs were less than half the SEM for the Speaking section.

Table 3

Speaking: Standard-Setting Results

Statistic	Round 1				Round 2			
	A1	A2	B1	B2	A1	A2	B1	B2
Mean	9.2	14.6	20.4	25.6	9.6	15.8	21.5	25.6
Minimum	7.0	12.0	18.0	22.0	7.0	12.0	18.0	23.0
Maximum	11.0	18.0	26.0	27.0	11.0	18.0	25.0	27.0
SD	1.2	2.0	2.3	1.5	1.0	1.5	1.8	1.1
SEJ	0.3	0.5	0.5	0.3	0.2	0.4	0.4	0.3

Final Evaluation Survey

Evidence addressing the validity of the standard-setting process (Kane, 1994) was collected from the final set of evaluation statements. Table 4 summarizes the panel's feedback regarding the general standard-setting process. The majority of panelists *strongly agreed* that the premeeting assignment was useful, that they understood the purpose of the standard-setting study, that the explanations and training provided were clear and adequate, that the opportunity for feedback and discussion between rounds was helpful, and that the standard-setting process was easy to follow. One panelist *disagreed* that the premeeting assignment was useful.

The panelists also were asked to indicate which of the following four factors most influenced their standard-setting judgments: the definition of the JQCs, the between-round discussions, the cut scores of the other panelists, and their own professional experience. The two most influential factors (very influential) were the definition of the JQC (16 panelists) and their own professional experience (12 panelists). Half of the panelists indicated that the between-round discussions were either very influential or somewhat influential, and 13 panelists indicated that the cut scores of other panelists were somewhat influential.

Table 4

Final Evaluations (in Percentage)

Statement	Strongly agree	Agree	Disagree	Strongly disagree
The premeeting assignment was useful preparation for the study.	78	17	6	0
I understood the purpose of this study.	100	0	0	0
The instructions and explanations provided by the facilitators were clear.	72	28	0	0
The training in the standard-setting methods was adequate to give me the information I needed to complete my assignment.	83	17	0	0
The explanation of how recommended cut scores are computed was clear.	67	33	0	0
The opportunity for feedback and discussion between rounds was helpful.	83	17	0	0
The process of making the standard-setting judgments was easy to follow.	67	33	0	0

Note. Number of panelists = 18.

Recall that the final cut scores are the Round 2 means, rounded to the next highest raw score. Table 5 provides final cut score recommendations for the three TOEFL Primary test sections in raw scores. Table 6 presents the panel's reported comfort level with these values. These reactions may be considered another source of validity evidence, addressing more specifically the outcomes of the standard-setting process. The majority of panelists reported being very comfortable with the cut scores—16 panelists indicated such for Listening and 15 panelists indicated such for Reading and Speaking. Two or three panelists were only somewhat comfortable with the cut scores for Listening and Reading, respectively; no panelists reported being *uncomfortable* with the Reading and Listening cut scores. Two panelists were only *somewhat comfortable* with the cut scores for Speaking, and one panelist indicated that he or she was somewhat uncomfortable. A few panelists added comments on the final evaluation, such as the following: "For Listening A2 and Reading A2, the difference between the maximum and minimum is quite big. I'd have felt more comfortable if these numbers had been closer."

Table 5

Final Recommended Cut Scores (Raw Points)

CEFR level	Listening (max. 57)	Reading (max. 57)	Speaking (max. 27)
A1	4	15	10
A2	32	39	16
B1	54	56	22
B2			26

Table 6

Comfort Level With Recommended Cut Scores (in Percentage)

Test section	Very comfortable	Somewhat comfortable	Somewhat uncomfortable	Very uncomfortable
Listening	89	11	0	0
Reading	83	17	0	0
Speaking	83	11	6	0

Note. Number of panelists = 18.

The results presented report the recommended cut scores in the raw score metric for each section, which is consistent with the process implemented; that is, panelists made judgments, and were given feedback on their judgments, in the raw score metric. In addition, panelists made judgments on more items than will appear on one form of the TOEFL Primary. However, we recognize that users of the TOEFL Primary will be interested in the CEFR cut scores as provided in the score report. Table 7 presents the final cut score recommendations in the scaled score metric included in the report for Step 1 and Step 2 of the Reading and Listening sections and for the Speaking section of the TOEFL Primary. Further information about the scale scores and the band levels that accompany and exemplify these scale scores can be found on the TOEFL Primary Web site (http://www.ets.org/toefl_primary).

Table 7

Final Recommended Cut Scores (Scaled Scores)

CEFR level	Listening (100–200)		Reading (100–200)		Speaking (0–27)
	Step 1	Step 2	Step 1	Step 2	
A1	102		102		10
A2	105	105	107	107	16
B1		113		114	22
B2					26

Conclusions

The purpose of this study was to link Listening, Reading, and Speaking scores on the TOEFL Primary test to the CEFR. Specifically, standard-setting procedures were used to identify the minimum test scores needed to reach the A1, A2, B1, and B2 levels of the CEFR. A variation of a yes–no Angoff was applied to the Listening and Reading sections, and a variation of a performance profile approach was applied to the Speaking section.

The responses to the final evaluation survey support the procedural validity of the standard-setting process. The majority of the panelists indicated, for example, that the standard-setting training had prepared them adequately to provide cut score judgments, that the provided instructions and explanations were clear, and that the process was easy to follow. Additional evidence of procedural validity was collected immediately following the specific training on the

standard-setting approaches. All panelists indicated that they were ready to proceed to make their first round of standard-setting judgments. In addition, at the conclusion of the study, the majority of panelists indicated that they were very comfortable with the recommended cut scores. Collectively, these results support the quality of the standard-setting implementation.

The SEJs, which estimated the uncertainty of the panelists' judgments, were relatively small in value, and all were close to or less than half the SEM for the respective test section. This indicates that the panelists were reasonably consistent in their judgments. However, due caution should be exercised when interpreting some of the cut score recommendations. The A1 cut score for Listening is 4 raw score points, which indicates that the Listening test may be inaccessible to students at the low end of A1. Similarly, but at the other end of the range, the B1 cut score for Listening was 54 raw points (out of 57), and for Reading, the B1 recommended cut score was 56 raw points (out of 57). This could be due to insufficient measurement power at the B1 level for Reading and Listening. Similarly, the B2 cut score for Speaking was 26 out of 27 raw points, which may indicate a lack of measurement power to make the distinctions between B1 and B2 for Speaking. An additional factor may be that the number of classification levels is high relative to the number of score points on these test sections.

As previously noted (Baron & Tannenbaum, 2011), decision makers, those who interpret and use cut score recommendations, are encouraged to evaluate them in relation to their specific needs. Cut scores are neither exact nor perfectly reliable values. They reflect the informed and reasoned decisions of the individuals who participate in the standard-setting process. In the present case, the cut scores offered as the final recommendations (Round 2) are an average of the 18 educators' recommendations. Decision makers may elect to adjust the recommendations, if doing so brings the cut scores into better alignment with their needs and circumstances (Cizek & Bunch, 2007; Geisinger & McCormick, 2010). Lowering a cut score—reducing the requirement to enter a CEFR level, for example—means more students will likely be classified into that level. This may be a legitimate goal. However, by lowering requirements, there is an increased risk that students who do not have the English-language skills expected at that level will be falsely classified at that level (a false positive decision). Conversely, raising a cut score makes it more difficult to enter a CEFR level. This increases the risk of falsely denying a student's classification into a higher CEFR level (a false negative decision). There is no "correct" decision to be made; there will always be some degree of misclassification that occurs. We suggest that decision makers first consider how they plan to use the CEFR levels and then review the cut

score recommendations from this study, and evaluate if the current recommendations support the intended decisions or if adjustments may be needed. One final note: If adjustments are made, decision makers should have reasonable rationales for the adjustments. Setting cut scores is analogous to forming a policy, which, as Kane (2001) noted, should be based on sound and supported argument (Baron & Tannenbaum, 2011).

References

Alderson, J. C. (2007). The CEFR and the need for more research. *Modern Language Journal, 91*(4), 659–663.

Baron, P. A., & Tannenbaum, R. J. (2011). *Mapping the TOEFL® Junior™ Test onto the common European Framework of Reference* (ETS Research Memorandum No. RM-11-07). Princeton, NJ: ETS.

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* (pp. 1–30). Maple Grove, MN: JAM Press.

Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education, 17*, 59–88.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* Thousand Oaks, CA: Sage.

Clauser, B. E., Mee, J., Baldwin, S. G., Margolis, M. J., & Dillon, G. F. (2009). Judge's use of examinee performance data in an Angoff standard-setting exercise for a medical licensing examination: An experimental study. *Journal of Educational Measurement, 46*, 390–407.

Cohen, A. S., Kane, M. T., & Crooks, T. J. (1999). A generalized examinee-centered method for setting standards on achievement tests. *Applied Measurement in Education, 12*, 343–366.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.

Council of Europe. (2009). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment—A manual.* Retrieved from http://www.coe.int/t/dg4/linguistic/manuel1_en.asp

Council of Europe. (n.d.). *ELP checklists for young learners: Some principles and proposals.* Retrieved from http://www.coe.int/t/dg4/education/elp/elp-reg/Source/Templates/ELP_Language_Biography_Checklists_for_young_learners_EN.pdf

Geisinger, K. F., & McCormick, C. A. (2010). Adopting cut scores: Post-standard-setting panel considerations for decision makers. *Educational Measurement: Issues and Practice, 29*, 38–44.

Gulliksen, H. (1987). *Theory of mental tests.* Hillsdale, NJ: Lawrence Erlbaum Associates.

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.

Hasselgreen, A. (2005). Assessing the language of young learners. *Language Testing, 22*(3), 337–354.

Hasselgreen, A. (2012). Adapting the CEFR for the classroom assessment of young learners' writing. *Canadian Modern Language Review, 69*(4), 415–435.

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 Associates.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: Praeger.

Kane, M. (2012). Validating score interpretations and uses. *Language Testing, 29*(1), 3–17.

Kantarcioğlu, E., Thomas, C., O'Dwyer, J., & O'Sullivan, B. (2010). Benchmarking a high-stakes proficiency exam: The COPE linking project. In W. Martyniuk (Ed.), *Studies in language testing: Aligning tests with the CEFR* (pp. 102–118). Cambridge, UK: Cambridge University Press.

O'Neill, T. R., Buckendahl, C. W., Plake, B. S., & Taylor, L. (2007). Recommending a nursing-specific passing standard for the IELTS examination. *Language Assessment Quarterly, 4*, 295–317.

Papageorgiou, S. (2010). Investigating the decision-making process of standard setting participants. *Language Testing, 27*(2), 261–282.

Schneider, G., & Lenz, P. (2000). *European Language Portfolio: Guide for developers*. Retrieved from http://www.coe.int/t/dg4/education/elp/elp-reg/Source/Publications/Developers_guide_EN.pdf

Tannenbaum, R. J. (2010). *Setting standards on the TOEIC® writing and speaking assessments for internationally trained nurses* (Research Memorandum No. RM 10-15). Princeton, NJ: Educational Testing Service.

Tannenbaum, R. J., & Baron, P. A. (2010). *Mapping TOEIC® test scores to the STANAG 6001 language proficiency levels* (Research Memorandum No. RM 10-11). Princeton, NJ: Educational Testing Service.

Tannenbaum, R.J., & Baron, P.A. (2013). *Mapping the TOEFL® Junior™ Comprehensive Test onto the Common European Framework of Reference*. Manuscript in preparation.

Tannenbaum, R. J., & Katz, I. R. (2013). Standard setting. In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology: Vol 3. Testing and assessment in school psychology and education* (pp. 455–477). Washington, DC: American Psychological Association.

Tannenbaum, R. J., & Wylie, E. C. (2008). *Linking English language test scores onto the Common European Framework of Reference: An application of standard setting methodology* (TOEFL iBT Series No. TOEFL ibt-06). Princeton, NJ: Educational Testing Service.

Weir, C. J. (2005). Limitations of the Common European Framework of Reference for Languages (CEFR) for developing comparable examinations and tests. *Language Testing, 22*(3), 281–300.

Wendt, A., & Woo, A. (2009). *A minimum English proficiency standard for the test of English as a Foreign Language Internet-based test (TOEFL iBT®)* (NCLEX Psychometric Research Brief No. 1-10). Chicago, IL: National Council of State Boards of Nursing.

Zieky, M. J., Perie, M., & Livingston, S. A. (2008). *Cutscores: A manual for setting standards of performance on educational and occupational tests.* Princeton, NJ: Educational Testing Service.

Appendix A

Standard-Setting Agenda

Day 1: Wednesday, July 10

Registration, receive materials

Welcome and overview

Listening: Review and discuss

Develop JQC definitions for CEFR levels A1, A2, and B1

Lunch

Training on standard-setting method; training evaluation

Round 1 judgments

Break

Round 1 discussion and Round 2 judgments

Reading Comprehension: Review and discuss

Adjourn for the day

Day 2: Thursday, July 11

Sign in and receive materials

Develop JQC definitions for reading comprehension A1, A2, and B1

Round 1 judgments

Lunch

Round 1 discussion and Round 2 judgments

Break

Speaking: Review and discuss

Small group: Develop JQC definitions for A1, A2, B1 and B2

Adjourn for the day

Day 3: Friday, July 12

Sign in and receive materials

Speaking: Finalize JQC definitions for A1, A2, B1, and B2

Training on standard-setting method; training evaluation

Round 1 judgments

Lunch

Round 1 discussion and Round 2 judgments

Complete final evaluations of standard-setting process

Content alignment survey

Appendix B

Panelists and Countries

Panelist	Country
Hidehiko John Amano	Japan
Alla Chekanova	Russia
Christina Ilve Dehman	Sweden
Vasiliki Dimitrakaki	Greece
Keenya Hofmaier	France
Thy Anh Le Huynh	Vietnam
Belyavskaya Lyubov Ivanovna	Kazakhstan
Diane Lafayette	France
Antonieta Heyden Megale	Brazil
Martha Patricia Concha Mendoza	Mexico
Miroslava Mišová	Slovakia
Nisreen S. Naji	Jordan
Rafael Ruiz Pleguezuelos	Spain
Maria Isabel Diaz Ruiz	Peru
Hoki Song	South Korea
Miyuki Takahashi	Japan
Ana Eduarda Toscano	Brazil
Alex Xiuming Zhang	China

Table C1

Modified CEFR Descriptors: Listening

CEFR level	Guiding statement and associated can-do statements
A1	Can recognize familiar words and very basic phrases concerning oneself, family, and immediate concrete surroundings when speech is very slow and carefully articulated, with long pauses to assimilate meaning. *Students at this level can:* Recognize short, simple directions in school and at home (e.g., stand up, close the book) Understand familiar everyday expressions and very basic, formulaic phrases
A2	Can understand short, simple messages and conversations when clearly articulated. *Students at this level can:* Catch the main point in short, clear, simple messages, announcements, and instructions Generally identify the topic of discussion around oneself, when conducted clearly Derive the idea of the overall meaning of short texts and utterances on everyday topics of a concrete type despite not knowing all of the words
B1	Can understand straightforward factual information about common everyday and school-related topics, identifying both general messages and specific details, provided speech is clearly articulated. *Students at this level can:* Follow straightforward short talks on familiar topics provided these are delivered in clearly articulated speech Understand multistep directions that may be embedded in longer text Extrapolate the meaning of occasional unknown words from accessible contexts Generally follow the main points of extended discussion around him or her, provided speech is clearly articulated Understand straightforward factual information about common everyday topics, identifying both general messages and specific details, provided speech is clearly articulated Understand the main points of straightforward, short talks and narratives on familiar matters regularly encountered in school and at home

Table C2

Modified CEFR Descriptors: Reading

CEFR level	Guiding statement and associated can-do statements
A1	Can understand very short, simple texts, perhaps a single word or phrase at a time, by picking up familiar names, words, and basic phrases and rereading as required. *Students at this level can:* Understand the general idea of short, simple informational texts, perhaps as much as a few sentences long, and descriptions of familiar, immediate topics, and are greatly assisted by pictures that help to explain the text Understand some words, labels, and phrases in simple texts that are found in familiar settings and daily life Follow written instructions expressed in short, imperative phrases and sentences, especially if pictures are included to aid comprehension Understand words, phrases, and simple sentences, especially those that contain common and courteous everyday expressions, in simple friendly messages, such as letters and e-mails
A2	Can understand short, simple texts on familiar matters of a concrete type that consist of high-frequency everyday language used in familiar contexts such as at home and in school. Can understand short, simple texts written in sentence form that contain the high-frequency vocabulary. *Students at this level can:* Understand sentences containing the most common words, including some shared international words Understand short, simple texts written in common everyday language Find specific information presented in simple everyday material such as advertisements, brochures, menus, and schedules and on signs and notices in public places such as streets, restaurants, railway stations, and school Understand simple instructions for everyday situations, such as school assignments and tasks at home Understand short, simple personal correspondence, such as letters and e-mails Identify specific information in simple written material such as letters, stories, and textbooks
B1	Can understand straightforward texts, controlled for language construction and accessibility, on familiar topics. *Students at this level can:* Recognize significant points in straightforward informational articles on school subjects Find and understand general information in everyday material, such as friendly correspondence Understand the description of events, feelings, and wishes in personal letters and stories well enough to correspond about them with a friend or acquaintance Search texts written in sentences to locate specific information Understand and connect important ideas and details, such as cause-effect relationships, about straightforward instructions for everyday situations, such as school assignments and tasks at home Identify the main ideas in texts expressing an opinion

Table C3

Modified CEFR Descriptors: Speaking

CEFR level	Guiding statement and associated can-do statements
A1	Can use a few words and formulaic expressions, typically speaking with a lot of repetition, pauses, and hesitation at a slower rate of speech. *Students at this level can:* Say a few simple words or phrases to describe people and places Ask and answer simple questions related to immediate need in familiar settings (e.g., asking for the directions to a bathroom) Use basic greetings and other polite expressions (e.g., hello, good-bye, see you tomorrow, thank you)
A2	Can participate in simple, routine conversations requiring listener effort for comprehension. *Students at this level can:* Give short, simple descriptions of familiar and concrete topics, typically using short phrases and simple sentences (e.g., describing a picture or a favorite sport) Describe what they do and did in everyday routine situations Have a short, simple conversation on familiar topics
B1	Can participate in short conversations on familiar topics on less concrete topics (e.g., a movie they watched) in less routine situations without difficulty most of the time. *Students at this level can:* Speak fluently about their daily routines, their families, and familiar objects and things Retell a story or event with sufficient detail Provide reasons or opinions about events that they experienced (e.g., a story that he or she read, a sporting event in which he or she participated) using simple language
B2	Can participate in conversations on a wide range of familiar and unfamiliar topics. B2 speakers can do things that B1 speakers do with much improved accuracy and fluency. B2 speakers' speech generally poses little strain for the listener. *Students at this level can:* Participate in conversations on familiar and unfamiliar topics with native speakers Participate in classroom discussions with ease Compare and contrast competing ideas Elaborate reasons and opinions

Appendix D

Just Qualified Candidate Descriptions

Listening

JQC A1

1. Can recognize basic instructions and simple phrases with visual support.
2. Can understand familiar expressions, provided speech is very slowly and clearly articulated.

JQC A2

1. Can understand two- to three-step instructions and directions with visual support.
2. Can identify the overall subject matter of short texts and utterances on familiar topics of concrete type without knowing all the words.
3. Can identify appropriate responses to short talks on familiar topics.
4. Can understand the main idea of messages, including unknown or unexpected vocabulary concerning familiar topics.
5. Can understand standard language clearly articulated.

JQC B1

1. Can understand multistep instructions without visual support.
2. Can understand the overall topic and supporting ideas in longer texts.
3. Can understand short talks and extended conversations on familiar topics when language is clearly articulated.
4. Can understand topics pertaining to culture (e.g., geography), hobbies, and topics related to academic and personal interests.
5. Can infer meaning of a wider range of vocabulary, including synonyms and unfamiliar words in accessible contexts.

Reading

JQC A1

1. Can understand simple words and phrases.
2. Can understand the meaning of short, simple sentences in short texts (e.g., e-mails, letters) on familiar topics by recognizing isolated words.
3. Can understand basic instructions and descriptions with visual aids.

JQC A2

1. Can understand short, simple texts with short, simple sentences containing high-frequency, familiar words about home and school.
2. Can identify the main idea in short, simple, clear passages (e.g., e-mail, stories, textbooks) regarding everyday life.
3. Can identify specific simple information presented in short texts such as ads, brochures, letters, e-mails, signs, schedules, and instructions.
4. Can understand a wider range of words, phrases, and short, simple sentences in everyday expressions.
5. Can understand the meaning of the most common words.

JQC B1

1. Can understand the main idea of longer texts, on a wider variety of topics (feelings, opinions, events, wishes), on subjects related to his or her interests.
2. Can understand a series of short sentences with linking words.
3. Can locate specific details in longer texts.
4. Can infer meaning of less familiar words.

Speaking

JQC A1

1. Can introduce and describe himself or herself using simple sentences and basic greeting protocols.
2. Can give simple information about personal, basic, routine, familiar topics.
3. Can name familiar and concrete objects using single words and a limited repertoire of expressions.

4. Can ask simple questions with prompting from or help of the interlocutor.

5. Has problems with pronunciation.

JQC A2

1. Can ask simple, short questions on familiar topics and answer appropriately questions of the same nature.

2. Can handle short social exchanges with the help of the other person involved in the communication.

3. Can give short, simple descriptions of familiar and concrete objects (pictures, places, people, and activities) typically using short phrases with minor errors in grammar.

4. Speech is generally intelligible, but the delivery may be slow and hesitant.

5. Can deliver a short prepared talk on a familiar subject.

JQC B1

1. Can mostly participate in short conversations on less concrete topics related to his or her own experience, with minor errors.

2. Can retell a story or event with enough detail for the listener to get the message.

3. Can express opinions and feelings on familiar topics using simple language with some minor errors that do not affect comprehension.

4. Can give a short talk on a familiar subject with some lapses in pronunciation, some hesitation, and inaccuracy, which do not impede comprehension.

JQC B2

1. Demonstrates a good range of appropriate vocabulary and degree of control of some complex grammatical forms with some interferences and inaccuracies.

2. Participates in an extended exchange on less familiar topics, with some inaccuracies, but generally errors do not interfere with communication; can provide clarification when necessary.