Measuring the Power of Learning.™

Research Memorandum
ETS RM–15-13

Mapping Scores From the *TOEFL Junior*® Comprehensive Test Onto the Common European Framework of Reference (CEFR)

Richard J. Tannenbaum

Patricia A. Baron

December 2015

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 Scores From the TOEFL Junior® Comprehensive Test Onto the Common European Framework of Reference (CEFR)

Richard J. Tannenbaum and Patricia A. Baron
Educational Testing Service, Princeton, New Jersey

December 2015

Corresponding author: R. J. Tannenbaum, E-mail: rtannenbaum@ets.org

Suggested citation: Tannenbaum, R. J., & Baron, P., A. (2015). *Mapping scores from the TOEFL Junior® Comprehensive test onto the Common European Framework of Reference (CEFR)* (Research Memorandum No. RM-15-13). Princeton, NJ: Educational Testing Service.

Action Editor: Don Powers

Reviewers: Spiros Papageorgiou and Priya Kannan

Abstract

A standard-setting study was conducted to link scores on the *TOEFL Junior*® Comprehensive test of listening, reading, writing, and speaking to the Common European Framework of Reference (CEFR) levels. The CEFR describes 6 levels of language proficiency organized into 3 bands: A1 and A2 (basic user), B1 and B2 (independent user), and C1 and C2 (proficient user). "The [CEFR] provides a common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, textbooks, etc. across Europe. It describes . . . 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" (Council of Europe, 2001, p. 1). The TOEFL Junior comprehensive test measures academic and social English-language skills of middle-school students learning English as a second or foreign language (ESL/EFL). The test consists of multiple-choice questions for reading and listening, and constructed-response tasks for writing and speaking. In this study, we focus on 3 levels of the CEFR (A2, B1, and B2). 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 writing and speaking sections. A total of 18 educators from 15 countries served on the standard-setting panel. The results of the study are minimum scores for each of the 4 test sections that are recommended for classifying test takers according to the levels of the CEFR.

Key words: CEFR, *TOEFL Junior*®, standard setting, cut scores

A standard-setting study was conducted to link scores on the *TOEFL Junior®* Comprehensive test of reading, listening, writing, and speaking to the Common European Framework of Reference (CEFR) levels. The CEFR describes six levels of language proficiency organized into three bands: A1 and A2 (basic user), B1 and B2 (independent user), C1 and C2 (proficient user). "The [CEFR] provides a common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, textbooks, etc. across Europe. It describes . . . 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" (Council of Europe, 2001, p. 1).

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. 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 that 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; Papageorgiou & Tannenbaum, in press). This report describes the standard-setting study conducted to link the TOEFL Junior Comprehensive test scores to the CEFR.

The TOEFL Junior Comprehensive Test

The TOEFL Junior Comprehensive test is relevant to those students ages 11+ who are learning English as a second or foreign language (ESL/EFL). The test measures the academic and social English-language skills that are typical of instructional environments (https://www.ets.org/toefl_junior/about). The test consists of multiple-choice questions for reading and listening and constructed-response tasks for writing and speaking. Four scores are reported, one for each section.

Listening Section

This section includes 28 operational questions. The listening prompts include single speakers, short conversations, and academic listening (i.e., listening to a presentation). Students

(test takers) are given 36 minutes to complete the section and may take notes while they listen to the prompts. Each prompt is played only once.

Reading Section

This section includes 28 operational questions. The types of reading prompts include correspondence, nonlinear information, fiction, and journalistic and expository texts. Students (test takers) are given 41 minutes to complete the section.

Writing Section

This section includes four tasks scored using 4-point rubrics. The first task requires students (test takers) to edit a short text. A second task requires students to respond to questions in the form of an e-mail. A third task requires students to express an opinion about a given topic. The fourth task requires students to listen to a presentation and to write an explanation of the presentation. Students may take notes, but the listening prompt is played only once. The maximum number of raw points for this section is 16. Students are given 39 minutes to complete the section (32 minutes are allocated for active writing.)

Speaking Section

This section includes four tasks scored using 4-point rubrics. The first task requires students (test takers) to read a passage out loud. The second task requires students to narrate a pictured sequence of events. In the third and fourth tasks, students listen to prompts. In the third task, students listen to a short discussion and then must describe the main points. In the fourth task, students listen to a presentation and must explain the content of the presentation. Students may take notes, but the listening prompts are played only once. The maximum number of raw points for this section is 16. Students are given 18 minutes to complete the section (4 minutes are allocated for active speaking).

The Common European Framework of Reference (CEFR)

The CEFR is, by purpose and design, a general framework that is neither language specific nor prescriptive. As noted by North (one of the chief "architects" of the CEFR), Martyniuk, and Panthier (2010), "The CEFR is a concertina-like reference tool that . . . educational professionals can merge or sub-divide, elaborate or summarize, adopt or adapt according to their context" (p. 4). The flexible nature of the CEFR means that there is no one

single correct interpretation that may be applied to it; North (2014) reinforced this point: "There is not and never will be an authorised interpretation of the CEFR. That openness is the secret of its success" (p. 5). The accessibility of the CEFR also means, however, that it is not a blueprint for assessment design or development. It is not surprising, therefore, that the TOEFL Junior Comprehensive test is not necessarily a point-by-point reflection of the English-language skills and expectations offered by the CEFR. This is not a limitation of the test, but it does mean that evidence is needed regarding where, for which levels of the CEFR, the test is considered adequately aligned before engaging in a standard-setting process to map test scores to the CEFR (Council of Europe, 2009; Tannenbaum & Cho, 2014). Therefore, before conducting the standard-setting study, Educational Testing Service (ETS) language assessment experts identified the specific CEFR levels that were considered most closely aligned with the TOEFL Junior Comprehensive listening, reading, writing, and speaking sections. Each section was judged to address CEFR levels A2, B1, and B2. Therefore, the process of standard setting focused only on those levels.

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 (A2, B1, and B2). A Yes/No Angoff standard-setting approach was applied to the listening and reading sections, which include multiple-choice items. The Yes/No Angoff 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 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 (Baron & Papageorgiou, 2014; Kantarcioğlu, Thomas, O'Dwyer, & O'Sullivan, 2010). A Performance Profile approach was applied to the writing and speaking sections (Baron & Papageorgiou, 2014; Tannenbaum, 2010; Tannenbaum & Baron, 2010; Tannenbaum & Cho, 2014; Tannenbaum & Wylie, 2008; Zieky, Perie, & Livingston, 2008). The Performance Profile approach includes actual samples of student (test taker) responses—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. Fifteen educators reported being either full-time or part-time ESL/EFL teachers; three reported holding administrative positions. Thirteen reported having at least 5 years of experience teaching students within the age range of interest (11+ years); 17 reported having at least 5 years of experience teaching ESL/EFL. (See Appendix B for a list of the panelists and the countries they represent.)

Premeeting Assignment: Familiarization With the Common European Framework of Reference (CEFR) and the TOEFL Junior Comprehensive 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 Junior Comprehensive test. The assignment was intended as part of a calibration of the panelists to a shared understanding of the minimum requirements for each of the targeted CEFR levels (A2, B1, and B2) for listening, reading, writing, and speaking. They were provided with tables from the CEFR that related to the specific skills measured by the test. For example, the writing section addresses skills relevant to writing essays and reports but does not address creative writing skills; therefore, the CEFR table addressing reports and essays was included in the review, but the table addressing creative writing was not. The educators were asked to review the selected CEFR tables 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 A2, B1, and B2 levels. This review was done for each of the four skills (listening, reading, writing, and speaking). However, because the CEFR is not age specific, and due to some concerns about the application of the descriptors to the experiences of young learners (Hasselgreen, 2012; Schneider & Lenz, 2000), the educators were advised to modify the extant language of the CEFR descriptors, if needed, to ensure proper alignment with the age range covered by the TOEFL Junior Comprehensive test. The educators brought their completed assignment to the standard-setting study.

The day before the study, each educator took all four sections of the TOEFL Junior Comprehensive test at an authorized test center. The test was delivered online, in the same way it is presented to actual test takers. (All educators had signed a nondisclosure/confidentiality form before having access to the test.) The experience of taking the test is necessary for the educators

to understand the scope of what the test measures and the difficulty of the questions and tasks on the test.

Procedures

Standard Setting

Recent reviews of research on standard-setting approaches reinforce 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 develop 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; Hambleton & Pitoniak, 2006; Tannenbaum & Cho, 2014). The approaches used in this study adhere to all of these guidelines.

Yes/No Angoff approach. A Yes/No Angoff approach (Impara & Plake, 1997) was applied to the listening and reading sections. Each section includes 28 operational questions that are worth one point each. As noted above, 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 Angoff approach also supported a direct estimation for each of the three targeted levels of the CEFR (A2, B1, and B2) without burdening or fatiguing panelists. 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 A2 level, then logic dictates that a "yes" decision must be made for the B1 and B2 levels. If a "yes" decision was first made for the B1 level, then, again, a "yes" decision must be made for the B2 level. The same procedures were followed for listening and reading. Standard setting occurred first for the listening section and was repeated for the reading section. We describe the details of the Yes/No Angoff 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 perception of the general difficulty of the tested content for students. The panelists were asked to identify and discuss content that most students (ages 11+) learning ESL/EFL (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 (A2, B1, and B2). This was a continuation of the premeeting assignment. A student (test taker) 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 B1 level. We chose to start with this level as it represents the entrance into the independent user band of the CEFR and is well defined by the CEFR tables. Once defined, this level then serves as an anchor point for defining the A2 and B2 levels.

Panelists worked in three small groups, with each group independently defining the B1 level for the test section. Copies of the applicable CEFR tables were provided to each panelist; and panelists were encouraged to refer to their premeeting assignment notes. Panelists also had access to JQC definitions developed from a previous study that mapped the TOEFL Junior Standard test to the CEFR (Baron & Tannenbaum, 2011). The previous version measured listening, reading, and language form and meaning skills. A whole-panel discussion of the small-group definitions was facilitated and concluded with a consensus definition for the B1 JQC. Definitions of the JQCs for A2 and B2 were accomplished through whole-panel discussion, using the agreed upon B1 description as a starting point. (See Appendix C 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 A2 JQC, then for the B1 JQC, and then for the B2 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 represented 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 (A2, B1, and B2) were displayed by panelist ID number. Summary statistics of the panel's cut scores were also displayed: the panel average

(mean), minimum and maximum, and the 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. Questions for which less than 67% of the panelists reached agreement were flagged for further 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. Panels were asked to share their decision rationales for the flagged questions and any additional questions they wanted to discuss. Panelists were encouraged during the discussion to refer to the JQC descriptions to support their rationales. Questions were discussed in sets. After discussing the Round 1 judgments for a set, panelists made Round 2 judgments for that set. Panelists were informed that they were not required to make changes to their Round 1 judgments but had the option of making changes, if they desired.

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 (Clauser, Mee, Baldwin, Margolis, & Dillon, 2009). The question-level data available for the TOEFL Junior Comprehensive test was from a pilot-test administration ($N = 497$), not an operational administration; and more than half of the test takers were from one country. While we are confident that the *p*-value information accurately reflects how these pilot participants performed, we had some reservations about the generalizability of the *p*-value data to a more representative cohort of students (test takers) that take the test under operational conditions. Given that panelists may over correct their judgments to be more in-line with *p*-value data, we elected not to include this kind of question-level data.

The same type of feedback presented at the conclusion of Round 1 was presented at the conclusion of Round 2. In addition, the panelists were provided with impact data. Impact data reflect the percentage of pilot participants who would be classified into each of the three CEFR levels if the Round 2 average cut scores were applied to their test scores. While we chose not to provide *p*-value data, we had less concern about the use of the pilot scores in the aggregate. We believed that the panelists needed some feedback about the potential consequences of their cut-score recommendations (Reckase, 2001). Nonetheless, the panelists were informed that the

scores were based on a pilot administration and that the classification results were not definitive. The panelists engaged in a discussion of the Round 2 results and classification estimates and then made a final set of judgments (Round 3).

In Round 3, panelists were asked to make holistic cut-score decisions for the overall test section. Specifically, panelists were asked to review the JQC definitions for A2, B1, and B2 CEFR levels and to adjust their individual Round 2 A2, B1, and B2 cut scores. As before, panelists were not required to make adjustments.

Performance Profile approach. A Performance Profile approach (Zieky et al., 2008) was applied to the writing and speaking sections. These sections require students (test takers) to produce responses and so offer direct evidence of a student's English language proficiency. The same procedures were followed for writing and speaking. Standard setting occurred first for the writing section and was repeated for the speaking section. 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 what would and would not likely challenge most students (ages 11+) learning ESL/EFL. Following this discussion, the panelists defined the minimum skills needed to reach each of the targeted CEFR levels—the JQC descriptions for A2, B1, and B2. The first JQC to be defined was for the B1 level. Panelists worked in three small groups, with each group independently defining the B1 level for the test section. Copies of the applicable CEFR tables 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 B1 JQC. Definitions of the JQCs for A2 and B2 were accomplished through whole-panel discussion, using the agreed upon B1 description as a starting point. (See Appendix C for the writing and speaking JQC descriptions.)

Panelists were trained in the variation of the Performance Profile 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.[1] Then they went through three rounds of operational judgments, with feedback and discussion between rounds.

In this approach, panelists reviewed the tasks and corresponding scoring rubrics. They then reviewed 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. For example, the writing section includes four tasks; a student's response to each task is a performance profile. Profiles were sampled from pilot test responses to represent the most frequently occurring score patterns across the range of total scores. Twenty-four writing section profiles were selected, ranging from a score of 4 to 15.5 total raw points. (No student in the pilot-test administration earned all 16 points; half-points were possible due to the averaging of the scores on the first two writing tasks.) Thirty-four speaking section profiles (audio files) were selected, ranging from a score of 4 to 16 total raw points.[2] The profiles were presented in increasing score order. The writing samples were provided in a binder, and the speaking samples were played aloud. Each panelist was also provided with a printed sheet (one for the writing section and one for the speaking section) that showed the task scores and the total score for each profile to facilitate their judgment process.

Each panelist was asked to review the JQC descriptions for A2, B1, and B2 for the particular test 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 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 A2 to B1 to B2 levels.

Three rounds of judgments took place, with feedback and discussion between rounds, similar to that which occurred for 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 average recommendations, minimums and maximums, and standard deviations. Panelists shared their judgment rationales. Panelists had the opportunity to adjust their Round 1 judgments in Round 2. The Round 2 feedback included the percentage of pilot participants who would be classified into each of the three CEFR levels if the Round 2 average cut scores were applied to their test scores. The panelists engaged in a discussion of the Round 2 results and classification estimates and then made a final set of judgments (Round 3).

Results

The first set of results summarizes the panel's standard-setting judgments by round for each of the test sections. The Round 3 average cut scores are considered the panel's final recommendations. The operational cut scores are based on the Round 3 averages, rounded to the next highest raw score. Because an operational cut score is intended to reflect the minimum acceptable score to enter a CEFR level, averages are rounded up. For example, suppose the recommended B1 cut score for the listening test was 20.3, that indicates, on average, that panelists believed more than 20 points were needed to enter the B1 level; in this instance, therefore, the appropriate cut score would be 21. 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 28 points. The panel's average cut-score recommendation for A2 increased at Round 2 and was relatively unchanged at Round 3. The average cut scores for B1 and B2 were consistent across the three rounds. 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 by the square root of the number of panelists (Cizek & Bunch, 2007). In order to reduce the impact on misclassification rates (false positives and false negatives), Cohen, Kane, and Crooks (1999) suggested that an SEJ should be no more than half the value of the standard error of measurement (SEM). The SEM for the listening section was 2.1 raw points. All SEJs were less than half the SEM for the listening section.

Table 1. Listening: Standard-Setting Results

	Round 1			Round 2			Round 3		
Levels	A2	B1	B2	A2	B1	B2	A2	B1	B2
Mean	9.1	20.6	26.9	10.2	20.8	26.8	10.9	20.3	26.3
Minimum	4.0	11.0	23.0	7.0	13.0	23.0	8.0	15.0	24.0
Maximum	14.0	27.0	28.0	15.0	25.0	28.0	15.0	24.0	28.0
SD	3.1	3.9	1.6	2.9	3.4	1.7	2.1	2.1	1.3
SEJ	0.7	0.9	0.4	0.7	0.8	0.4	0.5	0.5	0.3

Note. SEJ = standard error of judgment.

Reading

Table 2 summarizes the result for the reading section. The maximum raw score for reading is 28 points. The panel's average cut-score recommendation for A2 was consistent across the three rounds. The average cut score for B1 increased at Round 2 but then decreased at Round 3. The average cut score for B2 was consistent across Rounds 1 and 2 but decreased at Round 3. The standard deviations tended to decrease across the rounds. The SEM for the reading section was 2.2 raw points. All SEJs were less than half the SEM for the reading section.

Table 2. Reading: Standard-Setting Results

	Round 1			Round 2			Round 3		
Levels	A2	B1	B2	A2	B1	B2	A2	B1	B2
Mean	8.6	22.7	27.2	8.4	23.2	27.2	8.6	21.5	26.5
Minimum	5.0	16.0	21.0	4.0	16.0	21.0	6.0	16.0	21.0
Maximum	16.0	28.0	28.0	13.0	28.0	28.0	13.0	26.0	28.0
SD	3.2	3.1	1.7	2.7	3.0	1.7	2.1	2.5	1.8
SEJ	0.8	0.7	0.4	0.6	0.7	0.4	0.5	0.6	0.4

Note. SEJ = standard error of judgment.

Writing

Table 3 summarizes the result for the writing section. The maximum raw score for writing is 16 points. The panel's average cut-score recommendations for A2, B1, and B2 were consistent across the three rounds. The standard deviations decreased in Round 2 and remained unchanged in Round 3. The SEM for the writing section was 1.5 raw points. All SEJs were less than half the SEM for the writing section.

Table 3. Writing: Standard-Setting Results

	Round 1			Round 2			Round 3		
Levels	A2	B1	B2	A2	B1	B2	A2	B1	B2
Mean	5.9	9.4	12.8	5.9	9.4	12.8	5.9	9.4	12.7
Minimum	4.0	7.0	11.0	5.0	8.5	12.0	5.0	8.5	11.0
Maximum	7.0	11.0	14.0	7.0	11.0	14.0	6.5	11.0	14.0
SD	0.6	1.0	0.9	0.5	0.8	0.7	0.5	0.8	0.8
SEJ	0.2	0.3	0.2	0.1	0.2	0.2	0.1	0.2	0.2

Note. SEJ = standard error of judgment. One panelist did not participate in the standard-setting for writing due to illness.

Speaking

Table 4 summarizes the result for the speaking section. The maximum raw score for speaking is 16 points. The panel's average cut-score recommendation for A2 increased at Round 2 and remained the same at Round 3. The average cut scores for B1 and B2 were consistent across the three rounds. The standard deviations tended to decrease across the rounds. The SEM for the speaking section was 1.3 raw points. All SEJs were less than half the SEM for the speaking section.

Table 4. Speaking: Standard-Setting Results

	Round 1			Round 2			Round 3		
Levels	A2	B1	B2	A2	B1	B2	A2	B1	B2
Mean	6.8	10.0	13.1	7.3	10.3	13.2	7.3	10.3	13.3
Minimum	5.0	8.0	11.0	6.0	9.0	12.0	6.0	9.0	12.0
Maximum	8.0	11.0	15.0	8.0	12.0	15.0	8.0	12.0	15.0
SD	1.0	0.8	1.1	0.6	0.8	1.0	0.6	0.8	0.9
SEJ	0.2	0.2	0.3	0.1	0.2	0.2	0.1	0.2	0.2

Note. SEJ = standard error of judgment.

Final Evaluation Survey

Evidence addressing the validity of the standard-setting process (Kane, 1994) was collected from the final set of evaluation statements. Table 5 summarizes the panel's feedback regarding the general standard-setting process. The majority of panelists marked *strongly agree* 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. No panelist noted *disagree* regarding any of the evaluation statements.

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). These results were expected, given the central role of the JQC definitions in the standard-setting process and the reliance on professional judgment throughout the process.

Table 5. Final Evaluations

Statement	Strongly agree *N*	Agree *N*
The pre-meeting assignment was useful preparation for the study.	15	3
I understood the purpose of this study.	15	3
The instructions and explanations provided by the facilitators were clear.	15	3
The training in the standard-setting methods was adequate to give me the information I needed to complete my assignment.	15	3
The explanation of how recommended cut scores are computed was clear.	16	2
The opportunity for feedback and discussion between rounds was helpful.	17	1
The process of making the standard-setting judgments was easy to follow.	12	6

Recall that the final cut scores are the Round 3 averages, rounded to the next highest raw score. Table 6 presents the final cut-score recommendations for the four TOEFL Junior Comprehensive test sections. Table 7 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—12 and 13 panelists indicated such for listening and reading, respectively; 11 indicated the same for writing and speaking. Approximately one third of the panelists noted they were only *somewhat comfortable* with the cut scores; no panelists reported being *uncomfortable* with the cut scores.

Table 6. Final Recommended Cut Scores

Section	Raw points A2	B1	B2
Listening (max. 28 raw points)	11	21	27
Reading (max. 28 raw points)	9	22	27
Writing (max. 16 raw points)	6	10	13
Speaking (max. 16 raw points)	8	11	14

Table 7. Comfort Level With the Recommended Cut Scores

	Very comfortable		Somewhat comfortable	
Section	*N*	%	*N*	%
Listening	12	67%	6	33%
Reading	13	72%	5	28%
Writing	11	61%	7	39%
Speaking	11	61%	7	39%

Conclusions

The purpose of this study was to link scores on the TOEFL Junior Comprehensive test of listening, reading, writing, and speaking to the Common European Framework of Reference (CEFR) levels. Specifically, standard-setting procedures were used to identify the minimum test scores needed to reach the A2, B1, and B2 levels of the CEFR. A Yes/No Angoff was applied to the listening and reading sections, and a Performance Profile approach was applied to the writing and speaking sections.

The responses to the final evaluation survey support the validity of the standard-setting process. The majority of the panelists indicated, for example, that the standard-setting training had prepared them, that the provided instructions and explanations were clear, and that the process was easy to follow. Additional evidence of 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 estimate the uncertainty of the panelists' judgments, were relatively small in value, and all were less than half the SEM for the respective test section. This indicates that the panelists were reasonably consistent in their judgments. It is worth noting that the B2 cut scores for each of the test sections were quite high (see Table 6). This finding suggests that the TOEFL Junior Comprehensive test may be a more reasonable measure of the A2 and B1 levels of the CEFR than of the B2 level.

Poststudy Adjustments

The process of setting a standard typically begins with the informed judgments (recommendations) of a panel of experts, but it does not necessarily stop there (Geisinger & McCormick, 2010). Decision makers, those who interpret and use cut-score recommendations, are encouraged to evaluate them, and to adjust them, in relation to their specific needs and values, and other relevant information, including scores from other assessments (Geisinger & McCormick, 2010; Kane, 2001). This recognition is consistent with the understanding that setting a standard (depicted by a cut score) is more similar to the formation of a policy than it is to statistical estimation (Cizek & Bunch, 2007; Kane, 2001; Tannenbaum & Katz, 2013). As Kane and Tannenbaum (2013) remarked: "It is not an accident that standards are said to be 'set' rather than 'estimated'" (p. 177). Cut scores are neither fixed nor correct values; they, like policies, are constructed (Zieky, 2001). And, the ultimate objective of setting standards, like forming policies, is to propose decision rules that are reasonable and appropriate for their intended use.

It is with this objective in mind, that subsequent to the standard setting for the TOEFL Junior Comprehensive test, an analysis was conducted to inform adjustments to the cut-score recommendations. There are two versions of the TOEFL Junior test: standard and comprehensive. TOEFL Junior Standard measures listening, reading, and language form and meaning skills. Scores on TOEFL Junior Standard had previously been mapped to the CEFR (Baron & Tannenbaum, 2011). Those results and the results from the current mapping study were jointly evaluated through a variation of an equipercentile equating approach in an ETS confidential memorandum (ETS, 2012).[3] Approximately 1,000 students took both versions of the TOEFL Junior test and minor adjustments were made to the cut scores to bring the proportion of students reaching the (listening and reading) cut scores across the two test versions into more consistent alignment. Table 8 presents revised cut scores for TOEFL Junior Comprehensive listening and reading, expressed in scale score values; the values for writing and speaking were not adjusted because there is no corresponding measure of those skills on the TOEFL Junior Standard test. The ranges of the scale scores are also provided in the table. Note that for writing and speaking the scale scores are the same as the raw scores.

Table 8. Final Cut Scores in Scale Score Values

	Scale values		
	A2	B1	B2
Listening (140–160 points)	143	150	157
Reading (140–160 points)	143	151	157
Writing (0–16 points)	6	10	13
Speaking (0–16 points)	8	11	14

References

Baron, P. A., & Papageorgiou, S. (2014). *Mapping the TOEFL® Primary™ test onto the Common European Framework of Reference* (Research Memorandum No. RM-14-05). Princeton, NJ: Educational Testing Service.

Baron, P. A., & Tannenbaum, R. J. (2011). *Mapping the TOEFL Junior® test onto the Common European Framework of Reference* (Research Memorandum No. RM-11-07). Princeton, NJ: Educational Testing Service.

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 Language: Learning, teaching, assessment*. Cambridge, UK: Cambridge University Press.

Council of Europe. (2009). *Relating language examinations to the common European framework of reference for languages: Learning, teaching, assessment.* Strasbourg, France: Council of Europe.

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.

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. (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. 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.

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

Kane, M. T., & Tannenbaum, R. J. (2013). The role of construct maps in standard setting. *Measurement*: *Interdisciplinary Research & Perspectives, 11,* 177–180.

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.

North, B. (2014). *The CEFR in practice*. Cambridge, UK: Cambridge University Press.

North, B., Martyniuk, W., & Panthier, J. (2010). Introduction: The manual for relation language examinations to the Common European Framework of Reference for Languages in the context of the Council of Europe's work on language education. In W. Martyniuk (Ed.), *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual* (pp. 1–17). 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., & Tannenbaum, R. J. (in press). Situating standard setting within argument-based validity. *Language Assessment Quarterly*.

Reckase, M. D. (2001). Innovative methods for helping standard-setting participants to perform their task: The role of feedback regarding consistency, accuracy, and impact. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 159–173). Mahwah, NJ: Lawrence Erlbaum.

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 RM-10-11). Princeton, NJ: Educational Testing Service.

Tannenbaum, R. J., & Cho, Y. (2014). Critical factors to consider in evaluating standard-setting studies to map language test scores to frameworks of language proficiency. *Language Assessment Quarterly, 11,* 233–249.

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 Research Report No. TOEFLiBT-06). Princeton, NJ: Educational Testing Service.

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,* 1–10.

Zieky, M. J. (2001). So much has changed: How the setting of cutscores has evolved since the 1980s. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 19–52). Mahwah, NJ: Lawrence Erlbaum.

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:

Registration, Receive materials

Welcome and Overview

Listening: Review and discuss

Develop JQC definitions for CEFR levels A2, B1, and B2

Lunch

Training on standard setting method; training evaluation

Round 1 judgments

Break

Round 1 discussion and Round 2 judgments

Adjourn for the Day

Day 2:

Sign in and receive materials

Round 2 feedback for Listening

Round 3 judgments

Reading: Review and discuss

Develop JQC definitions for A2, B1, and B2

Lunch

Round 1 judgments

Break

Round 1 discussion and Round 2 judgments

Break

Round 2 feedback and Round 3 judgments

Adjourn for the Day

Standard-Setting Agenda (Continued)

Day 3:

Sign in and receive materials

Writing: Review and discuss

Develop JQC definitions for A2, B1, and B2

Training on standard setting method; training evaluation

Lunch

Round 1 judgments

Break

Round 1 discussion and Round 2 judgments

Break

Round 3 judgments

Adjourn for the Day

Day 4:

Sign in and receive materials

Speaking: Review and discuss

Develop JQC definitions for A2, B1, and B2

Round 1 judgments

Break

Round 1 discussion and Round 2 judgments

Lunch

Round 3 judgments

Final evaluations of process

Closing comments

End of Study

Appendix B. Panelists and Countries

Panelist	Country
Yüksel Altunkaynak	Turkey
Ioannis Bardis	Greece
Karine Bonechi	France
Özlem Önder Çelik	Turkey
Denise Giraud Dupui	Mexico
Vania Ledesma Espino	Mexico
Stéphanie Gharbi	France
Dang Hiep Giang	Vietnam
Carmen Maria Hester	Brazil
Mark Leverage	Kuwait
Natalia Matveyeva	Russia
Miroslava Mišová	Slovakia
Mario Alberto Sanabria Moreira	Costa Rica
Isna Shorbirin	Indonesia
Angela Margaret Walshe	Italy
Robert E. Woodhead	Thailand
Chien-Fen Yeh	Taiwan
Tang Zhenhua	China

Appendix C. Just Qualified Candidate (JQC) Descriptions

Listening

JQC A2

1. understands discrete sentences on familiar and factual information
2. understands brief, clear expected announcements, when delivered slowly
3. identifies the topic of a spoken text
4. requires repetition on less familiar topics
5. understands instructions and directions when delivered in a list format, and with visual clues
6. understands basic vocabulary beyond immediate needs
7. recognizes simple tone of voice (emotion and intent)
8. identifies the topic of a spoken text

JQC B1

1. understands the main points/general meaning of clear speech on familiar matters (school, family, leisure)
2. understands and follows a short sequence of instructions and directions, when speech is clear and relatively slowly delivered
3. understands the majority of audio/TV broadcasts on topics of personal interest when delivered relatively slowly and clearly
4. understands the main idea/gist of a conversation between native speakers about topics which are of interest
5. catches key words and some details about places, time in everyday conversations, short narratives
6. infers meaning about familiar matters
7. guesses the meaning of unknown words from context
8. understands main ideas when presented in seminar format and supported by visual stimulus, body language, etc.

9. understands familiar pronunciation
10. understands different tones, moods and emotions expressed by peer group

JQC B2

1. understands most details at normal, standard speed
2. understands reasonably familiar topics
3. understands simple nuance
4. follows the main points and most details in conversations between native speakers

Reading

JQC A2

1. understands information in short, simple, clear, predictable, familiar texts (e.g., short-emails, simple regulations) on familiar topics that include simple, commonly used conjunctions and high-frequency vocabulary
2. locates information from lists
3. makes simple inferences

JQC B1

1. understands the main facts of a text, which addresses about his/her filed of interest
2. scans for information and picks out some details from longer tests, beyond personal experience
3. understands the meaning of unfamiliar words in familiar contexts
4. recognizes the flow of an argument in medium-length texts
5. comprehends explicit sequences of events
6. infers simple, concrete information
7. applies appropriate reading strategies across different types of texts
8. identifies author's intention and purpose through tone and vocabulary on familiar topics

JQC B2

1. understands the main points and most details in different types of texts outside his/her field of interest
2. understands more complex structures and a wider range of vocabulary in context
3. understands implicit ideas in familiar contexts
4. identifies different points of view in texts beyond his/her field of interest
5. understands the use of some figurative speech

Writing

JQC A2

1. writes simple phrases and sentences linked with simple connectors (and, but, because) in fields of personal interest
2. controls vocabulary when related to concrete matters, everyday activities and personal matters
3. describes using limited vocabulary
4. uses basic structures (present, simple past and simple future), with some errors

JQC B1

1. generates written language (essays, letters, stories, reports) provided the subject is about personal matters by using simple sentences and simple linking words
2. operates with a range of tenses on a communicative level, although not always accurately
3. describes events of personal interest and expresses feelings and opinions in simple connected sentences
4. understands the writing purpose, and organizes a text logically and structures ideas effectively on familiar matters
5. uses reasonably precise and varied vocabulary on topics of personal interest and everyday life

6. uses accurate grammar based on modeled language and patterns, with some inconsistencies
7. summarizes factual information
8. supports opinion with some details (some lexical mistakes occur but they do not impede comprehension)
9. understands the writing purpose (the writer, the audience, the message)
10. uses spelling and punctuation with reasonable accuracy; inaccuracies in mechanics do not impede comprehension

JQC B2

1. writes cohesive text with developed argument with support materials
2. writes more complex text with some errors that do not impede comprehension
3. uses, on occasion, more common idiomatic expressions, but not always appropriately

Speaking

JQC A2

1. begins to use intonation appropriate to situations
2. sequences events in time and logical order about everyday activities
3. engages in short conversations provided they occurs with a sympathetic listener who prompts and encourages
4. relies on circumlocution
5. uses short simple sentences without supporting details

JQC B1

1. speech is generally accurate, demonstrating production of final sounds, word endings, time references, etc.
2. mistakes in pronunciation, intonation and L1 interference do not cause misunderstanding or impede communication
3. summarizes information from familiar written or spoken sources

4. uses intonation to convey meaning
5. describes familiar subjects reasonably fluently, using structures, and pertinent vocabulary at an acceptable speed
6. expresses himself/herself on a range of topics about predictable matters with prior preparation or support material
7. expresses opinion using some supporting details
8. uses reasonably precise and varied vocabulary of a moderate range, on topics outside his/her field of interest
9. sustains a conversation, even though hesitation and pauses occur
10. narrates a story linking sentences where inaccuracies don't impede comprehension

JQC B2

1. produces spontaneous speech on familiar topics
2. uses a wide range of vocabulary on familiar topics
3. uses more complex structures where errors do not impede meaning
4. elaborates on familiar topics using supporting details, but may require some prompting
5. produces complex speech occasionally
6. demonstrates ease of expression with minimal pauses and hesitations

Notes

[1] One panelist requested additional training, before signing the evaluation form.

[2] Eleven speaking profiles from across the score scale were initially played aloud. Panelists requested that three additional profiles be played to help them come to their final decisions.

[3] This confidential memorandum is not included in the reference section of this report.