Listening. Learning. Leading.®

Research Memorandum
ETS RM–11-07

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

Patricia A. Baron

Richard J. Tannenbaum

May 2011

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

Patricia A. Baron and Richard J. Tannenbaum[1]
ETS, Princeton, New Jersey

May 2011

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Technical Review Editor: Daniel Eignor

Technical Reviewers: Samuel Livingston and Michael Zieky

Abstract

The purpose of this study was to conduct a standard setting to link scores on the TOEFL® *Junior*™ 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)*, 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" (CEFR, Council of Europe, 2001, p. 1). The TOEFL Junior test measures the academic and social English-language skills representative of English-medium instructional environments of middle school students. The test consists of multiple-choice questions in Listening Comprehension, Language Form and Meaning, and Reading Comprehension sections. Three levels of the CEFR: A2, B1, and B2, were judged as most clearly aligned with TOEFL Junior sections. A modified Angoff standard-setting approach was followed to identify the TOEFL Junior scores linked to these three CEFR levels. Fourteen language experts from nine countries served on the standard-setting panel. The results of this study provide policy makers with minimum scaled scores (cut scores) recommended by the panel for each of the three CEFR levels on the TOEFL Junior sections.

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

Acknowledgments

We extend our sincere appreciation to Zeineb Mazouz, our colleague from the ETS Global BV office for organizing the logistics of the meeting and managing the general well-being of the group. We also thank colleagues from the ETS Princeton office, Paul Rybinski, who provided assessment development expertise prior to and during the meeting, Lisa Costas, for organizing the materials, and Craig Stief, for his work on all the rating forms, analysis programs, and on-site scanning.

Table of Contents

List of Tables

The purpose of this study was to conduct a standard setting to link scores on the TOEFL® *Junior*™ 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)*, 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" (CEFR, Council of Europe, 2001, p. 1). The TOEFL Junior test measures the academic and social English-language skills representative of English-medium instructional environments of middle school students. The test consists of multiple-choice questions in Listening Comprehension, Language Form and Meaning, and Reading Comprehension sections.

The TOEFL Junior test was neither specifically designed 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. Standard-setting panelists cannot develop cut scores on a test for levels of knowledge and skill that are not included in the test design. Therefore, before conducting the standard-setting study, ETS testing experts identified the specific CEFR levels that were most clearly aligned with the TOEFL Junior Listening Comprehension, Language Form and Meaning, and Reading Comprehension sections. Each section was judged to address CEFR levels A2 through B2. 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 of the three sections of the test to reach each of the targeted CEFR levels (A2, B1, B2). For each section of the test, the general process of standard setting was conducted in a series of steps that will be elaborated upon below. A modified Angoff standard-setting approach was followed to identify the TOEFL Junior scores linked to the A2 through B2 levels of the CEFR (Cizek & Bunch, 2007; Zieky, Perie, & Livingston, 2008). The specific implementation of this approach followed the work of Tannenbaum and Wylie (2008) in which cut scores were constructed linking *Test of English for International Communication*™ (TOEIC®) to the CEFR. Similar studies have been recently conducted using this approach (Baron & Tannenbaum, 2010; Tannenbaum & Baron, 2010). Recent reviews of research on standard-setting approaches reinforce a number of core principles for best practice: careful selection of panel members/experts and a sufficient

number of panel members to represent varying perspectives, sufficient time devoted to ensure development of a common understanding of the domain under consideration, adequate training of judges, 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 & Katz, in press). The approach used in this study adheres to these principles.

Panelists

Fourteen individuals from nine countries served on the panel. All had expertise in English language instruction, and/or assessment. Thirteen taught English as a foreign or second language to students between ages 11 and 14 (middle school), and one taught high school. All had at least five years of experience teaching English as a foreign or second language. The panelists/experts were familiar with the CEFR, with the general population of students who would likely take TOEFL Junior, and with the TOEFL Junior test, the latter due to their taking the test as a provision of serving on the panel. Table 1 provides a description of the self-reported demographics of the panelists. (See the Appendix A for panelist affiliations.)

Premeeting Assignment

Prior to the standard-setting study, the panelists were asked to complete two activities to prepare them for work at the study. All panelists were asked to complete an assignment related to the CEFR and to take the TOEFL Junior 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 Comprehension, Language Form and Meaning, and Reading Comprehension. They were provided with selected tables from the CEFR, and asked to respond to the following questions based on the CEFR and their own knowledge and experience teaching English as a foreign or second language to students: What should you expect students ages 11–14 who are at the beginning of each CEFR level to be able to do in English? What in-class behaviors would you observe to let you know the level of the student's ability in listening and reading comprehension and in language form and meaning? The panelists were asked to consider characteristics that define students with "just enough" English skills to enter into each of the three CEFR levels, and to make notes and bring those to the workshop to use as a starting point for discussion. This homework assignment was useful as a familiarization

tool for the panelists, in that they were beginning to think about the minimum requirements for each of the CEFR levels under consideration.

Each expert also took the TOEFL Junior test prior to the standard-setting study; they had each signed a nondisclosure/confidentiality form before having access to the test. 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 questions on the test.

Table 1

Panelist Demographics

		Number
Gender	Female	11
	Male	3
Function	Teacher	12
	Testing Specialist	1
	Coordinator	1
Experience teaching middle school (ages 11–14)	Less than 5 years	1
	5–10 years	4
	More than 10 years	8
	Other*	1
Experience teaching English as a foreign or second language	Less than 5 years	0
	5–10 years	3
	More than 10 years	11
Country	Brazil	1
	Colombia	2
	Denmark	1
	Egypt	1
	France	2
	Greece	1
	Italy	4
	Jordan	1
	Poland	1

* Thirty years' experience teaching high school students.

Standard-Setting Process

The general process of standard setting was conducted in a series of steps for each section: Listening Comprehension, followed by Language Form and Meaning, and finally Reading Comprehension. See Appendix B for the agenda. In the first step of the process for each section, the panelists defined the minimum skills needed to reach each of the targeted CEFR levels (A2, B1, B2). A test taker who has these minimally acceptable skills is referred to as a *just qualified candidate* (JQC). Following a general discussion on what the test section measures, the panelists worked in two small groups, with each group defining the skills of a candidate who just meets the expectations of someone performing at the B1 level.[2]

Panelists referred to their prestudy notes, and to a draft list of Can Do statements for each level which was provided as a starting point for discussion and development of the JQC. A whole-panel discussion of the small group lists was facilitated, and concluded with a consensus definition for the B1 level JQC. Definitions of the JQC for A2 and B2 levels were accomplished through whole-panel discussion, using the B1 descriptions as a starting point. 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. (See Appendix C for JQC Descriptions.)

A modified Angoff approach was implemented following the procedures of Tannenbaum and Wylie (2008), which included three rounds of judgments informed by feedback and discussion between rounds. Prior to judgments made on the first section (Listening Comprehension), the panelists were trained in the process and then given opportunity to practice making their judgments. At this point, they were asked to sign a training evaluation form confirming their understanding and readiness to proceed, which all did. In Round 1, for each test question, panelists were asked to judge the percentage of *just qualified candidates* for the A2 and B2 levels who would answer the question correctly. They used the following judgment scale (expressed as percentages): 0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100. The panelists were instructed to focus only on the alignment between the English skills demanded by the question and the English skills possessed by JQCs, and not to factor random guessing into their judgments. For each test question, they made judgments for each of the two CEFR levels (A2 and B2) before moving to the next question. After making judgments, panelists received feedback on individual and group judgments.

Feedback following Round 1 judgments was based on the panel's judgments and on test taker performance. The sum of each panelist's cross-item judgments (divided by 100) represents his or her recommended cut score. Each panelist's recommended cut score was provided to him/her.

The panel's recommended cut score, and the highest and lowest cut scores were compiled and presented to the panel to foster discussion. The panel-recommended cut score is computed by taking the average of the panelists' judgments. The average was then rounded to the next highest whole number; it is this whole number that represents the recommended cut score. Similarly, the highest and lowest cut scores presented to the panelists are first rounded to the next highest whole number before being presented to the panelists as feedback.

Panelists were then asked to share their judgment rationales. As part of the feedback and discussion, *p* values (percentage of test takers who answered each question correctly), were shared. The feedback was based on the performance data of test takers from the first administration of the TOEFL Junior (October 2010, $N = 8{,}529$). In addition, *p* values were calculated for candidates scoring at or above the 75th percentile on that particular section (i.e., the top 25% of candidates) and for candidates at or below the 25th percentile (i.e., the bottom 25% of candidates). Examining question difficulty for the top 25% of candidates and the bottom 25% of candidates was intended to give panelists a better understanding of the relationship between overall language ability for that TOEFL Junior test section and each of the questions. The partitioning, for example, enabled panelists to see any instances where a question was not discriminating, or where a question was found to be particularly challenging or easy for test takers at the different ability levels. After discussion, panelists made Round 2 judgments.

In Round 2, judgments were made again at the question level; panelists were asked to take into account the feedback and discussion from Round 1, and were instructed that they could make changes to their ratings for any question(s), for either A2 or B2 levels, or both. The Round 2 judgments were compiled, and feedback similar to that presented in Round 1 was provided. In addition, impact data from the October 2010 test administration were presented; panelists discussed the percentage of test takers who would be classified into each of the levels currently recommended (percent below A2, percent above B2, and the percent between the two current recommendations for cut scores for A2 and B2, which includes students who would be classified

at the A2 and B1 levels). At the end of the Round 2 feedback and discussion, panelists were given instructions to make Round 3 judgments.

In Round 3, panelists were asked to consider the cut scores for the overall section (e.g., Listening Comprehension). Specifically, panelists were asked to review the JQC definitions of all three levels and to decide on the recommended cut score for B1, taking into account the Round 2 cut score recommendations and discussions regarding the A2 and B2 levels. They were instructed that for Round 3, they should indicate their final cut score recommendation at the section level for A2 and B2, then locate the B1 cut score, using the A1 and B2 cut scores as references. The transition to a section-level judgment places emphasis on the overall constructs of interest (i.e., Listening Comprehension, Language Form and Meaning, and Reading Comprehension) rather than on the deconstruction of the constructs through another series of question-level judgments. This modification had been used in previous linking studies (e.g., Tannenbaum & Wylie, 2005, 2008), and posed no difficulties for the TOEFL Junior panelists.

At the conclusion of Round 3 judgments for each section, the process was repeated for the next test section, starting with the general discussion of what the section measured and a discussion of minimum skills needed to reach each of the targeted CEFR levels (JQC definitions), followed by three rounds of judgments and feedback. After final (Round 3) judgments were compiled for all three sections, the results of the standard setting were presented to the panel and final evaluations were completed.

Results

The first set of results summarizes the panel's standard-setting judgments for each of the TOEFL Junior test sections. The tables summarize the results of the standard setting for Levels A2 and B2 for Rounds 1 and 2, and for Levels A2, B1, and B2 for the final round of judgments. The results are presented in raw scores, which is the metric that the panelists used. The final panel-recommended cut score is computed by taking the average of the panelists' judgments. For Round 3, the panel-recommended cut score is based on the panelists' Round 3 holistic judgments. The average was then rounded to the next highest whole number; it is this whole number that represents the final recommended cut score for each round. Also included in each table is the standard error of judgment (SEJ), which indicates how close each recommended cut score is likely to be to a cut score recommended by other panels of experts similar in composition to the current panel and similarly trained in the same standard-setting method.[3] The

last set of results is a summary of the panel's responses to the end-of-study evaluation survey. (The scaled cut scores are provided in the conclusion section.)

TOEFL Junior Listening Comprehension. Table 2 summarizes the results of the standard setting for each round of judgments. The recommended cut score for A2 decreased at Round 2, and did not change at Round 3. The recommended cut score for B2 remained the same across three rounds. For both A2 and B2, the variability among the panelists decreased over three rounds, as can be seen by the decrease in the standard deviations (SD). This decrease in variability suggests that the panelists' individual judgments are converging after each round of judgment, feedback, and discussion. The SEJ, which is a function of variance, also decreased over rounds. The interpretation of the SEJ is that a comparable panel's cut score would be within one SEJ of the current cut score 68% of the time and within two SEJs 95% of the time. The SEJ for Listening Comprehension at Round 3 is less than one point for all three levels, which is relatively small, and provides some confidence that the recommended cut score would be similar were a panel with comparable characteristics convened.

Table 2

Listening Comprehension Standard Setting Results

	Round 1		Round 2		Round 3		
Levels	A2	B2	A2	B2	A2	B1	B2
Recommended cut score	9.00	24.00	8.00	24.00	8.00	17.00	24.00
Mean	8.37	23.49	7.95	23.54	7.74	16.21	23.57
Median	8.50	23.70	8.58	23.83	8.25	16.50	23.50
Minimum	2.15	19.75	2.20	20.80	2.20	11.00	21.00
Maximum	12.70	27.40	10.30	26.45	10.00	19.00	26.45
SD	3.03	2.05	2.28	1.74	2.03	2.08	1.45
SEJ	0.81	0.55	0.61	0.46	0.54	0.56	0.39

TOEFL Junior Language Form and Meaning. Table 3 summarizes the results of the standard setting for each round of judgments. The pattern for this section is similar to that seen for Listening Comprehension. The recommended cut score for A2 decreased at Round 2, and at Round 3, reverted to the recommendation made at Round 1. The cut score for B2 remained the same across three rounds. For both A2 and B2, the panelists' judgments converged across the three rounds of judgments, as seen in the decrease in the standard deviations (SD). As with the previous section, panelists' judgments are less disparate after feedback and discussion. The SEJs

similarly decreased across rounds. The Round 3 SEJ for all three levels is less than one point, which is relatively small.

Table 3

Language Form and Meaning Standard Setting Results

	Round 1		Round 2		Round 3		
Levels	A2	B2	A2	B2	A2	B1	B2
Recommended cut score	7.00	26.00	6.00	26.00	7.00	18.00	26.00
Average	6.23	25.30	6.00	25.30	6.17	17.23	25.24
Median	5.10	25.40	5.65	25.40	6.00	17.00	25.05
Minimum	2.05	21.50	2.05	22.35	3.00	14.00	23.00
Maximum	11.40	28.35	11.10	28.35	10.00	21.00	27.05
SD	3.17	1.98	2.58	1.75	1.98	2.17	1.24
SEJ	0.88	0.55	0.71	0.48	0.55	0.60	0.34

Note. One panelist did not participate in the standard setting for Language Form and Meaning due to illness.

TOEFL Junior Reading Comprehension. Table 4 summarizes the results of the standard setting for each round of judgments. The pattern of recommendations across rounds is similar to that observed for Listening Comprehension and Language Form and Meaning, in that there were only slight changes in the cut score recommendations across the three rounds. The recommendation for A2 remained the same for three rounds. For B2 recommended cut scores, panelists decreased their overall cut score by one point from Round 1 to Round 2, but returned to the Round 1 recommendation at Round 3. The variability (SD) in panelists' judgments decreased across three rounds for A2, and for B2, increased minimally from Round 1 to Round 2, and decreased at Round 3. The overall reduction in the standard deviation across three rounds for Reading Comprehension indicates some convergence in the final round of judgments. The pattern for SEJs is consistent with the SD pattern of results, as expected. The Round 3 SEJ for all three levels is less than one point.

End-of-study evaluation survey. Panelists responded to a final set of questions addressing the procedural evidence for validity of the standard setting process (Kane, 1994). The survey is a tool to gather evidence that the procedures have been implemented in a reasonable way, that is, panelists understood the purpose of the standard setting process, understood what they were doing, etc. Table 5 summarizes the panel's feedback regarding the general process.

The majority of panelists *strongly agreed* or *agreed* that the premeeting assignment was useful, that they understood the purpose of the study, that instructions and explanations provided were clear, that the training provided was adequate, that the opportunity for feedback and discussion was helpful, and that the standard setting process was easy to follow.

Table 4

Reading Comprehension Standard Setting Results

	Round 1		Round 2		Round 3		
Levels	A2	B2	A2	B2	A2	B1	B2
Recommended cut score	10.00	26.00	10.00	25.00	10.00	18.00	26.00
Average	9.58	25.18	9.04	24.95	9.42	17.71	25.01
Median	9.88	25.63	9.08	25.30	9.50	18.00	25.00
Minimum	4.25	19.60	5.15	19.95	6.00	14.00	21.00
Maximum	16.35	28.30	15.00	28.30	15.00	22.00	28.30
SD	3.36	2.16	2.86	2.20	2.59	2.23	1.86
SEJ	0.90	0.58	0.76	0.59	0.69	0.60	0.50

Table 5

Feedback on Standard Setting Process

	Strongly agree		Agree		Disagree		Strongly disagree	
	N	%	*N*	%	*N*	%	*N*	%
The homework assignment was useful preparation for the study.	11	79%	3	21%	0	0%	0	0%
I understood the purpose of this study.	14	100%	0	0%	0	0%	0	0%
The instructions and explanations provided by the facilitators were clear.	14	100%	0	0%	0	0%	0	0%
The training in the standard setting method was adequate to give me the information I needed to complete my assignment.	13	93%	1	7%	0	0%	0	0%
The explanation of how the recommended cut scores are computed was clear.	12	86%	2	14%	0	0%	0	0%
The opportunity for feedback and discussion between rounds was helpful.	12	86%	2	14%	0	0%	0	0%
The process of making the standard setting judgments was easy to follow.	11	79%	3	21%	0	0%	0	0%

Additional questions focused on how influential each of the following four factors was in their standard-setting judgments: the definition of the JQC, the between-round discussions, the cut scores of the other panelists, and their own professional experience. All panelists indicated these factors to be either *very influential* or *somewhat influential*. More than half of the panelists indicated that the definition of the JQC and their own professional experience were *very influential*; half of the panelists indicated that the between-round discussions were *very influential*, the other half indicating these discussions were only *somewhat influential*. All but one of the panelists indicated that the cut scores of other panelists were only *somewhat influential*.

Panelists were also asked to indicate their level of comfort with the final cut score recommendations; Table 6 summarizes these results. A majority of the panelists reported they were either *very comfortable* or *somewhat comfortable* with the recommended cut scores for the three sections. All fourteen of the panelists indicated this level of comfort with the cut scores for Listening Comprehension. Two panelists reported being *somewhat uncomfortable* with the recommended cut scores for Language Form and Meaning, and one panelist indicated being *somewhat uncomfortable* with the recommended cut scores for Reading Comprehension.

Table 6

Comfort Level With Recommended Cut Scores for TOEFL Junior

	Very comfortable		Somewhat comfortable		Somewhat uncomfortable		Very uncomfortable	
	N	%	*N*	%	*N*	%	*N*	%
Listening Comprehension	11	79%	3	21%	0	0%	0	0%
Language Form and Meaning	9	64%	3	21%	2	14%	0	0%
Reading Comprehension	10	71%	3	21%	1	7%	0	0%

As part of the final evaluation, panelists were given an opportunity to provide general comments. All panelists wrote brief comments; some themes were mentioned more than once, as follows. Three panelists indicated that they felt that the B2 cut scores were too high, specifically for Language Form and Meaning and Reading Comprehension. One panelist commented that the

A2 cut score for Reading Comprehension was too high; another panelist stated that they expected a higher A2 cut score for Listening Comprehension and Language Form and Meaning. Some panelists also suggested that the interactions of the panel indicated a great amount of knowledge about students' levels and the CEFR; others commented that the composition of the panel would have benefited from more variation across language skill level; specifically there were very few panel members with experience teaching highly able students, relative to the three CEFR levels addressed. Two comments emerged concerning the level of skills being tested; both of these panelists felt that the test may be more appropriate for older, more motivated students.

Conclusions

The purpose of this standard setting study was to recommend cut scores (minimum scores) for TOEFL Junior Listening Comprehension, Language Form and Meaning, and Reading Comprehension sections that correspond to the A2, B1, and B2 levels of the CEFR. A modified Angoff standard-setting approach was implemented. The panelists worked in the raw score metric during the study. Three rounds of judgments, with feedback and discussion, occurred to construct the cut scores. Feedback included October 2010 test administration data on how test takers performed on each of the questions and the percentage of test takers who would have been classified into each of the targeted CEFR levels. Following the completion of the standard–setting study, a scaling procedure was implemented, transforming test scores from the raw score metric to the new reporting scale. The scaling process makes it possible to compare test scores across different forms of the TOEFL Junior test. The scaled score range for each section of the test is 200 to 300. The scaling process was conducted such that the panel-recommended raw cut scores for each test section correspond to the same scaled cut scores. The assumption in the scaling is that the recommended cut scores will be accepted. Table 7 presents the scaled cut scores.

Table 7

Scaled Cut Scores for TOEFL Junior

TOEFL Junior scaled cut score for each section		
A2	B1	B2
210	250	280

The responses to the end-of-study evaluation survey support the quality of the standard setting implementation (evidence for procedural validity). The majority of panelists *strongly agreed* or *agreed* that they understood the purpose of the study, that instructions and explanation provided were clear, that the training provided was adequate, that the opportunity for feedback and discussion was helpful, and that the standard-setting process was easy to follow. Procedural evidence for validity reinforces the reasonableness of the recommended cut scores.

Setting Final Cut Scores

The standard-setting panel is responsible for recommending cut scores. Policymakers consider the recommendations, but are responsible for setting the final cut scores (Kane, 2002). In the context of the TOEFL Junior test, policymakers may be members of an academic institution that need to have a decision rule, for example, pertaining to the level of English instruction appropriate for current or incoming middle school students. Policymakers may also want to evaluate students' performance levels as a factor in curriculum design or staff recruitment. The needs and expectations of policymakers vary, and cannot be represented in full during the process of recommending cut scores. Policymakers, therefore, have the right and responsibility of considering both the panel's recommended cut scores and other sources of information when setting the final cut scores (Geisinger & McCormick, 2010). The recommended cut scores may be accepted, adjusted upward to reflect more stringent expectations, or adjusted downward to reflect more lenient expectations. There is no single correct decision; the appropriateness of any adjustment may only be evaluated in terms of meeting the policymaker's needs. Two sources of information often considered by policymakers when setting cut scores are the standard error of measurement (SEM) and the standard error of judgment (SEJ). The former addresses the reliability of test scores and the latter the reliability of panelists' cut score recommendations.

The SEM is a measure of the uncertainty of a test score; it takes into account that a test score—any test score on any test—is less than perfectly reliable. The SEM addresses the question: "How close of an approximation is the test score to the *true score*?" A test taker's score likely will be within one SEM of his or her true score 68% of the time and within two SEMs 95% of the time. The *scaled score* SEMs for TOEFL Junior Listening Comprehension, Language Form and Meaning, and Reading Comprehension sections are 10 points each.

The SEJ allows policymakers to consider the likelihood that the current recommended cut score would be recommended by other panels of experts similar in composition and experience to the current panel. The smaller the SEJ, the more likely that another panel would recommend a cut score consistent with the current cut score. The larger the SEJ, the less likely the recommended cut score would be reproduced by another panel. The SEJ, therefore, is sometimes considered a measure of credibility, in that there would be consistency; a recommendation may be more credible if that recommendation were likely to be offered by another panel of experts. An SEJ no more than one-half the size of the SEM is desirable because the SEJ is small relative to the overall measurement error of the test (Cohen, Kane, & Crooks, 1999). The SEJs in this study were in the raw score metric. We approximated the average scaled score change due to the SEJs by applying the raw-to-scale score conversions for each of the TOEFL Junior test sections. In all cases, the SEJ resulted in an average scaled score change of one-third of the scaled SEM. This SEJ value meets the criterion to be considered credible.

In addition to measurement error metrics (e.g., SEM, SEJ), policymakers should consider the likelihood of classification errors. That is, when adjusting a cut score, policymakers should consider whether it is more important to minimize a false positive decision or to minimize a false negative decision. A false positive decision occurs when the conclusion made from a test score is that someone has the required skill, but actually does not. A false negative occurs when the conclusion made from a test score is that someone does not have the required skills, but actually does. For example, a TOEFL Junior Reading Comprehension score may be used to determine whether a student should remain in a class or be moved into the next level of instruction. The nature of instruction and expectations for the students in the B1 level may differ substantially from the A2 level class. The level of frustration for a student may be high if placement into B1 is premature. In that instance, a policymaker may decide that it is more important to minimize a false positive decision, and, erring on the side of caution, elect to raise the cut score for B1 Reading. Raising the cut score reduces the likelihood of a false positive decision, as it increases the stringency of the requirement. (It also, however, means that some number of students who might have been at B1-level in Reading will now remain in the A2-level class and be denied access to the B1 classroom instruction.) Policymakers need to consider which decision error to minimize; it is not possible to eliminate these decision errors.

References

Baron, P. A., & Tannenbaum, R. J. (2010). *Mapping the Test de Français International™ onto the Common European Framework of Reference* (ETS Research Memorandum No. RM-10-12). Princeton, NJ: ETS.

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 Publications.

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, England: Cambridge University Press.

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 Publishers.

Kane, M. (1994). Validating performance standards associated with passing scores. *Review of Educational Research*, *64*, 425–461.

Kane, M. (2002). Conducting examinee-centered standard-setting studies based on standards of practice. *The Bar Examiner. 71*(4): 6–13.

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

Tannenbaum, R. J., & Katz, I. R. (in press). Standard setting. In K.F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology*. Washington, DC: American Psychological Association.

Tannenbaum, R. J., & Wylie, E. C. (2005). *Mapping English Language Proficiency Test Scores Onto The Common European Framework* (TOEFL Research Report No. RR-80). Princeton, NJ: ETS.

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 Rep. No. TOEFLibt-06, ETS Research Report No. RR-08-34). Princeton, NJ: ETS.

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

Notes

[1] Author names are listed alphabetically.

[2] Panelists were divided into diverse small groups based on their experience; groups were switched around for each discussion to allow for a more robust exchange of ideas.

[3] An SEJ assumes that panelists are randomly selected from a larger pool of panelists and that standard-setting judgments are independent. It is seldom the case that panelists may be considered randomly sampled, and only the first round of judgments may be considered independent. The SEJ, therefore, likely underestimates the uncertainty associated with cut scores (Tannenbaum & Katz, in press).

Appendix A

Panelists' Affiliations

Name	Affiliation
Maria De Lourdes R. Al Fadel	Jordan
Ewelina Agnieszka Bigda	Gimnazjum 44 – Warsaw, Poland
Montserrat Braso Gonzalez	Copenhagen French School – Denmark
Maria Grazia Carcano	Liceo Scientifico Statale "Falcone e Borsellino" – Arese, Italy
Nigel Downey	Hellenic American Union/Hellenic American University – Greece
Fatma El Sakhawy	Nefertari International American School – Cairo, Egypt
Laura Ferrari	IT Artemisia Gentileschi – Milan, Italy
Maia Jeanquier	Saint Haus Saint Clément – France
Lilian Itzicovitch Leventhal	Colégio I. L. Peretz – Brazil
Patrizio Mongiardo	Laura Conti Buccinasco – Milan, Italy
M. T. Richard	Lycée des Me'hers Sainte – Clotilde F. Strasbourg, France
Diana Londonō Romań	Colegio fe y Alegria la Pat – Manzinales, Colombia
Lucero Perilla Vélez	Santa Maria Goretti Institución – Montenegro, Colombia
Felice Verdelli	Collegio San Carlo – Milan, Italy

Note. Panelists' affiliations are listed as requested.

Appendix B

Agenda
November 15–17, 2010

Day 1: Monday, November 15
Start: 8:30 a.m. Finish: 5:30

Registration, Receive materials

Welcome and Overview

Listening Comprehension: Review and discuss

Develop Just Qualified Candidate (JQC) definitions for CEFR levels A2, B1, and B2

Lunch

Training on standard setting method; training evaluation

Round 1 judgments, levels A2 and B2

Afternoon break

Round 1 discussion and Round 2 judgments

Adjourn for the Day

Day 2: Tuesday, November 16
Start: 8:30 a.m. Finish: 5:30

Sign in and receive materials

Round 2 feedback for Listening Comprehension

Round 3 judgments: levels A2, B1, and B2

Language Form and Meaning: Review and discuss

Develop JQC definitions

Lunch

Round 1 judgments

Break

Round 1 discussion and Round 2 Judgments

Break

Round 2 feedback and Round 3 judgments

Reading Comprehension: Review and discuss

Develop JQC definition for B1 level

Adjourn for the Day

Day 3: Wednesday, November 17

Start: 8:30 a.m. Finish: 4:00

Sign in and receive materials

Develop Reading Comprehension JQC definitions for A2 and B2 levels

Round 1 Judgments

Break

Round 1 discussion and Round 2 judgments

Lunch

Round 3 judgments

Final evaluations of process

Panelists' feedback

Adjourn for the Day

Appendix C

Just Qualified Candidate Descriptions

Listening Comprehension

A2 Level
In very slow and carefully articulated spoken texts with long pauses, limited to very concrete and familiar topics and short simple messages, the just qualified A2 student:

- Identifies topic of discussion
- Picks out familiar key words and short phrases
- Recognizes past, present, and future tense (e.g., yesterday, tomorrow).
- Understands basic instructions
- Recognizes spelling

B1 Level
In clearly and slowly articulated extended spoken texts about familiar topics in concrete factual contexts, the just qualified B1 student:

- Understands the main points in a conversation
- Understands very simple and basic phrasal verbs
- Catches key words in general messages or short stories
- Catches a few of the details in a short text, such as quantities, time, place
- Identifies the explicit situation and the speakers and their explicit relationships
- Infers simple concrete information
- Recognizes the differences between formal and informal English
- Understands simple instructions and directions
- Is aware of present and past and future tenses and a basic aspect of present-perfect

B2 Level
In clearly articulated some extended spoken texts at normal speed, about familiar and some abstract topics, the just qualified B1 student:

- Understands specific detailed information
- Infers function, tone; relationship between speakers

- Understands the gist of conversation among multiple speakers
- Understands some common idioms and common phrasal verbs related to everyday activities
- Understands moderately syntactically complex spoken texts

Language Form and Meaning

A2 Level
On short, simple, clear texts on very concrete and familiar topics, the just qualified A2 student:

- Has limited control of simple grammatical structures
- Has basic vocabulary
- Uses present simple, present continuous, and past simple tenses
- Recognizes simple functions

B1 Level
On familiar topics in concrete factual contexts, the just qualified B1 student:

- Paraphrases words, sentences, and meaning
- Communicates some everyday needs and experiences with common vocabulary, but errors may occur without loss of meaning
- Uses appropriate basic grammatical structures (e.g., present, past, future, basic aspect of present perfect) and recognizes the use of 1 and 2 conditionals, passive voice in the present simple and past simple
- Is aware of the use of linkers
- Uses pronouns to give cohesion to text
- Guesses the meaning of vocabulary from the context

B2 Level
On some academic and non-academic topics, the just qualified B2 student:

- Understands more complex sentences
- Understands more lengthy and extended sentences

- Uses and has control of 1st and 2nd conditionals, and recognizes the use of the 3rd conditional
- Recognizes complex syntactic structures
- Uses a wider range of vocabulary, including collocations
- Constructs more complex and cohesive texts
- Uses self-correction and mistakes are less frequent

Reading Comprehension

A2 Level
In written short, simple, clear (linear and nonlinear) texts about familiar and concrete topics, the just qualified A2 student:

- Understands the main idea and some details with the help of familiar words to infer meaning
- Scans for specific information

B1 Level
In written medium-length texts about familiar topics in clear, concrete contexts, the just qualified B1 student:

- Identifies the main idea of the whole text and of each paragraph
- Identifies basic details in simple academic and nonacademic texts
- Infers the meaning of words from context
- Infers specific information in letters, e-mails, ads, events, timetables, short official documents, and nonlinear texts
- Understands the function of linkers and cohesive elements

B2 Level
In written extended-length texts about academic and nonacademic contexts, the just qualified B2 student:

- Uses variety of reading strategies in order to read different texts efficiently
- Understands the gist and some details
- Identifies meaningful details in more complex texts
- Infers implicit information

- Understands a wider range of vocabulary
- Understands more complex grammatical structures
- Understands the function of more complex linkers and cohesive elements