Listening. Learning. Leading.®

Research Memorandum
ETS RM–11-33

Mapping TOEFL® ITP Scores Onto the Common European Framework of Reference

Richard J. Tannenbaum

Patricia A. Baron

November 2011

Mapping TOEFL® ITP Scores Onto the Common European Framework of Reference

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

November 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: Donald Powers and E. Caroline Wylie

Abstract

This report documents a standard-setting study to map TOEFL® ITP scores onto the Common European Framework of Reference (CEFR). The TOEFL ITP test measures students' (older teens and adults) English-language proficiency in three areas: Listening Comprehension, Structure and Written Expression, and Reading Comprehension. This study focused on recommending the minimum scores needed to enter the A2, B1, and B2 levels of the CEFR. A variation of a modified Angoff standard-setting approach was implemented. Eighteen English-language educators from 14 countries served on the standard-setting panel. The results of this study provide policy makers with panel-recommended minimum scores (cut scores) needed to enter each of the three targeted CEFR levels.

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

Acknowledgments

We extend our sincere appreciation to Steven Van Schalkwijk, CEO of Capman Testing Solutions, for hosting the study and Rosalyn Campos, Capman Testing Solutions, for her support during the study. We also thank our colleagues from the ETS Princeton office, Dele Kuku, for organizing the study materials, and Craig Stief, for his work on 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 study to map TOEFL® ITP test scores onto 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 in a comprehensive way what language learners have to learn in order to use a language for communication and what knowledge and skills they have to develop so as to be able to act effectively" (CEFR, Council of Europe, 2001, p. 1). TOEFL ITP is a selected-response test that measures students' (older teens and adults) English-language proficiency in three areas: Listening Comprehension, Structure and Written Expression, and Reading Comprehension. TOEFL ITP content comes from previously administered TOEFL PBT (paper-based) tests. TOEFL ITP tests, therefore, are not fully secure and should not be used for admission purposes. College and universities, English-language programs, and other agencies may use TOEFL ITP test scores, for example, to place students into English-language programs, to measure students' progress throughout those programs, or to assess students' end-of-program English-language proficiency (http://www.ea.etsglobal.org/ea/tests/toefl-itp/).

The focus of this study was to identify for each test section the minimum scores (cut scores) necessary to enter the A2, B1, and B2 levels of the CEFR. Scores delineating these levels support a range of decisions institutions may need to make.

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, and B2). For each section of the test the general process of standard setting was conducted in a series of steps which will be elaborated upon below. A variation of a modified Angoff standard-setting approach was followed to identify the TOEFL ITP scores mapped 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 minimum scores (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 develop a common understanding of the domain under consideration, adequate training of panelists, 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

Directors of the TOEFL program, which includes TOEFL ITP, targeted four regions for inclusion in the current study: EMEA (Europe, Middle East, and Africa), Latin America, Asia Pacific, and the United States. These regions represent important markets for this test. Eighteen educators from 14 countries across the targeted four regions served on the standard-setting panel. Table 1 provides a description of the self-reported demographics of the panelists. Eight panelists were from EMEA, four from Latin America, four from Asia Pacific, and two from the United States. In summary, 11 were teachers of English as a second language (ESL) at either a private school or university; five were administrators, directors, or coordinators of an ESL school, department, or program; and two held different titles. Sixteen panelists had more than 10 years of experience in English-language instruction. (See Appendix A for panelist affiliations.)

Premeeting Activities

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 take the TOEFL ITP test (all three sections). Each panelist had signed a non-disclosure/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. The other activity 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, Structure and Written Expression, 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 of and experience teaching English as second or foreign language to students: What should you expect students 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, structure and written expression, and reading comprehension? 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.

Table 1

Panelist Demographics

Variable		*N*
Gender	Female	11
	Male	7
Function	ESL teacher at language school (private or university)	11
	Administrator, director, or coordinator of ESL school, program, or department	5
	Researcher of language assessment	1
	Director of a language and testing service	1
Experience	5–10 years	2
	More than 10 years	16
Country	Argentina	1
	Chile	1
	China	1
	Colombia	2
	France	2
	Germany	2
	Indonesia	1
	Italy	1
	Japan	1
	Kuwait	1
	Macedonia	1
	Spain	1
	Thailand	1
	United States	2

Standard-Setting Process

The general process of standard setting was conducted in a series of steps for each section: Listening Comprehension, followed by Structure and Written Expression, 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, and B2). A test taker (candidate) 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 three small groups, with each group defining the skills of a candidate who just meets the expectations of someone performing at the B1 level.[1] Panelists referenced their prestudy assignment notes and test-taking experience when constructing their small-group descriptions. A whole-panel discussion of the small group descriptions 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 variation of 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. The first two rounds focused on item-specific judgments for the A2 and B2 levels of the CEFR. In the third (final) round, holistic decisions (section-specific) were made first for the A2 and B2 levels and then for the B1 level. The B1 decision was made using the A2 and B2 decisions as reference points. This approach was used to reduce the cognitive load that would have been imposed if the panelists were to have conducted item-specific judgments for all three levels for each round of judgment. Before making their Round-3 judgments, the panelists were instructed to rereview the JQCs for each level. This was especially important for locating each B1 cut score so that the recommended cut score would be informed by its operational definition (B1 JQC), and not be assumed, by default, to be the average of the A2 and B2 cut scores.

Prior to the first round of judgments made on the first section (Listening Comprehension), the panelists were trained in the standard-setting 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. The sum of each panelist's cross-item judgments (divided by 100) represents his or her recommended cut score. After completing Round-1 judgments, panelists received feedback on their individual cut-score recommendations and on the panel's recommendations (the average of the panelists' recommendations).

The panel's recommended cut scores (for A2 and B2 CEFR levels), the highest and lowest cut-score recommendations, and the standard deviation of the cut-score recommendations were presented to the panel to foster discussion. Panelists were 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 more than 6,000 candidates who in 2010 had taken the form of TOEFL ITP reviewed at the standard-setting study. 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 scoring 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 ITP 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 candidates 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 2010 test administration were presented; panelists discussed the percentage of candidates who would be classified into each of the levels currently recommended—the percent at and above A2, and the percent at and above B2. In addition, the percent below A2 and the percent between A2 and B2 (which covers the A2 and B1 levels) were presented. 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 rereview the JQC definitions of all three CEFR levels and then to decide on the final recommended cut score first for A2, and then for B2. Once these two decisions were made, panelists then decided on the B1 recommended cut score. The A2 and B2 decisions, therefore, served as "anchors" for the B1 decision. The transition to a section-level judgment places emphasis on the overall constructs of interest (i.e., Listening Comprehension, Structure and Written Expression, 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 2008; Tannenbaum & Baron, 2010), and posed no difficulties for the TOEFL ITP 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 completing standard-setting judgments for all three test sections, the final (Round-3) panel-level cut-score recommendations were presented and each panelist completed an end-of-study evaluation.

Results

The first set of results summarizes the panel's standard-setting judgments for each of the TOEFL ITP 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, B2, and B1 for the final round of judgments. The results are presented in raw scores, which is the metric that the panelists used. Each panel-recommended cut score is computed by taking the mean of the panelists' individual recommendations. The Round-3 means were rounded to the next highest whole numbers to produce the final recommended cut scores. 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.[2] 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 ITP Listening Comprehension. Table 2 summarizes the results of the standard setting for each round of judgments. The recommended cut score for A2 was consistent across Rounds 1 and 2, and increased in Round 3. The recommended cut score for B2 also was consistent across Rounds 1 and 2, but decreased in Round 3. The B1 recommended cut score was located approximately 12 points above the A2 cut score and 13 points below the B2 cut score. The standard deviation (SD) of judgments for A2 decreased across the rounds; for B2, it decreased between Rounds 1 and 2, and then increased in Round 3. The standard error of judgment (SEJ) did not exceed one point in any instance. The interpretation of the SEJ is that a comparable panel's recommended cut score (for a CEFR level) would be within one SEJ of the current recommended cut score 68% of the time and within two SEJs 95% of the time. The Round-3 SEJs are relatively small, providing some confidence that the recommended cut scores would be similar were other panels with comparable characteristics convened.

Table 2

Listening Comprehension Standard-Setting Results

Levels	A2	B2	A2	B2	A2	B1	B2
	Round 1		Round 2		Round 3		
Mean	9.9	36.4	9.8	36.5	10.2	22.4	35.3
Median	9.2	37.3	9.1	37.2	10.0	23.8	37.0
Minimum	5.1	29.3	5.8	29.4	5.8	10.5	25.0
Maximum	17.3	43.0	17.1	42.8	15.0	26.0	40.0
SD	3.2	3.9	2.8	3.7	2.0	4.0	4.1
SEJ	0.7	0.9	0.7	0.9	0.5	0.9	1.0

TOEFL ITP Structure and Written Expression. Table 3 summarizes the results of the standard setting for each round of judgments. The recommended cut score for A2 increased across the three rounds. The recommended cut score for B2 was consistent across the three rounds. The B1 recommended cut score was located approximately 12 points above the A2 cut score and 10 points below the B2 cut score. The standard deviation (SD) of judgments for A2 decreased across the rounds; for B2, it decreased from Round 1 to Round 2, and then remained the same for Round 3. The standard error of judgment (SEJ) was less than one point in all instances. The Round-3 SEJs are relatively small, providing some confidence that the recommended cut scores would be similar were other panels with comparable characteristics convened.

Table 3

Structure and Written Expression Standard-setting Results

	A2	B2	A2	B2	A2	B1	B2
Levels	Round 1		Round 2		Round 3		
Mean	6.7	30.1	7.0	30.1	7.5	19.4	29.8
Median	6.7	30.1	7.0	30.1	8.0	20.0	30.0
Minimum	2.8	22.9	3.3	22.9	3.5	14.0	22.9
Maximum	14.0	36.3	14.0	35.6	11.0	24.0	35.6
SD	3.1	3.2	2.8	2.9	1.9	2.3	2.9
SEJ	0.7	0.8	0.7	0.7	0.5	0.6	0.7

TOEFL ITP Reading Comprehension. Table 4 summarizes the results of the standard setting for each round of judgments. The recommended cut score for A2 was consistent between Rounds 1 and 2, and then increased in Round 3. The recommended cut score for B2 decreased across the rounds. The B1 recommended cut score was located approximately 15 points above the A2 cut score and 15 points below the B2 cut score. The standard deviation (SD) of judgments for A2 and B2 decreased across the rounds. The standard error of judgment (SEJ) was less than one point in all instances. The Round-3 SEJs are relatively small, providing some confidence that the recommended cut scores would be similar were other panels with comparable characteristics convened.

Table 4

Reading Comprehension Standard-Setting Results

	A2	B2	A2	B2	A2	B1	B2
Levels	Round 1		Round 2		Round 3		
Mean	7.0	38.5	7.1	38.0	7.6	22.1	37.1
Median	6.4	39.6	7.0	39.1	7.6	21.5	38.0
Minimum	2.9	32.3	3.5	31.7	4.0	17.5	30.0
Maximum	11.8	43.4	11.2	42.4	11.0	27.0	41.2
SD	2.8	3.7	2.4	3.4	2.2	3.2	3.3
SEJ	0.7	0.9	0.6	0.8	0.5	0.7	0.8

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, i.e., panelists understood the purpose of the standard-setting study; the steps they were to follow to make their judgments; etc. Table 5 summarizes the panel's feedback regarding the general process. All panelists *strongly agreed* or *agreed* that the premeeting activities were useful, that they understood the purpose of the study, that the 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. No panelists indicated *disagree* or *strongly disagree*.

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. The definition of the JQCs and their own professional experience were the most influential; 15 panelists reported that both were *very influential*. Ten reported that the between-round discussions were *very influential*, and eight reported that the discussions were *somewhat influential*. Most panelists (13) reported that the cut scores of other panelists were *somewhat influential*.

Table 5

Feedback on Standard-Setting Process

	Strongly agree		Agree	
	N	%	*N*	%
The premeeting activities were useful preparation for the study.	14	78%	4	22%
I understood the purpose of this study.	13	72%	5	28%
The instructions and explanations provided by the facilitators were clear.	17	94%	1	6%
The training in the standard-setting method was adequate to give me the information I needed to complete my assignment.	13	72%	5	28%
The explanation of how the recommended cut scores are computed was clear.	12	67%	6	33%
The opportunity for feedback and discussion between rounds was helpful.	14	78%	4	22%
The process of making the standard-setting judgments was easy to follow.	12	67%	6	33%

Panelists were also asked to indicate their level of comfort with the final cut-score recommendations; Table 6 summarizes these results. All panelists reported they were either *very comfortable* or *somewhat comfortable* with the recommended cut scores for the three sections. Thirteen panelists reported being *very comfortable* with the cut scores for Listening Comprehension and Structure and Written Expression. Ten reported being *very comfortable* with the cut scores for Reading Comprehension. No panelists indicated *somewhat uncomfortable* or *very uncomfortable*.

Table 6

Comfort Level with the Recommended Cut Scores for TOEFL ITP

	Very comfortable		Somewhat comfortable	
	N	%	*N*	%
Listening Comprehension	13	72%	5	28%
Structure and Written Expression	13	72%	5	28%
Reading Comprehension	10	56%	8	44%

Conclusions

The purpose of this standard-setting study was to recommend cut scores (minimum scores) for TOEFL ITP Listening Comprehension, Structure and Written Expression, and Reading Comprehension sections that correspond to the A2, B1, and B2 levels of the CEFR. A variation of 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 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.

Table 7 presents the Round-3 (final) recommended cut scores for each test section in raw- and in scaled-score metrics. The reporting scale for TOEFL ITP Listening Comprehension ranges from 31 to 68 scaled points; for Structure and Written Expression, it ranges from 31 to 68; and for Reading Comprehension, it ranges from 31 to 67 scaled points. The A2 cut scores for Reading Comprehension and for Structure and Written Expression were very low, eight raw points each, which corresponds to 31 and 32 scaled points, respectively. These results suggest that the panel, overall, believes that these test sections pose a significant challenge for A2-level candidates. This is not surprising, given the panel's definition of the just qualified A2 candidate for these two English-language skills. The A2 JQC for Structure and Written expression was expected to recognize and use *simple* and *routine* structures, but still likely to make systematic errors; and was expected to understand and use sufficient vocabulary for *basic everyday* needs. The panelists commented that the questions on the Structure and Written Expression section exceeded these expectations.

The just qualified A2 candidate for Reading Comprehension was expected to understand *short* (1–2 paragraphs) of *simple* text that are on *familiar* topics (e.g., notes, emails, letters); to locate *explicit basic* information about daily or *everyday* needs; and to *sometimes grasp* the probable meaning of unfamiliar words in simple, short texts on familiar topics. The panelists commented that the passages on the Reading Comprehension section were not simple, short, or about everyday needs.

Table 7

Round-3 (Final) Recommended Cut Scores

Levels	A2		B1		B2	
	Raw	Scaled	Raw	Scaled	Raw	Scaled
Listening Comprehension	11	38	23	47	36	54
Structure and Written Expression	8	32	20	43	30	53
Reading Comprehension	8	31	23	48	38	56

The responses to the end-of-study evaluation survey support the quality of the standard-setting implementation (evidence for procedural validity). All panelists *strongly agreed* or *agreed* that the premeeting activities were useful; that they understood the purpose of the study; that the 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. Procedural evidence for validity reinforces the reasonableness of the recommended cut scores.

Setting Final Cut Scores

The 18 educators were responsible for recommending cut scores. Policymakers consider the recommendation, but are responsible for setting the final cut scores (Kane, 2002). In the context of the TOEFL ITP, policymakers may represent colleges and universities, English-language programs, and other agencies that use the test scores, for example, to place students into English-language programs, to measure students' progress throughout those programs, and to assess students' end-of-program English-language proficiency. 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* SEM for TOEFL ITP Listening Comprehension is 2.04, for Structure and Written Expression it is 2.51, and for Reading Comprehension it is 2.28.

The SEJ allows policymakers to consider the likelihood that the current recommended cut score (for each CEFR level) 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 cut scores consistent with the current cut scores. The larger the SEJ, the less likely the recommended cut scores would be reproduced by another panel. 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 ITP test sections. In all cases, the SEJ resulted in an average scaled score change less than one-half of the scaled SEM.

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. Raising a cut score reduces the likelihood of a false positive decision, but increases the likelihood of a false negative decision. The converse is true when a cut score is lowered. Policymakers need to consider which decision error it is more important to minimize.

Postscript

The current standard-setting study focused on recommending cut scores for TOEFL ITP Listening Comprehension, Structure and Written Expression, and Reading Comprehension

sections that correspond to the A2, B1, and B2 levels of the CEFR. A total *scaled* test score for TOEFL ITP is computed by converting each *raw* section score to its scaled-score equivalent, summing the three *scaled* sections scores and multiplying that sum by ten-thirds. Using the scaled recommended cut scores from Table 7, the scaled total cut scores for A2, B1, and B2 are: 337, 460, and 543. Subsequent to this study, several ministries of education expressed the need to have a TOEFL ITP recommended cut score for the C1 level of the CEFR. The C1 level reflects the beginning of Proficient User band. Because this the C1 level was not part of the current study, we estimated a C1 cut score using information from a previous standard-setting study mapping TOEFL PBT to the CEFR and analyses of test-taker responses used to create concordance tables between scores on TOEFL PBT and TOEFL CBT (computer-based test) and between scores on TOEFL CBT and TOEFL IBT (internet-based test). The estimated TOEFL ITP C1 cut score is 627 scaled points.

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

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

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., & 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., & 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] Panelists were divided into diverse small groups based on their experience. Panelist grouping was changed for subsequent JQC discussions to provide the opportunity for a more robust exchange of ideas.

[2] 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 and Katz, in press).

List of Appendices

Appendix A

Panelists' Affiliations

Anne Alibert	Institut National Polytechnique, France
Qatip Arifi	South East European University, Macedonia
Donna M. Brinton	The University of California at Los Angeles, USA
Maureen H. Burke	The University of Iowa, USA
Philip E. Cary	Universidad Santo Tomas, Chile
María Isabel Freyre	Instituto Cultural Argentino Norteamericano, Argentina
Marinela Garcia	Universidad Politécnica de Madrid, Spain
Diana Kartika Jahja	The Indonesian International Education Foundation, Indonesia
Zhang Jisheng	East China Normal University, China
Marjorana Karathanasis	Collegio San Carlo Scientific Lyceum, Italy
Alexis A. Lopez	Universidad de los Andes, Columbia
Susan Luther	Ohm University, Germany
Ahmad Y. Majdoubeh	Arab Open University, Kuwait
Kevin Metz	ESC Clermont, France
Gerardo Agudelo Pulido	Rosario University, Columbia
Miyuki Sasaki	Nagoya Gakuin University, Japan
Angelika Thorman	LTS Language and Testing Service, Germany
Jirada Wudthayagorn	Maejo University, Thailand

Appendix B

Agenda

Day 1: Tuesday, July 12
Start: 8:30 a.m. Finish: 5:30 p.m.

Registration and 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 and practice on standard-setting method, and training evaluation

Round 1 judgments.* Levels A2 and B2

Break

Round 1 feedback and discussion, and Round 2 judgments

Adjourn for the Day

Day 2: Wednesday, July 13
Start: 8:30 a.m. Finish: 5:30 p.m.

Sign in and receive materials

Round 2 feedback and discussion for Listening Comprehension

Round 3 judgments: Levels A2, B1, and B2

* There are three rounds of judgments. B1 judgments occur in Round 3.

Structure and Written Expression: Review and discuss

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

Lunch

Round 1 judgments

Break

Round 1 feedback and discussion, and Round 2 Judgments

Break

Round 2 feedback and discussion, and Round 3 judgments

Reading Comprehension: Review and discuss

Develop JQC definition for CEFR Level B1

Adjourn for the day

Day 3: Thursday, July 14
Start: 8:30 a.m. Finish: 4:00 p.m.

Sign in and receive materials

Develop Reading Comprehension JQC definitions for CEFR Levels A2 and B2

Round 1 Judgments

Break

Round 1 feedback and discussion, and Round 2 judgments

Lunch

Round 3 judgments

Final evaluations

End of study

Appendix C

Panel-developed Just Qualified Candidate Descriptions

Listening Comprehension

A2 Level

- Can identify/recognize general topic(s), and words and phrases, of most immediate priority in a familiar context.
- Can understand slowly and clearly spoken English.
- Can understand basic directions and instructions but may require repetition.

B1 Level

- Can understand main ideas of familiar topics spoken clearly.
- Can understand main ideas of a lecture in outline form on familiar topics when clearly spoken and visual cues are present.
- Can understand main ideas of short narratives – factual, concrete vocabulary, on familiar topics when clearly articulated (e.g., news, radio, documentaries).
- Can understand simple directions and instructions, delivered clearly.
- Can follow straightforward, concrete everyday life conversation spoken clearly.
- Can understand some key details but has difficulty with phrasal verbs/idioms.

B2 Level

- Can infer at the word and sentence level and can infer some content, some details when logically sequenced and signposted.
- Can understand a range of topics: academic, unfamiliar, abstract, when delivered at normal speed.
- Can follow main ideas and structure of complex arguments and complex language structures of somewhat familiar topics.
- Can recognize tone and attitude in most contexts.

Structure and Written Expression

A2 Level

- Can recognize and use simple and routine structures and may make systematic errors.
- Can understand and use sufficient vocabulary for basic everyday needs.

B1 Level

- Can understand simple and some compound/complex sentences associated with more predictable context.
- Can recognize and use simple and continuous tenses (past, present, future) in familiar contexts but still makes some mistakes (some of which may be influenced by L1), e.g., subject-verb agreement.
- Can utilize everyday familiar vocabulary; recognizes general academic vocabulary, but may make major errors with unfamiliar words and more complex topics.
- Can identify most parts of speech/morphology in simple sentences.

B2 Level

- Has good grammatical control of most simple and complex structures; may make mistakes that do not interfere with communication and comprehension.
- Has good range of vocabulary on most general topics and in his/her specialized field with some lexical gaps.
- Can use words accurately most of the time and mistakes rarely interfere with communication and comprehension.

Reading Comprehension

A2 Level

- Can understand short (1-2 paragraphs), simple texts on familiar topics (e.g., notes, emails, letters).
- Can locate explicit basic information regarding daily (everyday) needs.
- Can sometimes grasp the probable meaning of unfamiliar words in simple, short texts on familiar topics.

B1 Level

- Can understand the main ideas, supporting details, and conclusions of straightforward clearly written texts of moderate length (1-3 pgs) on subjects that are familiar.
- Can sometimes recognize different writing styles (narrative, expository, argumentative).
- Can use strategies to infer the meaning of unknown vocabulary in familiar topics (from context clues); still have difficulties with idioms/slang/phrasal verbs.
- Can scan for specific (relevant) information for a defined purpose.

B2 Level

- Can understand the main ideas, supporting details, and conclusions of a variety of moderately complex and longer texts.
- Can often identify author's purpose, tone, attitude, and intention.
- Can often infer meaning of unfamiliar vocabulary/content, and grasp implicit meaning.
- Can understand frequently used idioms, slang, and phrasal verbs.
- Can scan quickly.