

Compendium Study

Mapping *TOEIC*® and *TOEIC Bridge*™ Test Scores to the Common European Framework of Reference

Richard J. Tannenbaum and E. Caroline Wylie

September 2013

By themselves, scores on English-language proficiency tests may not provide sufficient information to score users. A score of 130 on a measure of speaking skills, for example, does not readily convey what a test taker with that score may be able to do outside of the test. Is it reasonable to anticipate that test takers with such a score can express their opinions, providing multiple reasons to support their opinions? Are they able to speak with sufficient clarity and fluency so that they are readily understood by a native speaker? How is a score user to know? This is a fundamental issue in assessment: understanding the meaning of scores.

One way of providing meaning to English-language assessment scores is associating them with operational definitions or descriptions of English-language skills (Tannenbaum, 2011). Kane (2012) reinforced this point in a discussion of argument-based validity: “We can add meaning to the scores by referencing them to . . . performance levels, benchmark performance levels, or achievement levels (e.g., as in NAEP or CEFR)” (p. 8). By associating a test score with an operational definition of language proficiency, the meaning of the score becomes more evident as the score implies that the test taker is likely able to demonstrate the described set of skills.

The purpose of our study was to map *TOEIC*® test scores (Listening, Reading, Speaking, and Writing) and *TOEIC Bridge*™ test scores (Listening and Reading) onto the Common European Framework of Reference (CEFR; Council of Europe, 2001). The CEFR describes a progression of language proficiency in reading, writing, speaking, and listening on a 6-level scale clustered in three bands: A1–A2 (basic user), B1–B2 (independent user), and C1–C2 (proficient user). The different levels are described by sets of “can-do” statements for each skill area; holistic (global) descriptors are also provided. For example, overall, an individual at the B2 level possesses the following qualities:

Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation.

Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party.

Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages or various options. (Council of Europe, 2001, p. 24)

A recognized approach for mapping test scores to language frameworks, such as the CEFR, is through the process of standard setting (Council of Europe, 2009; Kaftandjieva, 2010; Martyniuk, 2010; Tannenbaum & Wylie, 2008). The fundamental objective of standard setting in this context is to identify minimally acceptable scores (cut scores) needed to enter the different CEFR levels of interest. Two standard-setting approaches were implemented: a variation of an Angoff method (Brandon, 2004; Cizek & Bunch, 2007; Plake & Cizek, 2012; Tannenbaum & Katz, 2013) for the Listening and Reading tests, which include selected-response items; and a variation of a Performance Profile method (Morgan, 2004; Perie & Thurlow, 2012; Zieky, Perie, & Livingston, 2008) for the Speaking and Writing tests, which include constructed-response (performance-based) tasks. The methods and their implementation are described on the next page.

Methods

Panelists

Twenty-two experts representing 10 countries served on the standard-setting panel. They had been selected for their experience with English-language instruction, learning, and testing in the workplace and, in particular, for their knowledge of and experience with the TOEIC and *TOEIC Bridge* tests. They were selected to represent a range of European countries where the tests are administered and where the CEFR levels are used to inform decisions such as placement, training, and advancement. Table 1 provides a summary of the demographic characteristics of the experts.

Table 1
Demographics of the Experts

Category		Number ^a
Gender	Female	19
	Male	3
Position	ESL teacher	17
	Administrator of ESL	4
	ESL assessment	5
	Educational consultant	5
Country	Belgium	2
	France	6
	Germany	2
	Greece	3
	Hungary	3
	Italy	1
	Malta	1
	Poland	2
	Russia	1
	Slovakia	1

Note. ESL = English as second language.

^aExperts provided multiple responses to the position category, so the total number exceeds 22.

Procedures

Prior to the study. Although the panelists reported being familiar with the CEFR, they were asked to complete a refamiliarization assignment before attending the standard-setting study. The assignment served two related goals. The first was to reaffirm their understanding of the CEFR levels in general. The second was to begin the process of them thinking about the specific minimal set of skills needed to enter each level. Each CEFR level defines a range of skills; but, the focus of the study is to identify the TOEIC test score (cut score) needed to enter a level. The functional question is, “What is the lowest acceptable score corresponding to a CEFR level?” The standard-setting process, described below, focuses on identifying the score equivalent to this point of entry, and so the skills expected of a candidate just entering a level need to be defined. This candidate is referred to as a *just qualified candidate* (JQC).

The panelists were asked to review selected tables from the CEFR for each language modality (listening, reading, speaking, writing) and to note key characteristics or indicators from the tables that described an English-language learner (ELL) with *just enough skills* to enter each of the CEFR levels (i.e., the JQC). The tables were selected to provide the panelists with a broad understanding of what candidates are expected to be able to do for each of the language modalities. As they completed this prestudy assignment, they were asked to consider what distinguishes a candidate with just enough skills to be considered performing at a CEFR level from a candidate with almost, but not quite enough skills to be performing at that level. For example, they were asked to consider what the least able C2 speaker can do that the highest performing C1 speaker cannot do, what the least able C1 speaker can do that the highest performing B2 speaker cannot do, and so on. The assignment was intended as part of a calibration of the panelists to a shared understanding of the minimum requirements to enter each of the CEFR levels.

During the study. Panelists completed the standard-setting process first for the *TOEIC* test and then for the *TOEIC Bridge* test. Time was spent coming up with an agreed upon definition of the JQC for each of the targeted CEFR levels. The intent of the TOEIC test mapping was to associate scores with each of the six CEFR levels. The mapping for the *TOEIC Bridge* test focused only on the A1, A2, and B1 levels. Although both the TOEIC and *TOEIC Bridge* tests measure listening and reading, the *TOEIC Bridge* test is intended for test takers with less developed English-language skills.

The panelists worked in three small groups to define the skills of a JQC for A2, B2, and C2 levels; this was done separately for the listening, reading, speaking, and writing tests. Panelists referred to their prestudy assignments and to the CEFR tables (provided at the study). A whole-panel discussion followed and an agreed upon definition was established for each level. Definitions of a JQC for A1, B1, and C1 levels were accomplished through whole-panel discussion, using the A2, B2, and C2 descriptions as boundary markers; as before, the panelists also referred to their prestudy assignments and the relevant CEFR tables. The JQC definitions served as the frame of reference for the standard-setting judgments; that is, panelists were asked to consider the test items and tasks in relation to these definitions.

Standard setting for listening and reading. A variation of an Angoff approach (e.g., Tannenbaum & Katz, 2013) was implemented for reading and listening tests. These tests include selected-response items. Panelists were trained in the method and then given an opportunity to practice making judgments. The practice judgments were summarized, and the panelists were asked to share their judgment rationales. The practice and discussion helped ensure that they understood how to complete the standard-setting judgment task. At this point, panelists were asked to sign a training evaluation form confirming their understanding and readiness to proceed, which all panelists did. The standard-setting process was completed first for the listening test and then for reading test. In either case, the same method was followed. The same procedures were applied to the TOEIC and *TOEIC Bridge* tests; the only difference was that for the *TOEIC Bridge* test the focus was on the A1, A2, and B1 CEFR levels. The text below illustrates the process for the TOEIC tests.

Panelists completed three rounds of judgments, with feedback and discussion between rounds. For Round 1, panelists judge the probability that a JQC would answer each item correctly. The Round 2 and Round 3 judgments were made at the test level. In order to reduce the cognitive demand and fatigue associated with setting cut scores for all levels of the CEFR, panelists initially focused on the A2, B2, and C2 levels. Once the cut scores for these levels were identified (at the end of Round 3), panelists located the A1, B1, and C1 cut scores in relation to the A2, B2, and C2 cut scores.

Specifically, in Round 1, each panelist independently judged the probability (likelihood) that a JQC would answer each item correctly. Judgments were rendered within items across levels; that is, a panelist judged the probability for a JQC A2, for a JQC B2, and for a JQC C2 for one item before repeating for the next item. Panelists used the following judgment scale (expressed as probabilities): 0, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 95, 100. The panelist believed that the higher the probability, the easier an item was for a JQC. (For *TOEIC Bridge* items, the JQCs were for A1, A2, and B1.)

The sum of each panelist's cross-item judgments represents his or her recommended cut score for a CEFR level. Each panelist's recommended cut score was provided to the panelist. The panel's average (recommended cut score) and the highest and lowest cut scores were presented to the panel to foster discussion. Panelists were asked to share their judgment rationales. As part of the feedback and discussion, item difficulty information (p -values, the percentage of test takers from previous administrations answering the item correctly) were shared. In addition, p -values were calculated for candidates scoring at or above the 75th percentile on the test (i.e., the top 25% of test takers) and for candidates at or below the 25th percentile (i.e., the bottom 25% of test takers). Examining item difficulty for the top 25% and bottom 25% of test takers gives panelists a better understanding of the relationship between overall language ability for that modality (e.g., listening) and each of the items. The partitioning, for example, enabled panelists to see any instances where the same item was comparably difficult regardless of test takers' overall language ability or where an item was found to be particularly challenging or easy for test takers at the different ability levels.

Before making their Round 2 judgments, panelists considered the rationales of their peers and the item difficulty information. For this round, judgments were made not at the item level but at the overall test level; that is, panelists were asked to consider if they wanted to recommend a different

test-level cut score for the A2, B2, and C2 levels. (For *TOEIC Bridge* items, overall judgments were for A1, A2, and B1 levels.) The transition to test-level judgments introduced a shift from discrete items to the overall language construct of interest. This holistic approach seemed more relevant and appropriate than did deconstructing the construct through another series of item-level judgments. Panelists had no difficulty with the holistic approach; this approach had also been used in a previous CEFR mapping study (Tannenbaum & Wylie, 2005). After making their second round of judgments, feedback similar to that provided in Round 1 was shared; in addition, the percentage of test takers classified into each of the three levels was presented and discussed. (The Round 2 recommended cut scores were applied to score distributions from the previous administrations to compute the classifications.) The panelists then had a final opportunity to change their test-level recommended cut scores. The average across panelists is the recommended cut score for a level. These final judgments were shared with the panelists. Exclusively for TOEIC items, they were then asked to “slot in” the A1, B1, and C1 level cut scores. Specifically, they reviewed and discussed all the JQC definitions (A1 through C2) and the current set of cut scores, and then each panelist identified the test-level A1, B1, and C1 cut scores. The panelists’ cut-score recommendations were averaged.

Standard setting for writing and speaking. A variation of a Performance Profile approach (e.g., Zieky et al., 2008) was implemented for the speaking and writing tests (exclusive to the TOEIC tests). These tests include constructed-response (performance-based) tasks. Consistent with the model followed for the *TOEIC*® Listening and Reading tests, three rounds of judgments occurred, with feedback and discussion informed by data (average task scores, partitioned as described above, and classification information). Panelists were trained in the method, practiced making judgments, discussed those judgments, and then signed off on their readiness to proceed. The standard-setting process was completed first for the writing test and then for the speaking test.

A performance profile consists of a test taker’s scored response to each test task, and the total score earned (a weighted sum of the task scores). The writing test generates eight task scores and the speaking test 11 task scores. Writing and speaking profiles from more than 20 test takers were sampled for use in the study. The samples represented a range of total test scores, from 6 to 25.4 raw points for the writing test and from 5.8 to 23.7 raw points for the speaking test. The writing responses were presented in a binder for the panelists, and the speaking responses were broadcast to the panelists. The profiles were presented in increasing total-score order. Screen shots of the test tasks and scoring rubrics were provided to each panelist. The fundamental standard-setting judgment was for each panelist to review the performance profiles and to decide on the total score most likely to be earned by JQCs at the A2, B2, and C2 levels. As before, the A1, B1, and C1 cut scores were slotted, after Round 3, in relation to the A2, B2, and C2 cut scores.

Adjustments based on perceived alignment to the CEFR. A formal alignment between the TOEIC tests, *TOEIC Bridge* tests, and the CEFR was not part of this study. An alignment process includes panelists judging the extent of overlap between the test content and the skills covered by the CEFR. Approaches to alignment vary, but the common goal is to collect evidence that documents the extent to which there is convergence between the two sources (test and CEFR). During the standard-setting process, some panelists expressed reservations about the extent to which the TOEIC tests

sufficiently measured the highest levels (C1 and C2) of the CEFR. (Concerns were not raised about alignment for the *TOEIC Bridge* test.) In order to account for this, cut scores were only computed if at least 67% of the panel (15 of 22 panelists) offered cut-score recommendations for these levels.

Results

Table 2 presents the final set of scaled recommended cut scores for each of the TOEIC tests: Listening, Reading, Speaking, and Writing. The missing numbers for C2 and for C1 of the Reading test are due to less than 67% of the panelists offering cut scores for these levels. Too few panelists thought that any of the tests sufficiently measured the C2 level of the CEFR. The majority of the panel similarly did not believe that the reading test measured the C1 level. For the Writing and Speaking tests, the C1 cut score was 200, which is the maximum scaled score obtainable. This indicates that these tests really do not adequately measure the C1 level. A cut score signifies the score needed to enter a level; if the highest score possible is needed, that means that the test is not challenging enough for test takers at that level. The same applies to the Listening test for the C1 level. Although 490 is not the maximum, it is so very close to it that the same conclusion is warranted: The Listening test is not challenging enough for test takers at the C1 level.

Table 2

Scaled Recommended Cut Scores—TOEIC

Level	Listening (Max. 495 points)	Reading (Max. 495 points)	Speaking (Max. 200 points)	Writing (Max. 200 points)
A1	60	60	50	30
A2	110	115	90	70
B1	275	275	120	120
B2	400	385	160	150
C1	490	-	200	200
C2	-	-	-	-

Table 3 presents the final set of scaled recommended cut scores for each of the *TOEIC Bridge* tests: Listening and Reading. Panelists recommended cut scores for all three levels of the CEFR. The cut scores to enter B1 were near the maximum score, suggesting that test takers with listening and reading skills likely greater than entry-level B1 may not be challenged sufficiently by the *TOEIC Bridge* test and so, perhaps, should be directed to take the TOEIC test as a more appropriate measure.

Table 3

Scaled Recommended Cut Scores—TOEIC Bridge

Level	Listening (Max. 90 points)	Reading (Max. 90 points)
A1	46	46
A2	64	70
B1	84	86

At the conclusion of the standard-setting study (TOEIC and *TOEIC Bridge* tests), the panelists completed an evaluation form. This form served the purpose of collecting information about the perceived quality of the standard-setting process. Panelists were asked to rate the clarity with which various aspects of the study were presented and were asked to indicate overall their level of comfort with the full set of recommended cut scores. The majority of panelists indicated that the homework assignment was useful preparation, that the purpose of the study and instructions were clear, training was sufficient, and the feedback/discussion process was helpful. Additional prompts asked panelists about what was most influential in their decision-making process. The definition of the JQC and the panelists' own professional experience were the two most influential factors. Finally, each panelist was asked to indicate his or her level of comfort with the final results. The most frequent response was *very comfortable*.

Discussion

The difficulty of linking test scores to the CEFR should not be underestimated. The CEFR, according to Weir (2005), does not provide sufficient information about how contextual factors affect performance across the levels or adequately delineate how language develops across the levels in terms of cognitive or meta-cognitive processing. Milanovic (2009) reinforced "that the CEFR itself is deliberately underspecified and incomplete" (p. 3). As he pointed out, the can-do statements were intended to be illustrative of the nature of the levels rather than precise definitions of the levels. This may lead to difficulties in consistency interpreting differences across the CEFR levels (Alderson et al., 2006; Papageorgiou, 2010). Some of this difficulty was evident during the panelist discussions of the CEFR when developing the just-qualified descriptions; panelists noted that the descriptive language of the CEFR was not consistently applied across the levels, making it more difficult for them to differentiate among the levels.

The linking difficulty, however, also is a function of the tests, specifically with the extent to which the tests are aligned with the CEFR. It is more likely that tests developed to map to the CEFR would pose less of a linking challenge than tests not so designed, relying solely on a post-hoc approach, as was the present case. Although tests considered in this study measured the four major language skills, all covered by the CEFR, the items and tasks on the tests were not specifically developed to measure these skills necessarily as depicted by the CEFR, and, in fact, the TOEIC and *TOEIC Bridge* Listening and Reading tests existed before the CEFR. Although this did not preclude recommending cut scores for some of the levels, it clearly was the reason for the challenges faced with the C1 and C2 levels. The value of using level descriptors to inform test development, thereby increasing alignment and the potential meaningfulness of cut scores, was noted by Bejar, Braun, and Tannenbaum (2007).

Although not all targeted CEFR levels were mapped, there was positive evidence of procedural validity. The majority of panelists for each test reported that they were adequately trained and prepared to conduct their standard-setting judgments and that the standard-setting process was easy to follow. Panelists reported that the definition of the JQC most influenced their judgments and that they were able to use their professional experience to inform their judgments. Furthermore, the majority of panelists reported that they were very comfortable with the recommended cut scores.


Procedural validity is an important criterion against which to evaluate the quality of the standard-setting process (Cizek & Bunch, 2007; Kane, 2001; Tannenbaum & Katz, 2013).

External validity evidence is also desirable and most often takes the form of convergence with other sources of information (Hambleton & Pitoniak, 2006; Kane, 2001; Tannenbaum & Katz, 2013). In the present case, for example, convergent evidence could be obtained from ESL teacher ratings of their students' English-language proficiency in terms of the CEFR (Council of Europe, 2009). Although a convergence of evidence would lend further support of the reasonableness of the panel-based cut scores, the meaning of a divergence of evidence is less clear, given that there is no true cut score. "Differences in results from two different procedures would not be an indication that one was right and the other wrong; even if two methods did produce the same or similar cut scores, we could only be sure of precision, not accuracy" (Cizek & Bunch, 2007, p. 63). With this in mind, the cut scores from this study should be considered estimates; they are not absolutes. Potential users of these cut scores are advised to consider their specific needs and circumstances as well as other relevant information that was not part of this study but may be germane to determinations of the English-language proficiency of their test takers. It is reasonable for users to adjust these recommended cut scores, considering, for example, the standard error of the tests (Geisinger & McCormick, 2010), to better accommodate their needs.

References

- Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S., & Tardieu, C. (2006). Analysing tests of reading and listening in relation to the common European framework of reference: The experience of the Dutch CEFR construct project. *Language Assessment Quarterly: An International Journal*, 3(1), 3–30.
- 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: Intellectual growth and standard setting* (pp. 1–30). Maple Grove, MN: Journal of Applied Metrics 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.
- Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment (CEFR)*. Cambridge, United Kingdom: Cambridge University Press.
- Council of Europe. (2009). *Relating language examinations to the common European framework of reference for languages: Learning, teaching, assessment (CEFR)*. Retrieved from http://www.coe.int/t/dg4/linguistic/manuel1_en.asp#P75_5622

- Geisinger, K. F., & McCormick, C. M. (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.
- 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.
- Kaftandjieva, F. (2010). *Methods for setting cut scores in criterion-referenced achievement tests: A comparative analysis of six methods with an application to tests of reading in EFL*. Arnhem, The Netherlands: CITO.
- Kane, M. (2012). Validating score interpretations and uses. *Language Testing*, 29, 3–17.
- Martyniuk, W. (Ed.). (2010). *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual*. Cambridge, United Kingdom: Cambridge University Press.
- Milanovic, M. (2009). Cambridge ESOL and the CEFR. *Research Notes*, 37, 2–5.
- Morgan, D. L. (2004, June). *The performance profile method (PPM): A unique standard setting method as applied to a unique population*. Paper presented at the annual meeting of the Council of Chief State School Officers, Boston, MA.
- Papageorgiou, S. (2010). Investigating the decision-making process of standard setting participants. *Language Testing*, 27(2), 261–282.
- Perie, M., & Thurlow, M. (2012). Setting achievement standards on assessments for students with disabilities. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 347–377). New York, NY: Routledge.
- Plake, B. S., & Cizek, G. J. (2012). Variations on a theme: The modified Angoff, extended Angoff, and yes/no standard setting methods. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 181–199). New York, NY: Routledge.
- Tannenbaum, R. J. (2011). *Relating English language test scores to definitions of language proficiency: A discussion of methods to set standards*. Paper presented at the annual meeting of the Japan Association of College English Teachers, Fukuoka, Japan.
- 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. (2005). *Mapping English-language proficiency test scores onto the Common European Framework* (TOEFL Research Report No. RR-80). Princeton, NJ: Educational Testing Service.
- 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 Report No. 06). Princeton, NJ: Educational Testing Service.
- Weir, C. J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing*, 22, 281–300.
- 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.