Research Report

Mapping the Test de Français International™ onto the Common European Framework of Reference

Patricia A. Baron

Richard J. Tannenbaum

September 2010

ETS RM-10-12

Mapping the *Test de français international*™ onto the Common European Framework of Reference

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

September 2010

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: Irvin Katz and Michael Zieky

Abstract

The Common European Framework of Reference (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). This study linked scores on the *Test de français international*™ (TFI™) to four levels of the CEFR: A2, B1, B2, and C1. The TFI measures reading and listening skills in French, and consists of multiple-choice questions. A combination of a modified Angoff standard setting approach, and a holistic judgment was followed to identify the TFI scores linked to the CEFR levels. Sixteen language experts from seven countries served on the standard setting panel.

Key words: CEFR, TFI, standard setting, cut scores

Acknowledgments

We extend our sincere appreciation to Laure Mounier, our colleague from the ETS Global office in Paris, for her facilitation of the standard setting meeting. We also offer our gratitude to our other ETS Global colleagues, Zeineb Mazouz and Françoise Azak. Zeineb rapidly, yet calmly, translated between French and English during the study. Françoise organized the accommodations, meeting logistics, and materials. Finally, we thank our colleague Craig Stief for his work on all the rating forms, analysis programs, and on-site scanning.

Table of Contents

List of Tables

Page

Background

The Common European Framework of Reference (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 purpose of this study was to conduct a standard setting study to link scores on the *Test de français international*™ (TFI™) to the CEFR.

The TFI measures listening and reading skills in French. It is designed for test takers whose native language is not French, that is, French language learners. The test measures general listening and reading skills that people may need to use in daily life and across a range of workplace settings (TFI Examinee Handbook, ETS, 2008). Each test section (Listening and Reading) includes 90 multiple-choice questions. The TFI was neither specifically designed to measure the range of proficiency levels addressed by the CEFR nor to measure listening and reading 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 represented on the test. Therefore, before conducting the standard setting study, ETS testing experts identified the specific CEFR levels that were most clearly aligned with the TFI Listening and Reading sections. Each section was judged to address A2 through C1. The process of standard setting focused only on those levels, and a separate set of cut scores was constructed for each of the two sections, Listening and Reading.

Method

A modified Angoff standard setting approach (Cizek & Bunch, 2007; Zieky, Perie, & Livingston, 2008), coupled with a holistic judgment, was followed to identify the TFI scores linked to the A2 through C1 levels of the CEFR. The specific implementation of this approach followed the work of Tannenbaum and Wylie (2008). In that study, cut scores were constructed linking *Test of English for International Communication*™ (TOEIC®) to the CEFR. In the current study, the modified Angoff approach was employed in Round 1; in Rounds 2 and 3 panelists made holistic judgments on section scores (Listening and Reading). Recent reviews of

research on standard setting approaches reinforce a number of core principles for best practice: careful selection of panel members and a sufficient number of panel members to represent varying perspectives, sufficient time devoted to develop a common understanding of the domain under consideration, adequate training of 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; Cizek, 2006; Hambleton & Pitoniak, 2006). The approach used in this study adheres to these principles.

The TFI standard setting was conducted in French by a bilingual (French/English) facilitator who is experienced working with French educators. All standard setting materials were developed by the authors of this report, translated from English to French, and reviewed prior to the study with the bilingual facilitator. The two authors of this report and a second bilingual facilitator were present throughout the study. This allowed for translation of the technical and procedural questions when necessary during the implementation, in order for the authors to respond, because only a small number of panelists spoke English.

Panelists

Sixteen individuals from seven countries served on the panel. All had expertise in French language development or assessment. Nine taught French as a second language and four were either directors or heads of a language development school. Twelve had at least 10 years of experience in their current function. Because the TFI measures French language proficiency, the largest number of panelists was from France (10 of 16). The panelists were familiar with the CEFR, the TFI, and with the general population of test takers required to take the TFI. Table 1 provides a description of the self-reported demographics of the panelists. (See the Appendix for panelist affiliations.)

Pre-meeting Assignment

Prior to the standard setting study, the experts were given an assignment to review selected tables from the French version of the CEFR for Reading and Listening, and to note key characteristics or indicators from the tables that described a French language learner (candidate) with *just enough skills* to be performing at each of the targeted CEFR levels (A2, B1, B2, and C1). The tables were selected to provide the experts with a broad understanding of what learners are expected to be able to do for each of the two language areas. The experts were asked to consider what distinguishes a candidate with *just enough* skills to be considered performing at a

CEFR level from a candidate with *not quite enough* skills to be performing at that level. To help facilitate completion of the assignment, we provided the experts with descriptions of candidates believed to be at the beginning of each targeted CEFR levels for listening and reading constructed during the Tannenbaum and Wylie (2008) study. The experts were encouraged to review both sources of information and to modify the descriptions from the Tannenbaum and Wylie study, as needed, based on their own interpretation of the CEFR levels and their experiences in the field of French language development and assessment. The pre-meeting assignment was an opportunity for panelists to review relevant parts of the CEFR, and was intended as the first stage in the calibration of the experts to a shared understanding of the minimum requirements for each of the targeted CEFR levels.

Table 1

Panelist Demographics

		Number
Gender	Female	13
	Male	3
Function	Teacher	9
	Director/Head of language department or school	4
	Education consultant	1
	Project officer	1
	Language assessment expert	1
Experience	Less than 10 years	4
	10–20 years	6
	More than 20 years	6
Country	Belgium	1
	Canada	1
	France	10
	Iran	1
	Romania	1
	Russia	1
	Venezuela	1

Each expert also was provided with an opportunity to take the TFI before arriving at the standard setting study. Each expert had signed a non-disclosure/confidentiality form before having access to the test. The experience of taking the test is necessary for the experts to understand the scope of what the test measures and the difficulty of the questions on the test.

Judgment Process

During the study, the experts (panelists) defined the minimum skills needed to reach each of the targeted CEFR levels (A2, B1, B2, C1). 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 and B2 levels; this was done separately for Reading and Listening. This candidate was referred to as a *just qualified candidate* (JQC). Experts referred to their pre-study assignment and to the CEFR tables for each of the two skill areas. A whole-panel discussion occurred for each level and a final definition for each level was established. Definitions of the JQC for A1 and C1 levels were accomplished through whole-panel discussion, using the B1 and B2 descriptions as *boundary markers*. These definitions 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.

A modified Angoff approach was implemented following the procedures of Tannenbaum and Wylie (2008). 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. Then they went through three rounds of operational judgments, with feedback and discussion between rounds, for the B1 and B2 levels. In Round 1, for each test question, panelists were asked to judge the percent of 100 *just qualified candidates* (B1 and B2) who would know the correct answer. 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 skill demanded by the question and the skill possessed by JQCs, and not to factor guessing into their judgments. The panelists made their judgments for a question for each of the two CEFR levels (B1 and B2) before moving to the next question.

The sum of each panelist's cross-question judgments, divided by 100, represents the panelist's recommended cut score, i.e., the number correct across 90 questions. Each panelist's recommended cut score was provided to the panelist. The panel's average (panel's recommended

cut score), and the highest and lowest cut scores (unidentified) were compiled and presented to the panel to foster discussion. Panelists were then asked to share their judgment rationales. As part of the feedback and discussion, P+ values (percentage of test takers from a recent administration[1] who answered each question correctly), were shared. 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 experts a better understanding of the relationship between overall language ability for that TFI 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.

In Round 2, judgments were made, not at the question level, but at the overall level of the section; that is, panelists were asked to consider if they wanted to recommend a different section-level (e.g., Listening) score for B1 and/or B2. The transition to a section-level judgment places emphasis on the overall constructs of interest (i.e., Listening and Reading) rather than on the deconstruction of the constructs through another series of question-level judgments. This modification had been used in previous linking studies (Tannenbaum & Wylie 2005; 2008), and posed no difficulties for the TFI panelists. After making their second round of judgments, feedback similar to that in Round 1 was provided, but in addition, the percentage of test takers from a recent administration who would be classified into each of the two levels (B1 and B2) was presented and discussed. The panelists then had an opportunity to make a final (Round 3) section-level judgment.

The final (Round 3) judgments were compiled and shared with the panelists. They were then asked to recommend cut scores for the A2 and C1 levels. Specifically, they were asked to review the A2, B1, B2 and C1 descriptions of *just qualified candidates* and to identify the minimum section-level scores for candidates just performing at the A2 and C1 levels (Tannenbaum & Wylie, 2008). Their judgments were bounded by the now-established B1 and B2 cut scores. The panelists, as a group, discussed where to locate the A2 and C1 cut scores; but then each panelist made an individual judgment regarding the A2 and C1 cut scores. The average of the individual (A2 and C1) recommendations was computed.

Results

The first set of results summarizes the panel's standard setting judgments for the TFI Reading and Listening sections. The results are presented in raw scores, which is the metric that the panelists used. Also included is the standard error of judgment (SEJ), which indicates how close the cut score is likely to be to the current cut score for other panels of experts similar in composition to the current panel and similarly trained in the same standard setting method. This is followed by a summary of responses to an end-of-study evaluation survey, which provides evidence of process-based validity, or how well the study was conducted. (The scaled cut scores are provided in the conclusion section.)

Reading Section

Table 2 summarizes the results of the standard setting for Levels B1 and B2 for each round of judgments. The average (mean) cut score for B1 decreased at Round 2, and then increased in Round 3, to a score more consistent with the Round 1 cut score. The cut score for B2 decreased at Round 2 and increased somewhat by Round 3, but was not as high as the cut score recommendation at Round 1. For both B1 and B2, the variability among the panelists decreased over three rounds, as can be seen by the decrease in the standard deviations (SD). 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 Reading at Round 3 is less than two points for both B1 and B2 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

Reading: B1 and B2 Standard Setting Results

	Round 1		Round 2		Round 3	
Levels	B1	B2	B1	B2	B1	B2
Average	28.2	49.5	26.8	47.2	28.8	48.4
Median	26.5	49.6	25.5	48.4	28.0	46.7
Minimum	17.2	32.6	17.7	34.5	20.0	38.0
Maximum	53.9	69.8	41.2	60.0	35.0	63.0
SD	10.0	10.8	7.3	7.7	5.6	7.2
SEJ	2.5	2.7	1.8	1.9	1.4	1.8

Table 3 summarizes the results of the standard setting judgments for Reading for Levels A2 and C1. These judgments were made after the Round 3 cut scores for B1 and B2 had been presented. The recommended A2 cut score is approximately 14 raw points lower than the B1 recommendation and the C1 cut score is approximately 22 raw points higher than the B2 cut score. The SEJ for A2 and C1 levels is less than two raw points.

Table 3

Reading: A2 and C1 Standard Setting Results

Levels	A2	C1
Average	15.3	70.3
Median	15.0	70.0
Minimum	10.0	62.0
Maximum	20.0	82.0
SD	2.7	5.5
SEJ	0.7	1.4

Listening Section

Table 4 summarizes the results of the standard setting for Levels B1 and B2 for each round of judgments. The pattern of recommendations across rounds, as well as the pattern of changes in variability, is consistent with that observed for Reading. For both B1 and B2 recommended cut scores, panelists decreased their overall cut score, making it easier to enter into each level, at Round 2, and then increased them somewhat in Round 3. For Listening, the amount of increase at Round 3 did not result in recommendations as high as Round 1. The panelists' judgments converged across the three rounds of judgments, as seen in the decrease in the standard deviations. The SEJs similarly decreased across rounds. The Round 3 SEJ for B1 and B2 levels is less than two raw points.

Table 4

Listening: B1 and B2 Standard Setting Results

	Round 1		Round 2		Round 3	
Levels	B1	B2	B1	B2	B1	B2
Average	27.0	49.1	25.3	48.5	26.5	48.7
Median	26.1	49.4	25.0	48.0	25.9	47.5
Minimum	10.7	29.1	20.0	38.0	20.0	38.0
Maximum	45.3	66.5	38.0	66.0	38.0	66.0
SD	9.5	11.5	4.7	7.4	4.7	7.2
SEJ	2.4	2.9	1.2	1.8	1.2	1.9

Table 5 summarizes the results of the standard setting judgments for Listening for Levels A2 and C1. The recommended A2 cut score is approximately 13 raw points lower than the B1 recommendation and the C1 cut score is approximately 18 raw points higher than the B2 cut score. The SEJ for A2 and C1 levels is less than or equal to one raw point.

Table 5

Listening: A2 and C1 Standard Setting Results

Levels	A2	C1
Average	13.8	66.6
Median	15.0	65.5
Minimum	8.0	60.0
Maximum	18.0	75.3
SD	2.6	4.1
SEJ	0.7	1.0

End-of-Study Evaluation Survey

Panelists responded to a final set of questions addressing the procedural validity (Kane, 1994) of the standard setting process. Table 6 summarizes the panel's feedback regarding the general process. The majority of panelists *strongly agreed* or *agreed* that the pre-meeting assignment was useful, 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.

Additional questions focused on how influential each of the following four factors was in their standard setting judgment: the definition of the JQC, the between-round discussions, the cut scores of the other panelists, and their own professional experience. All panelists indicated their own professional experience was *very influential*, and the majority also indicated that each of the other three factors was *very influential*. Nonetheless, nearly a third of the panelists also indicated that the cut scores of the other panelists were only *somewhat influential*.

Table 6

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	69%	5	31%	0	0%	0	0%
I understood the purpose of this study.	14	88%	2	13%	0	0%	0	0%
The instructions and explanations provided by the facilitators were clear.	8	53%	6	40%	1	7%	0	0%
The training in the standard setting methods was adequate to give me the information I needed to complete my assignment.	11	69%	5	31%	0	0%	0	0%
The explanation of how the recommended cut scores are computed was clear.	4	27%	9	60%	2	13%	0	0%
The opportunity for feedback and discussion between rounds was helpful.	12	75%	4	25%	0	0%	0	0%
The process of making the standard setting judgments was easy to follow.	4	25%	12	75%	0	0%	0	0%

Note. Percentages are based on the number of panelists providing a response.

Finally, each panelist was asked to indicate their level of comfort with the final cut score recommendations; Table 7 summarizes these results. Fourteen of the 16 panelists reported being *very comfortable* or *somewhat comfortable* with the cut score results for Listening, with two panelists reporting that they were *somewhat uncomfortable*. All the panelists reporting being *very comfortable* or *somewhat comfortable* with the Reading cut scores, with slightly more than half of the panelists reported being *very comfortable*.[2]

Table 7

Comfort Level with Final Recommended Cut Scores

	Very Comfortable		Somewhat Comfortable		Somewhat Uncomfortable		Very Uncomfortable	
	N	%	*N*	%	*N*	%	*N*	%
Reading	8	53%	7	47%	0	0%	0	0%
Listening	8	50%	6	38%	2	13%	0	0%

Note. Percentages are based on the number of panelists providing a response.

Conclusions

The purpose of this study was to recommend cut scores (minimum scores) for TFI Reading and Listening sections that correspond to the A2, B1, B2, and C1 levels of the CEFR. A modified Angoff standard setting approach with a holistic component 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 for the B1 and B2 levels. The feedback included 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. The A2 and C1 levels were constructed using the final (Round 3) cut scores for B1 and B2 as references. Table 8 presents the final scaled score recommendations.

Table 8

Scaled Cut Scores for TFI

	Reading (max. 495 points)	Listening (max. 495 points)
CEFR Level		
A2	105	85
B1	185	160
B2	305	300
C1	430	395

The responses to the end-of-study evaluation survey support the quality of the standard setting implementation. The vast 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.

Half of the panelists reported that they were *very comfortable* with the recommended cut scores; the remainder of panelists reported they were *somewhat comfortable* with the Reading cut score recommendations and a majority reported being *somewhat comfortable* with the Listening recommendations. The panelists were provided an opportunity to offer open-ended written comments regarding the standard setting process and reactions to the recommended cut scores. Twelve experts wrote brief comments in this portion of the evaluation. Three issues emerged. One issue was that the TFI was not a complete measure of French language proficiency, because it does not measure French Writing and Speaking skills. Panelists noted that TOEIC® does include Writing and Speaking, and suggested that these skills should be added to TFI.

A second concern was that at times panelists needed more clarification regarding the standard setting task, which required translations from French to English and back to French, which was somewhat distracting for the panelists, and not always as timely as they would have desired. The TFI standard setting was conducted in French. All of the materials had been translated from English to French, and instructions and training regarding the standard setting process were presented in French by a bilingual (French/English) facilitator. The authors of this report were present to respond to technical/procedural questions, as needed, but speak only English and only a small number of panelists spoke English, so the translation process was necessary.

The last issue also had been raised during the standard setting discussions. The concern was that if only a total combined score (the sum of Reading and Listening scores) is reported that will likely lead to misunderstandings. For example, the recommended A2 scaled cut score for Reading is 105 and for Listening is 85 (Table 8). However, concluding that a combined score of at least 190 marks A2 proficiency is not accurate. Different combinations of scores on Reading and Listening may result in a combined score of 190; for example, a test taker may earn 130 scaled points on the Reading section (exceeding the recommended cut score), but only earn 60 points on the Listening section (below the recommended cut score). Panelists suggested that the recommended cut scores be reported separately for Reading and Listening.

Setting Final Cut Scores

The standard setting panel is 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 TFI, policymakers may be members of an academic institution that need to have a decision rule, for example, pertaining to admissions into a program of study that is conducted in French. Policymakers may also be members of an organization that need a decision rule, for example, addressing placement into a training program that is conducted in French.

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 to 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 "correct" decision; the appropriateness of any adjustment may only be evaluated in terms of its meeting the policymaker's needs. Two critical sources of information to consider when setting cut scores are the standard error of measurement (SEM) and the standard error of judgment (SEJ). The former addresses the reliability of TFI test scores and the latter the reliability of panelists' cut score recommendations.

The SEM allows policymakers to recognize that a test score—any test score on any test—is less than perfectly reliable. A test score only approximates what a test taker *truly* knows or *truly* can do on the test. The SEM, therefore, 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 TFI Reading is 22 points and is also 22 points for Listening.

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, may be considered a measure of credibility, in that 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, & Crook, 1999). In this study, the SEJs were below

two raw points; on average, this corresponds to about 11 or less scaled points or no more than one-half the size of the scaled SEMs.

In addition to measurement error metrics (e.g., SEM, SEJ), policymakers should consider the likelihood of classification error. 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 a test taker's score suggests one level of ability, but the person's actual level of ability is lower (i.e., the person does not possess the required skills). A false negative occurs when a test taker's score suggests that they do not possess the required skills, but that person nevertheless actually does possess those skills. For example, a TFI Reading score may be used by a company to place an employee into a specific position that requires at least B2 proficiency. The nature of that position may be such that not having at least a B2 level of proficiency means the person cannot carry out the core responsibilities of that position, which leads to unacceptable negative consequences. In that instance, policymakers may decide to minimize a false positive decision, and, erring on the side of caution, elect to raise the cut score for B2 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 employees who might have been at B2 Reading will now be denied access to that position. Policymakers need to consider which decision error (false positive or false negative) to minimize; it is not possible to eliminate both types of decision errors simultaneously.

References

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

Cizek, G.C. (2006). Standard setting. In S.M. Downing & T.M. Haladyna (Eds.), *Handbook of Test Development* (pp. 225–258). Mahwah, NJ: Lawrence Erlbaum Publishers.

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*(4), 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 and Psychological Measurement*, *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. (1994). Validating performance standards associated with passing scores. *Review of Educational Research*, *64*, 425–461.

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

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, RR-08-34). Princeton, NJ: ETS.

Tannenbaum R.J., & Wylie, E.C. (2005). *Mapping English Language Proficiency Test Scores Onto The Common European Framework* (TOEFL Research Report RR-80). 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] The P+ data are based on 1568 non-native French speakers around the world who took the test from October 18, 2006 to November 10, 2006. The candidates were either adults who worked in a French speaking workplace or who were learning the French language.

[2] One of the 16 panelists did not provide a response to the question regarding comfort level with the Reading cut scores.

Appendix

Panelists' Affiliations

Name	*Affiliation*
Brigitte Ringot	FFBC/ Ecole des Mines de Douai
Patrick Goyvaerts	TOEIC BELNED – ToTaal Communicatie WIPAL bvba
Anne Lhopital	Institut National des Sciences Appliquées de Lyon (INSA)
Anna Le Verger	L'Université de Technologie de Compiègne
Aline Mariage	École des Ponts ParisTech
Claudine Mela	Berlitz France
Geneviève Clinton	Arts et Métiers ParisTech
Chantal Libert	Université Paris Ouest Nanterre La Défense
Alexandra Hull	INP-ENSEEIHT
Călina-Christina Popa	Global English Inc./Groupe Renault – Automobile Dacia
Roxana Bauduin	Institut des Langues et D'Etudes Internationales – Université de Versailles Saint-Quentin-en-Yvelines
Andrey Mikhalev	Université Linguistique de Pyatigorsk (Russie)
Christine Candide	Ministēre de L'immigration, de L'intégratiar, de L'identité Nationale et du Développement Solidaire (MiiiNDS)
Rokhsareh Heshmati	Ecrimed Formation et Université de Cergy – Pontoise
I. Thomas	Universidad del Lulia

Note. One panelist did not wish to be listed in the final report.