Measuring the Power of Learning.®

Research Report
ETS RR–16-17

Setting Language Proficiency Score Requirements for English-as-a-Second-Language Placement Decisions in Secondary Education

Patricia A. Baron

Spiros Papageorgiou

June 2016

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Setting Language Proficiency Score Requirements for English-as-a-Second-Language Placement Decisions in Secondary Education

Patricia A. Baron & Spiros Papageorgiou

Educational Testing Service, Princeton, NJ

The purpose of this study was to collect recommendations for minimum score requirements (cut scores) on the *TOEFL Junior*® English-language proficiency test in order to guide decisions on the placement of learners into English as a second language (ESL) support classes. The TOEFL Junior test, intended primarily for students ages 11 and above, measures the academic and social English-language skills representative of English-medium instructional environments in secondary education. The study focused on three ESL placement decisions that affect international students applying for admission to English-medium independent or private secondary schools: (a) deny admission due to insufficient English-language skills, (b) admit conditionally with placement into ESL support, and (c) admit unconditionally (for students who have sufficient language skills at the time of admission to attend mainstream classes without ESL support). A combination of a modified Angoff standard-setting approach and a modified Performance Profile standard-setting approach was followed to identify the TOEFL Junior scores for schools to include in their decision-making process. A total of 16 language experts from private schools in 12 U.S. states served on the standard-setting panel. The results of this study provide policymakers with the minimum cut scores recommended by the panel for each section of the two testing modes: the TOEFL Junior Standard and the TOEFL Junior Comprehensive. Recommendations for the use of these scores and limitations of the results are discussed.

Keywords *TOEFL Junior*®; ESL placement; standard setting; cut scores

doi:10.1002/ets2.12102

The purpose of this study was to collect recommendations for minimum score requirements (cut scores) on a language proficiency test, specifically the *TOEFL Junior*® test, in order to guide decisions on the placement of students into English as a second language (ESL) support classes. The study was motivated by needs expressed by English-medium, independent (or private) U.S. secondary schools that admit international students. These schools have three options related to language ability when considering the applications of international students who speak English as a second language. The three options can be described as follows:

- Deny admission: Students are not admitted because they do not have sufficient English-language skills to cope with the language demands of instruction, and they are unlikely to develop such skills within a reasonable amount of time during the school year, even if ESL support is provided.
- Admit conditionally/Needs ESL support: Students are placed into an ESL support class or program because they have not yet developed all the language skills needed to cope with the language demands of instruction, but they are expected to make progress in the development of such skills within a reasonable amount of time during the school year (provided they receive ESL support).
- Admit unconditionally/No ESL support: Students are admitted into mainstream classes without any requirement to attend ESL support classes, because the student already possesses the language skills needed to cope with the language demands of instruction.

These three levels of ESL placement decision guided our study, for which teachers representing a range of U.S. private schools were recruited to participate in a standard-setting workshop. Schools were recruited to represent a range of contexts in which ESL support decisions are made. For example, the recruited schools differed by their school type, age range of students, number of levels of ESL support classes provided, and criteria required for conditional and unconditional

Corresponding author: P. Baron, E-mail: pbaron@ets.org

admission. As part of the standard-setting workshop, participants discussed and developed definitions of a hypothetical student who has just enough language ability to be put into a given placement level. These hypothetical students are referred to as *borderline students*; they describe students on the border between two levels. For example, the first borderline student definition describes students on the border between the levels not admitted and admit conditionally/needs ESL support. Borderline students were defined for entry into two levels: admit conditionally/needs ESL support and admit unconditionally/no ESL support. The definitions of these two levels, and the minimum scores (cut scores) required to meet them, are the results of this study. As noted in an earlier work (Tannenbaum & Baron, 2010), linking test scores to performance level descriptors makes the interpretation of the test scores more transparent and helps schools determine the alignment of the test scores to institutional placement criteria.

Using the TOEFL Junior Test in Student Placement

The TOEFL Junior test, intended primarily for students ages 11 and above, measures the academic and social English-language skills representative of English-medium instructional environments in secondary education. The test provides information to schools that can guide student placement and progress monitoring in classrooms that use English for content instruction and in English-language programs that build students' academic English skills. The TOEFL Junior test is available in two testing modes: the paper-based TOEFL Junior Standard and the computer-based TOEFL Junior Comprehensive.

The TOEFL Junior Standard test consists of 126 multiple-choice items divided among three sections: listening comprehension, language form and meaning, and reading comprehension. Each section contains 42 items, and the total duration of the test is 1 hour and 55 minutes. Section scores are reported on a scale ranging from 200 to 300 points, with 5-point intervals. A total score is also reported as the sum of the section scores, ranging from 600 to 900.

The TOEFL Junior Comprehensive test consists of four sections: reading comprehension and listening comprehension (each with 36 selected-response items) as well as speaking and writing (each with four constructed-response tasks). The total duration of the test is 2 hours and 14 minutes. The scores for the reading comprehension and listening comprehension sections are reported on a scale from 140 to 160 (with 1-point intervals). The scores for the speaking and writing sections range from 0 to 16 and are linked to the rubrics used to score speaking or writing tasks. The section and total scores of both modes of TOEFL Junior are accompanied by performance descriptors, which provide fine-grained information on what test takers are able to do.

For the purposes of our study, we examined the listening comprehension, language form and meaning, and reading comprehension sections of the TOEFL Junior Standard test and the speaking and writing sections of the TOEFL Junior Comprehensive test — five test sections in total. The listening and reading sections of TOEFL Junior Standard and TOEFL Junior Comprehensive are based on the same test specifications, although there are some slight differences in the way they are operationalized in each test (So et al., 2015). Because a scale alignment study has already linked the score scales across the two tests for each section (Educational Testing Service [ETS], 2012), there was no need to conduct separate cut score studies for the reading and listening sections of both TOEFL Junior Standard and TOEFL Junior Comprehensive. Rather, the present study collected recommendations for cut scores on the TOEFL Junior Standard test, and cut scores for the same sections of the TOEFL Junior Comprehensive test are determined based on the correspondence table from the scale alignment study (ETS, 2012). More information about the design of the TOEFL Junior tests can be found in Papageorgiou, Baron, and Tannenbaum (2015), Papageorgiou and Cho (2014); and So (2014). A detailed account of the theoretical and empirical foundations of TOEFL Junior is presented in So et al. (2015).

Earlier Research and Focus of the Present Study

The outcomes of a standard-setting study, such as the one presented in this report, are minimum scores (cut scores) needed to reach defined performance levels. Prior to this study, Papageorgiou and Cho's (2014) research offered preliminary validity evidence for the use of the TOEFL Junior Standard test scores to support ESL placement decisions about newly arrived students in secondary education contexts. However, the study did not offer recommendations for setting minimum score requirements on TOEFL Junior for ESL placement. As the authors explain, the sample size was particularly small for the quantitative approach they adopted in their analysis. Moreover, Papageorgiou and Cho (2014) explained that the use of different sets of cut scores by different schools might be dictated by contextual differences such as (a) the number of

level distinctions in ESL support classes, (b) class size, (c) opportunities to rectify misplacement decisions, (d) the content focus of the ESL classes, and (e) the proficiency levels of incoming students.

With this caveat in mind (i.e., that contextual factors will inevitably affect the use and appropriateness of cut scores), the focus of the present study was to recommend a set of TOEFL Junior cut scores related to ESL support for international students. The student group of interest was incoming international students in English-medium middle and high schools. Cut scores were recommended for two types of admissions decisions that related to ESL placement described earlier: admit conditionally/needs ESL support and admit unconditionally/no ESL support. For ease of reference, we will henceforth refer to these two decisions as *Needs ESL Support* and *No ESL Support*.

The present study employed a standard-setting methodology (Cizek & Bunch, 2007) in order to identify minimum score requirements for all five TOEFL Junior test sections. As discussed earlier, reading and listening cut score judgments were made on items in the TOEFL Junior Standard test sections, and cut scores for the reading and listening sections of the TOEFL Junior Comprehensive test were calculated based on those recommended for the TOEFL Junior Standard test with reference to the results of earlier scale alignment work (ETS, 2012). Panelists identified the aspects of English-language proficiency that are considered when making ESL placement decisions at their schools and then determined which scores on the TOEFL Junior Listening Comprehension, Language Form and Meaning, Reading Comprehension, Speaking, and Writing sections align to their local ESL placement criteria.

Method

Panelists

Panelists were recruited from U.S. private international schools identified as having a need to determine the amount of ESL support that should be provided to their international students. Introductory letters explaining the purpose of the project, criteria for participation, and details about the study including honorarium payment were sent to 201 institutions to ensure the range of panelist diversity needed for this work. Twenty-five curriculum vitae were received; ESL teachers and teachers of English-language arts in schools with ESL support programs were considered. Twenty educators were selected based on their experience and training working with students in the age range that TOEFL Junior targets (11+); four individuals were unable to attend. Thus, 16 individuals representing 12 states across the United States served on the panel. This number of experts allowed for a reasonable expectation of consistency in judgments (Tannenbaum & Kannan, 2015) and was higher than the generally agreed upon minimum of 10 panelists (Raymond & Reid, 2001).

The panelists represented different levels and types of school currently using or planning to adopt TOEFL Junior for making decisions about ESL support for international students. Three panelists were working in schools that included elementary and middle school grades; nine panelists in schools that included middle and high school grades, and three in schools that included high school and post–high school college-prep classes. Of the 16 schools represented, two were private religious schools, two were private military preparatory schools, and 12 were private secular schools. Based on the acceptance rates listed on the Admissions Quest website, eight of the schools had acceptances rates lower than 70%, whereas seven had acceptance rates of 70% or higher. The acceptance rate for one school was not listed on the website and was not counted in these totals.

All panelists had expertise in English-language instruction and/or assessment. Ten panelists taught ESL to students between the ages of 10 and 16, and six were ESL department chairs or coordinators. One panelist was a learning specialist, and one was an ESL tutor. (Overlap in the total panelists is due to some panelists performing dual roles as both teachers and administrators.) All had at least 6 years of overall teaching experience and at least 4 years of ESL teaching experience. Three panelists had over 30 years of general teaching experience and over 20 years of experience teaching ESL. All but three of the panelists had prior familiarity with the *TOEFL*® family of assessments (http://www.ets.org/toefl_family), and all panelists became familiar with the TOEFL Junior tests by taking them prior to the standard-setting workshop as a requirement of serving on the panel. Table 1 provides some of the self-reported demographics of the panelists. (See Appendix A for panelist affiliations and Appendix B for more panelist self-reported demographics.)

Standard-Setting Task and Methods

The standard-setting task for the panelists was to recommend the minimum scores on each of the five TOEFL Junior test sections discussed above, in order to decide whether incoming students should be placed into ESL support classes (Needs

Table 1 Panelist Demographics

Demographic	Number
Gender	
Female	14
Male	2
Ethnicity	
Hispanic	3
White, non-Hispanic	13
Function	
Teacher	7
Department director or coordinator	5
Learning specialist	1
Teacher and ESL tutor	1
Teacher and coordinator	2
School level	
Elementary school/middle school	
Grades 4 – 9	1
Grades K – 9	2
Middle school/high school	
Grades 6 – 12	2
Grades 7 – 12	2
Grades 9 – 12	5
High school/post high school	3
All grades (K – 12)	1
Institution type	
Military college preparatory	2
Private (religious)	2
Private (secular)	12
Approximate acceptance rate of institutions[a]	
50 – 59%	5
60 – 69%	3
70 – 79%	4
80 – 89%	3

[a]Acceptance rate percentages are based on the rates listed on the website www.admissionsquest.com. The number of schools totals to 15 because the acceptance rate for one school was not listed on this website.

ESL Support) or not (No ESL Support). For each test section, the general process of standard setting was conducted in a series of steps that will be elaborated upon here. Two standard-setting methods were implemented. For the test sections containing selected-response items (listening comprehension, language form and meaning, and reading comprehension), a modified Angoff procedure was followed (Cizek & Bunch, 2007; Plake & Cizek, 2012; Zieky, Perie, & Livingston, 2008). For the test sections containing constructed-response items (speaking and writing), a variation of the Performance Profile method was followed (Baron & Papageorgiou, 2014; Hambleton, Jaeger, Plake, & Mills, 2000; Tannenbaum & Baron, 2010; Tannenbaum & Cho, 2014; Zieky et al., 2008).

The modified Angoff method is a commonly used procedure in which panelists make individual probability judgments for each question they examine. For each test question, panelists identified the probability (on a scale of zero to 1.0 with intervals of 0.1) that each of the two borderline students would know the correct answer. These individual judgments were then totaled to yield two final cut scores for each test section. The Performance Profile method is a holistic method that requires panelists to make decisions or judgments based on an examinee's score profiles, or overall performance, rather than on each separate test item or task. This method has been used in standard-setting studies for English-language proficiency assessments and U.S. K – 12 statewide assessments (e.g., Baron & Papageorgiou, 2014; ETS, 2008). Panelists review actual samples of student responses across multiple tasks, such as multiple speaking samples, and consider the performance at each raw score represented by the profiles of responses across tasks. They mark the raw score representing the expected knowledge and skills at each performance level, using the definitions of borderline students at each level, as with the modified Angoff judgments. More specifics regarding materials and procedures for each of these two methods are included here.

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 them to represent varying perspectives, sufficient time devoted to ensure that the panelists develop a common understanding of the domain under consideration, adequate training of the 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 & Cho, 2014; Tannenbaum & Katz, 2013). The approach used in this study adheres to these principles. The general process of standard setting was conducted in a series of steps for each section: reading comprehension; language form and meaning, listening comprehension, speaking, and writing. (See Appendix C for the standard-setting workshop agenda.)

Preworkshop Assignments

Prior to the standard-setting study, panelists were sent a letter that described the purpose of the standard setting, explained panelists' role in the process, and detailed the security procedures followed at the standard-setting sessions. Additionally, the panelists were asked to complete a preworkshop activity to prepare them for their work. For each of the five sections of the test, the panelists were asked to consider the expectations they would have for a student with "just enough" English skill for the two types of decisions discussed earlier (*needs ESL support* and *no ESL support*). They were asked to bring their notes to the standard-setting workshop to assist in developing borderline student definitions. 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 admission decisions under consideration.

Borderline Student Definitions

Having taken all five sections of the test on Day 1, panelists' first activity on Day 2 was a discussion of the reading comprehension test section. The goal was to begin to think about and articulate perceptions of test difficulty for the intended test takers. The panelists were asked to identify and discuss test content that most students aged 11+ (that is, the typical age for TOEFL Junior test takers) learning English as a foreign or second language (a) would find particularly challenging and (b) would not necessarily find challenging.

Following this discussion, the panelists received training in the concept of borderline students; a borderline student, in the context of the standard-setting process, is defined as a student who has the minimally acceptable skills needed to reach the targeted ESL placement level (i.e., *Needs ESL Support* or *No ESL Support*). Typically, the next step in a standard-setting study is for the panel to immediately begin to develop these borderline student definitions. Owing to the diverse nature of the panel in this study and the differences across the participating schools' criteria for ESL placement in the admissions process, a group discussion was facilitated to acknowledge these differences and reach an agreement on how to reach consensus. The key idea was to determine if, in fact, schools have criteria that identify those students who will not or should not be admitted because they will not likely be successful in an English-medium academic environment. If schools have a policy whereby some students are admitted conditionally and some students are not admitted, then it makes sense for the standard-setting panel to identify those criteria and develop a cut score.

The result of this discussion was that most schools represented did have a process by which they determine international applicants' English-language skills, and for the most part, there were minimum requirements for admission based on this determination. It was acknowledged, however, that because of the diversity represented by the panel they would work toward a common definition that made sense for most schools. The goal was to identify what differentiates a student for whom ESL support would be helpful from a student for whom ESL support services would not be sufficiently helpful to allow the student to succeed in academic work at most schools. Consensus was determined to be defined as "most panelists agree and no panelists strongly object." The process of defining two borderline student definitions then proceeded as follows.

Panelists referred to their prestudy assignment notes in order to develop the borderline student definitions. Panelists were also provided with a table based on the work of Alderson et al. (2006), which indicated aspects of language proficiency for different proficiency levels (see Appendix D). This table was provided to the panelists not as a definitive taxonomy but as a starting point for discussion and development of the borderline student definitions. Panelists were instructed to use this table only to the extent that it would be a helpful tool and to disregard it if they felt it was limiting their discussions or restricting their thinking.

Panelists worked in three small groups to draft the borderline student definition for *No ESL Support* for reading comprehension. Each small group provided a list of "can-do" statements defining the reading skills of an incoming international ESL student who just meets the expectations of someone who can be admitted without any need for ESL support classes.

A whole-panel discussion of the small group draft definitions was facilitated and concluded with a consensus definition for the borderline student for *No ESL Support*. Definitions of the borderline student for *Needs ESL Support* were accomplished through whole-panel discussion using the *No ESL Support* definition as a starting point. The borderline student definition for both ESL placement decisions was repeated for all five TOEFL Junior test sections and served as the frame of reference for standard-setting judgments. (The borderline student definitions for all five test sections can be found in Appendix E.)

Angoff Standard-Setting Approach

Following the panel discussion of the difficulty of each selected-response test section and development of the borderline student definitions discussed above, panelists were trained in the modified Angoff standard-setting process and given an opportunity to practice their judgments. The facilitator first asked panelists to make judgments on four practice reading test items and discuss the rationale behind their judgments. The facilitator guided this instructional discussion and provided clarification on the procedure as needed. Each panelist was asked to complete an evaluation form indicating the extent to which the training was clear and whether or not the panelist was ready to proceed. Two panelists asked for clarification regarding the modified Angoff procedure, and the facilitator answered these questions. Once all panelists indicated their readiness to proceed, they were instructed to independently review the items and record their judgments on a rating form.

The modified Angoff approach was implemented in three rounds of judgments informed by feedback and discussion between rounds (Baron & Tannenbaum, 2011). In Round 1, for each test item, panelists were asked to judge the percentage of borderline students for *Needs ESL Support* and *No ESL Support* who would answer the question correctly. They used a judgment scale from 0 to 100 with 5-point increments. The panelists were instructed to focus only on the alignment between the English-language skills demanded by the test question and the English-language skills possessed by borderline students and not to factor random guessing into their judgments.

After completing their first round of judgments, panelists received feedback on individual and group judgments. The sum of each panelist's cross-item judgments (divided by 100) represented his or her recommended cut score. Each panelist's recommended cut score was provided to him/her. The panel-recommended cut score and the groups' highest and lowest cut scores were compiled and presented to the panel to foster discussion. Panelists were also presented with the group median cut score and the standard deviation. The panel-recommended cut score was computed by taking the average of the panelists' judgments. The average was then rounded to the next highest whole number; this whole number represented the recommended cut score. Similarly, the highest and lowest cut scores and median cut score presented to the panelists were first rounded to the next highest whole number before being presented to the panelists as feedback.

Panelists were then asked to share their judgment rationales and consider any changes to their judgments. Owing to fatigue on Day 1 of the standard-setting workshop, panelists completed Round 2 judgments for reading on the morning of the next day. This allowed them to consider the additional empirical feedback prior to completing Round 2 judgments. The empirical feedback provided was item data, specifically p values, or the percentage of test takers who answered each question correctly. The feedback was based on the performance data of test takers from one form of the TOEFL Junior Standard test administered from November 2012 to June 2014 ($N = 21{,}113$). 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). Providing item difficulty for the top 25% of candidates and the bottom 25% of candidates gave panelists a better understanding of the relationship between overall language ability for that TOEFL Junior test section and each of the test questions. It was explained that these data can show where test questions are discriminating or where a question was found to be particularly challenging or easy for test takers at different ability levels. Panelists were instructed to use the p values as a guide when considering relative difficulty of items, not as an indicator of the probability that a borderline student would get an item correct. Because panelists were making judgments for two cut scores for each item, they were reminded that their judgments would by definition be lower for *Needs ESL Support* than for *No ESL Support*. After discussion, panelists were asked to make Round 2 judgments.

In Round 2, judgments were made again at the test question level. Panelists were asked to take into account the feedback and discussion from Round 1 and were informed that they could make changes to their ratings for any question(s), for either cut score, or for both. The Round 2 judgments were compiled, and feedback similar to that presented in Round 1 was provided. In addition, impact data from the TOEFL Junior Standard test administration were presented in the form of percentage of test takers who would be placed into the two categories related to ESL support (i.e., impact data). These data include the percent of students who would fall below the cut score for *Needs ESL Support* (in other words, students who would not be admitted), the percent of students who would fall above the cut score for *No ESL Support*, and the percent of students who would fall between the two cut score recommendations. 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 provide one cut score recommendation for the overall section (e.g., reading comprehension) instead of item-level judgments. Specifically, panelists were asked to review the borderline student definitions for *Needs ESL Support* and *No ESL Support* and to decide on the recommended cut scores, taking into account the Round 2 cut score recommendations and group discussions. The transition to a holistic-level judgment places emphasis on the overall constructs of interest (i.e., listening comprehension, language form and meaning, and reading comprehension) rather than on the deconstruction of the constructs through another series of question-level judgments. Each panelist considered the test section (e.g., reading comprehension) and compared the demands of the section overall to the set of reading comprehension skills described in the borderline student definition. Panelists discussed the overall expected reading level of the borderline student and, considering the level of challenge represented by the reading comprehension section, discussed the reasonableness of the round 2 cut score relative to the expectations. This modification had been used in previous standard-setting studies (e.g., Baron & Tannenbaum, 2011) and posed no difficulties for the TOEFL Junior panelists. The three-round process was repeated with the other two selected-response test sections: language form and meaning, and listening comprehension. Standard-setting judgments were based on the 30 operational items in each section.

Performance Profile Standard-Setting Approach

A variation of a Performance Profile approach was applied to the speaking and writing sections, which require test takers to produce responses to specific test tasks. In this method, panelists considered profiles of student responses across the test tasks to make holistic standard-setting judgments. Standard setting was accomplished first for the speaking section and then was repeated for the writing section. Procedures for these two sections were the same.

Panelists were trained on the process prior to making operational judgments, and all panelists completed an evaluation indicating they were ready to proceed. Panelists next reviewed the tasks and corresponding scoring rubrics and then reviewed samples of student responses to the tasks. A student's set of responses to the tasks forms a profile; the sum of the task scores is that student's total (section) score. For each section, speaking and writing, students respond to four tasks; a student's response across the four tasks is a performance profile. Students can achieve a maximum total score of 16. Each total score can be achieved through various combinations, or profiles, of task scores. Using performance data from the administration of the TOEFL Junior Comprehensive test from July 2012 through February 2014 ($N = 2{,}339$), a total of 44 profiles for speaking and 43 profiles for writing were sampled to represent the most frequently occurring score patterns across the range of total scores. The profiles were presented in increasing score order. Table 2 provides examples of task score combinations on a profile sheet.

For speaking, audio files representing one sample of each total speaking score profile, ranging from 4 to 16, were played for the panel. Additional speaking files were played upon request by the panel as they refined their judgments for each cut score. In addition to listening to speaking samples, each panelist was provided with a printed student profile sheet to facilitate the judgment process. Similarly, panelists received a binder of the 43 student writing samples and a printed student profile sheet for the writing section.

To make cut score judgments, each panelist was asked to review the borderline student descriptions for both *Needs ESL Support* and *No ESL Support* for the particular test section in which they were working and then to review the performance profiles. The standard-setting judgment was for each panelist to decide on the total score that a borderline student for *Needs ESL Support* and a borderline student for *No ESL Support* would most likely earn. Because the borderline student for *No ESL Support* represents a higher performance expectation than the borderline student for *Needs ESL Support*, cut scores should increase as one advances from one level to the next. That is, section scores representing the skills of students who

Table 2 Sample of Student Profiles for Speaking

Candidate	Task 1 Read aloud	Task 2 Picture narration	Task 3 Listen-speak	Task 4 Listen-speak	Total score
A	1	1	1	1	4
P	2	3	2	2	9
Q	3	2	2	2	9
S	2	2	2	3	9
U	3	3	2	2	10
AA	3	2	3	3	11
AR	4	4	4	4	16

Note. Table entries include scores on individual tasks (1 through 4) and the sum of task scores.

are placed into ESL support classes would be expected to be lower than section scores of students who are placed into academic classes without ESL support.

Two rounds of judgments took place with feedback and discussion between rounds. After Round 1, each panelist's individual cut-score recommendations were displayed with a summary of the panel's average recommendations, the minimum and maximum, median, and standard deviation. Impact data were also shown after Round 1 to inform panelists about the percent of students who would be classified into each of the ESL placement levels based on the Round 1 cut scores. Panelists shared their judgment rationales. Panelists had the opportunity to adjust their Round 1 judgments in Round 2. Owing to the holistic nature of the judgments in this method, no Round 3 judgments took place. Final recommended cut scores were the panelists' mean cut scores following Round 2.

Feedback and Discussion

At the final debriefing, panelists learned what the final recommended cut scores were for all five tests (listening comprehension, language form and meaning, reading comprehension, speaking, and writing), as well as the resulting impact data. At the end of the last day, panelists were asked to complete a final evaluation form that asked questions about the process, the importance of various factors in the process, and which factors influenced their judgments. Panelists were also asked to indicate their level of confidence in the final set of recommended cut scores constructed during the process. They were given an opportunity to consider all of the cut scores across test sections and provide their opinion of the recommended scores in writing (e.g., too low, just right, too high).

Results

The first set of results summarizes the panel's standard-setting judgments for each of the five TOEFL Junior test sections. The tables summarize the results of the standard setting for the *Needs ESL support* and *No ESL Support* levels for all rounds of judgment. The results are presented in raw scores, which is the metric that the panelists used. In a later section of this report, the results are also presented in scaled scores, which is what test takers receive in their score reports. The final panel-recommended cut scores are computed by taking the average (mean) of the panelists' judgments. The panel-recommended cut scores for the selected-response test sections (reading comprehension, language form and meaning, and listening comprehension) are based on the mean of panelists' Round 3 holistic judgments on the TOEFL Junior Standard test form. The panel-recommended cut scores for the constructed-response test sections (speaking and writing) are based on the panelists' Round 2 judgments on the TOEFL Junior Comprehensive test form. For all final recommended cut scores, the group average was rounded to the next highest whole number, which is common practice in standard setting, as this is the next obtainable test score. It is this whole number that represents the final recommended cut score for each round.

Also included in each table is the standard error of judgment (SEJ), which is an estimate of the uncertainty in the panelists' judgments that is computed by dividing the standard deviation of the judgments by the square root of the number of panelists (Cizek & Bunch, 2007). The SEJ can be interpreted as an indication of 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 methods. A comparable panel's cut score would be within one SEJ of the cut score 68% of the time and within two SEJs 95% of the time. The last set of results is a summary of the panel's responses to the end-of-study evaluation survey.

Table 3 Reading Comprehension Standard-Setting Results

	Needs ESL Support cut score			*No ESL Support* cut score		
	Round 1	Round 2	Round 3	Round 1	Round 2	Round 3
Mean	9.0	9.0	9.0	26.0	25.0	25.0
Median	8.3	8.3	8.8	25.7	25.0	25.0
Minimum	3.2	4.2	4.0	20.6	23.0	22.0
Maximum	11.9	11.4	11.0	28.9	27.7	27.0
SD	2.3	2.0	1.7	2.1	1.4	1.4
SEJ	0.6	0.5	0.5	0.5	0.3	0.4

Note. ESL = English as a second language; *SD* = standard deviation; SEJ = standard error of judgment.

Table 4 Language Form and Meaning Standard-Setting Results

	Needs ESL Support cut score			*No ESL Support* cut score		
	Round 1	Round 2	Round 3	Round 1	Round 2	Round 3
Mean	9.0	9.0	9.0	26.0	26.0	26.0
Median	7.8	8.0	9.0	25.4	25.3	25.0
Minimum	2.7	5.7	6.0	23.0	23.3	23.0
Maximum	12.9	12.8	10.0	28.5	27.1	27.0
SD	2.9	2.0	1.3	1.6	1.1	1.0
SEJ	0.7	0.5	0.3	0.4	0.3	0.3

Note. ESL = English as a second language; *SD* = standard deviation; SEJ = standard error of judgment.

Reading Comprehension

Table 3 summarizes the results for the reading comprehension section for each round of judgments. The maximum raw score for this section is 30 raw score points. The cut score recommendation for *Needs ESL Support* was equal across three rounds. Cut score recommendations for *No ESL Support* decreased slightly from Round 1 to Round 2 and remained the same in Round 3. The variability (standard deviation [SD]) in panelists' judgments for the *Needs ESL support* cut score decreased across all rounds, suggesting some convergence in the final round of judgments. For the *No ESL Support* cut scores, *SD* decreased from Round 1 to Round 2 and then remained constant from Round 2 to Round 3. A similar pattern was observed in SEJ, as expected. Although SEJ did increase marginally from Round 2 to Round 3 for *No ESL Support*, the Round 3 SEJ for both cut scores is still less than 1 point, which is relatively small, and provides some confidence that the recommended cut score would be similar were a panel with comparable characteristics convened.

Language Form and Meaning

Table 4 summarizes the results for the language form and meaning section for each round of judgments. The maximum raw score for this section is 30 raw score points. The pattern for this section is similar to that seen for reading comprehension. The recommended cut score (mean) for both *Needs ESL Support* and *No ESL Support* remained the same across all rounds; however *SD* for both cut scores decreased across the three rounds, once again showing a convergence of panelists' judgments. As with the previous section, panelists' judgments were less disparate after feedback and discussion. Similarly, the SEJs generally decreased or remained the same across rounds. The Round 3 SEJ for both cut scores is less than 1 point, which is relatively small, and provides some confidence that the recommended cut score would be similar were a panel with comparable characteristics convened.

Listening Comprehension

Table 5 summarizes the results for the listening comprehension section for each round of judgments. The pattern for this section is similar to that seen for the other two selected-response sections. The recommended cut score for *Needs ESL Support* increased in Round 2 from 8.0 to 9.0 and did not change in Round 3. The recommended cut score for *No ESL Support* remained the same across the three rounds. For *Needs ESL Support*, the variability among the panelists decreased

Table 5 Listening Comprehension Standard-Setting Results

	Needs ESL Support cut score			*No ESL Support* cut score		
	Round 1	Round 2	Round 3	Round 1	Round 2	Round 3
Mean	8.0	9.0	9.0	25.0	25.0	25.0
Median	6.4	7.9	8.5	24.7	25.2	25.0
Minimum	4.6	5.6	6.0	22.5	22.5	22.0
Maximum	12.5	12.4	11.0	27.2	26.7	26.0
SD	2.7	2.1	1.3	1.4	1.0	1.1
SEJ	0.7	0.5	0.3	0.4	0.3	0.3

Note. ESL = English as a second language; *SD* = standard deviation; SEJ = standard error of judgment.

Table 6 Speaking Standard-Setting Results

	Needs ESL Support cut score		*No ESL Support* cut score	
	Round 1	Round 2	Round 1	Round 2
Mean	7.4	7.1	12.1	12.1
Median	7.0	7.0	12.0	12.0
Minimum	6.0	6.0	9.0	9.0
Maximum	9.0	9.0	14.0	13.0
SD	1.2	0.9	1.2	1.0
SEJ	0.3	0.2	0.3	0.2

Note. ESL = English as a second language; *SD* = standard deviation; SEJ = standard error of judgment.

over the three rounds, as can be seen by the decrease in the *SD*. For *No ESL support*, *SD* decreased from Round 1 to Round 2, but increased slightly from Round 2 to Round 3. The SEJ also decreased over rounds, but for *No ESL Support*, it remained the same from Round 2 to Round 3. The SEJ for listening comprehension at Round 3 is less than 1 point for both levels, which is relatively small, and provides some confidence that the recommended cut score would be similar were a panel with comparable characteristics convened.

Speaking

Table 6 summarizes the results of the two rounds of judgment for the speaking section. The maximum raw score for speaking is 16 points. The panel's average cut-score recommendation for *Needs ESL Support* decreased from 7.4 at Round 1 to 7.1 at Round 2. The average cut score for *No ESL Support* remained the same from Round 1 to Round 2. The *SD*s decreased across both rounds for both cut scores, indicating a convergence of panelists' judgments. All SEJs were less than 1 point for the speaking section.

Writing

Table 7 summarizes the results of the two rounds of judgment for the writing section. The maximum raw score for writing is 16 points. The raw score scale for this section increases by half points. Each of the four tasks received a maximum of 4 score points. Task 1 (editing) is scored using answer keys. There are 8 responses for the editing task and each correct answer receives one-half point. The panel's average cut score recommendations for *Needs ESL Support* decreased slightly from Round 1 to Round 2. For *No ESL Support*, the panel's recommendation increased from Round 1 to Round 2. The *SD* for Cut Score 1 and Cut Score 2 decreased across the rounds. All SEJs were less than 1 point for the writing section.

End-of-Study Evaluation Survey

Panelists responded to a final set of questions addressing the procedural evidence for validity of the standard-setting process (Hambleton & Pitoniak, 2006; Hambleton, Pitoniak, & Copella, 2012; Kane, 1994). The survey is a tool to gather evidence that the procedures have been implemented in a reasonable way (i.e., that panelists understood the purpose of

Table 7 Writing Standard-Setting Results

	Needs ESL Support cut score		*No ESL Support* cut score	
	Round 1	Round 2	Round 1	Round 2
Mean	5.9	5.8	11.7	12.3
Median	6.0	6.0	12.0	12.0
Minimum	4.0	4.0	9.0	10.0
Maximum	9.0	7.0	14.0	14.0
SD	1.5	0.8	1.3	1.0
SEJ	0.4	0.2	0.3	0.3

Note. ESL = English as a second language; *SD* = standard deviation; SEJ = standard error of judgment.

Table 8 Feedback on Standard-Setting Process

	Strongly agree	Agree	Disagree
The homework assignment was useful preparation for the study.	6	7	3
I understood the purpose of this study.	14	1	1
The instructions and explanations provided by the facilitators were clear.	15	1	0
The training in the standard-setting methods was adequate to give me the information I needed to complete my judgment task:			
Angoff method	13	3	0
Performance method	14	2	0
The explanation of how the recommended cut scores are computed was clear.	13	3	0
The opportunity for feedback and discussion between rounds was helpful.	16	0	0

Note. Total number of panelists responding = 16.

Table 9 Comfort Level With Recommended *Needs ESL Support* Cut Scores

	Very comfortable	Somewhat comfortable	Somewhat uncomfortable
Listening comprehension	10	2	2
Language form and meaning	11	1	2
Reading comprehension	9	5	0
Speaking	6	8	0
Writing	9	2	3

Note. Total number of panelists responding = 14. ESL = English as a second language.

the standard-setting process and how to execute their role in the work). Table 8 summarizes the panel's feedback regarding the general process. The majority of panelists strongly agreed or agreed that the preworkshop homework assignment was useful, that they understood the purpose of the study, that the instructions and explanations provided were clear, that the training provided for both methods was adequate, and that the explanation of how the recommended cut scores are computed was clear. No panelists selected the option strongly disagree for any statement. All panelists (100%) strongly agreed that the opportunity for feedback and discussion between rounds was helpful.

Panelists were also asked to indicate their level of comfort with the final cut score recommendations. Tables 9 and 10 summarize these results for both the *Needs ESL Support* and *No ESL Support* cut scores, respectively. A majority of the panelists reported they were either very comfortable or somewhat comfortable with the recommended cut scores for the five sections. No panelists selected the option very uncomfortable for any statement. For *Needs ESL Support*, a majority of panelists indicated this level of comfort for the reading comprehension and speaking sections. Two panelists were somewhat uncomfortable with the final recommended cut scores for both the listening comprehension section and language form and meaning section, and three panelists indicated they were somewhat uncomfortable with the final cut score recommendation for the writing section. Two panelists did not respond to the final evaluation questions for *Needs ESL Support* cut scores.

Typically in a standard-setting study, panelists are more homogeneous in their understanding and expectations of the student at the cut score—the borderline student. For example, teachers from the many public schools in the United States that are modifying curricula and selecting state assessments for Grades 3–8 based on the Common Core State Standards

Table 10 Comfort Level with Recommended *No ESL Support* Cut Scores

Test	Very comfortable	Somewhat comfortable	Somewhat uncomfortable
Listening comprehension	12	4	0
Language form and meaning	8	7	1
Reading comprehension	11	5	0
Speaking	11	5	0
Writing	11	4	1

Note. Total number of panelists responding = 16.

are familiar with a common set of performance level descriptors typically provided by their state. This provides the teachers with a common standard from which to begin. However, as discussed in the introduction, the context for this study was not typical. The panelists represented schools for which the ESL support resources varied widely, and the teachers in the panel were asked to describe the criteria that they believed were appropriate to identify the performance levels (*Needs ESL Support*, *No ESL Support*). This process is contrary to the typical standard-setting process, which is to align scores to an external framework whereby the test is judged in relation to external standards. Thus, the panelists' levels of comfort with the definition of a borderline student for the first cut score incorporates the panelists' diversity and the differences among the schools they represent. Some of this variability can be heard in these examples of responses to an open-ended evaluation question at the end of the study: One panelist described her concern about the judgment process:

> It was challenging to approach the borderline definitions in terms of 'should.' Doing so much admissions work, I evaluate and admit according to things a student cannot do, which is easier for the rest of admissions [officers] to understand.

Another panelist offered another point of view:

> I found the discussions of rationale for cut scores very beneficial to critique my own understanding of placement in [English language] EL classes and out of EL. I think the conference has helped me refine my own thinking. The experience has made me feel more competent to approach the new director of admissions and have a discussion about placement.

In response to a follow-up question regarding their level of comfort with Cut Score 1, one panelist stated:

> I think that the cut scores for speaking and writing are a bit too high, especially the speaking cut score. Speaking and writing are skills that are either passively acquired in immersion settings (speaking) or overtly focused on in all English and ESL classes (writing). So it seems too stringent to exclude so many students based on those cut scores we have decided on.

Another panelist wrote,

> I think some of our Cut Score 1 s [*Needs ESL Support*] may be too low. However, based upon our cut score descriptions, they are correct. I just think we may have made our Cut Score 1 description too easy.

For *No ESL Support*, all 16 of the panelists indicated they were very comfortable or somewhat comfortable with the cut scores for the listening comprehension, reading comprehension, and speaking sections. One panelist reported being somewhat uncomfortable with the recommended cut score for language form and meaning, and one panelist indicated being somewhat uncomfortable with the recommended cut score for writing. There were no comments regarding the rationales for panelists' level of comfort for the *No ESL Support* cut score.

As part of the final evaluation, panelists were given an opportunity to provide general comments about what they liked best and least about the workshop. All panelists wrote brief comments; some themes were mentioned more than once. Several panelists mentioned they found value in the group discussions, enjoyed collaborating with members of their professional community, and thought the facilitators did an excellent job of guiding the discussion. A few panelists

Table 11 Raw[a] and Scale Recommended Cut Scores for TOEFL Junior Test Sections

		Needs ESL Support		*No ESL Support*	
Test	Test section	Raw score	Scale score	Raw score	Scale score
TOEFL Junior Standard	Listening comprehension	9	220	25	285
	Language form and meaning	9	210	26	280
	Reading comprehension	9	210	25	275
TOEFL Junior Comprehensive	Speaking	8	8	13	13
	Writing	6	6	13	13
	Reading		144[b]		156[b]
	Listening		142[b]		156[b]

[a]Raw scores have been rounded up to the next achievable whole number score. [b]Based on a scale alignment study (ETS, 2012).

felt it was challenging to make overarching judgments about abilities of the borderline student when the test encompasses such a wide range of student ages. Some felt that the discussions sometimes got off topic because panelists began to discuss their own students as opposed to students in general. However, most acknowledged that there was not much that could be done about this and that the facilitators did a good job of keeping the discussions as productive as possible. Researchers have actually found through qualitative analysis of panel discussions that it is common for panelists' decision-making to be affected by students they know (Ferdous & Plake, 2005; Papageorgiou 2010; Papageorgiou, Baron, & Tannenbaum, 2015; Skorupski, 2012; Skorupski & Hambleton, 2005). It was particularly important in this workshop that the wide range of students be acknowledged explicitly in order for the panelists to reach consensus definitions of the borderline students.

Validity Evidence in Standard Setting

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

Overall Standard-Setting Study Results in Scale Score Metric

As part of score reporting, TOEFL Junior test scores are provided on the TOEFL Junior Standard score scale for listening and reading comprehension and for language form and meaning, which makes it possible to compare test scores across different forms of the TOEFL Junior test. As discussed earlier, the scaled score range for each section of the test is 200–300 for TOEFL Junior Standard. For TOEFL Junior Comprehensive, a 140–160 scale is used for the selected-response sections (listening and reading comprehension) and a 0–16 scale is used for the constructed-response sections (speaking and writing). Table 11 presents the cut scores on the reported TOEFL Junior score scales that correspond to the raw scores recommended by the standard-setting panel. Two points, which we mentioned earlier in this report, need to be pointed out. First, the raw scores were rounded up to the next achievable raw score, as is customary in applying cut score recommendations. Second, cut scores provided for the reading and listening sections on TOEFL Junior Comprehensive are calculated based on the cut score recommendations for TOEFL Junior Standard and the results of the separate scale alignment study (ETS, 2012).

Discussion and Implications

Placement Decisions in Context

The recommendations from this study are appropriate to the extent that an institution finds the scores appropriate for their use. The recommended cut score for placing students into middle school academic classes with ESL support is at the low end of the scale, which is an indication that the panel supported accepting students into school with fairly low levels of language skills. The recommended cut score for placing students into mainstream classes without ESL support was

higher, as expected. The standard-setting process results in recommendations that must be addressed by policymakers in the context of their particular need; decisions about cut scores can be made by drawing upon results of standard setting along with other information available to the decision makers. We recognize that, as also found in Papageorgiou and Cho (2014), contextual factors irrelevant to the test will affect ESL placement decisions. In this study, the institutions represented varied along many criteria relevant to determining cut scores.

Through recruiting and selection of panelists, this study intentionally included educators from a variety of schools, which provided preliminary indicators about how placement needs differ across a range of institutions. Some schools have very large ESL programs with the ability to support numerous students with varying degrees of English competency. These schools are typically willing to admit students with lower levels of proficiency, knowing that adequate ESL support is available. Other schools may only have one or two faculty members dedicated to ESL support and thus would not be able to meet the needs of an influx of students with a wide range of English-language proficiency. In the latter type of institution, the requirements for an international student applying for admission may be much more stringent, whereas in the former they would be more lenient.

In this study, panelists were asked to reach a consensus definition of the borderline students. However, due to the variety of students, range of ESL programs, and differing policies in place across the diverse schools represented by the educators on the panel, both the definition of the borderline student and the associated recommended cut scores are clearly a compromise for some of the panelists. We witnessed the process of compromise in the discussion throughout the workshop, and we also collected data in the final evaluations, which provided further articulation of the issue. For example, while many teachers described the ESL program at their school as having one class or one level of support, this was not the description of the support programs across all schools. One teacher repeatedly expressed her concern that the cut scores were not correctly placed for use by her school. She explained that in her institution there are multiple levels of support, and through discussion it was clear that her opinion was not in the majority. In the final evaluation results, we do not identify individual panelist responses, however, we expect that it would be difficult for her to "strongly agree" with the recommendation of the panel.

We also observed a difference in the degree of compromise on the two cut scores. We concluded based on the panel discussions over multiple days that there was much more agreement on the score needed to place out of an ESL class or program, because mainstream classroom requirements are more similar to each other across institutions and are not as subject to individual school differences. It was easier for panelists to agree upon the language skills needed to cope with the language demands in mainstream classrooms without any ESL support than it was for them to agree upon the more limited language skills of students who needed to attend ESL support classes. Policymakers should recognize this inconsistency when using the information in this report to set cut score requirements for admission and placement decisions.

We recommend due consideration in selecting cut scores on tests with the caveat that the cut scores presented in our study should only be treated as recommendations and that test scores should only be one part of the placement criteria. Decisions about the level of language proficiency that is acceptable for admission with or without ESL support are made at each institution and will represent a combination of factors unique to that school. School policy, resources, and values impact placement decisions, and therefore any recommendations for cut scores must be weighed in the context of institutional factors (Geisinger & McCormick, 2010).

Considerations When Setting Final Cut Scores

Given the above considerations, policymakers may want to adjust any or all of the cut scores recommended in this study. When such adjustments are considered appropriate, reasonable approaches often make use of two sources of information: the standard error of measurement (SEM) and the SEJ. The former addresses the reliability of test scores and the latter the reliability of panelists' cut score recommendations. Finally, policymakers should take into account the impact of classification errors (false positives and false negatives).

The SEM is a measure of the uncertainty of a test score and addresses the question: How close of an approximation is the test score to the *true score*? A test taker's score will be within one SEM of his or her true score 68% of the time and within two SEMs 95% of the time. The scale score cut score recommendations are in Table 11; raw and scale score SEMs for TOEFL Junior Standard and TOEFL Junior Comprehensive test sections are provided in Table 12 and are based on the most recent administrations of the tests.

Table 12 Raw and Scale Standard Error of Measurement (SEM) TOEFL Junior Test Sections

Test	Test section	Raw SEM	Scale SEM
TOEFL Junior Standard	Listening comprehension	2.31	10.41
	Language form and meaning	2.35	9.68
	Reading comprehension	2.31	9.60
TOEFL Junior Comprehensive	Speaking	1.31	1.31
	Writing	1.36	1.36
	Reading	2.24	2.13
	Listening	2.21	2.05

The SEJ is a measure of 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 (Baron & Tannenbaum, 2011). Cohen, Kane, and Crooks (1999) suggested that 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. For all five sections of the TOEFL Junior test, the raw score SEJ is less than .25 of the SEM, which meets this rule of thumb and provides evidence that the results of the study are acceptable.

Policymakers also need to consider which classification error to minimize; it is not possible to fully eliminate these decision errors (Baron & Tannenbaum, 2011). A false positive decision is one in which the conclusion made from a test score is that people have the required skill, but they actually do not. In the opposite case, a false negative decision is one in which the conclusion made from a test score is that people don't have the required skills, but they actually do. For example, once a student has been placed in an ESL class, a TOEFL Junior Reading Comprehension score may be used to determine whether the student should remain in an ESL class or be moved into the mainstream classes. The nature of instruction and expectations for the students in the mainstream level may differ substantially from those for students in an ESL class. Students may be frustrated if they are placed into a mainstream class prematurely. With that concern in mind, a policymaker may decide that it is more important to minimize false positive decisions and, erring on the side of caution, elect to raise the cut score for unconditional admission. Raising the cut score reduces the likelihood of a false positive decision, as it increases the stringency of the requirement. (It also, however, means that some number of students who might have been at mainstream level in reading will now remain in the ESL class and be denied access to the mainstream classroom instruction.)

Limitations and Areas for Future Research

It is important to point out some limitations of using language assessment scores for decisions that relate to student academic performance. In making such decisions, it should be recognized that the TOEFL Junior tests are assessments of language proficiency, and as such they cannot address other important criteria such as academic knowledge and preparation that are typically used when making admissions decisions. In other words, as other researchers have pointed out, English-language proficiency is a necessary but not sufficient condition for international students to succeed in academic contexts where English is the medium of instruction (Bridgeman, Cho, & DiPietro, 2015; Cho & Bridgeman, 2012). It must also be acknowledged that admissions and placement decisions should be based on an accumulation of evidence. That is, no one measure of knowledge or skills should be used as the sole criterion for an admission or placement decision.

We recognize the need for institutional score users to establish cut scores aligned to their own local needs. The results of the current study are based on a collection of educator recommendations and may apply to some contexts better than others. Each institution considering the use of language assessment scores for placing students into or out of ESL support must consider the institution's own context before choosing cut scores, and institutions should also reevaluate the effectiveness of chosen cut scores by reviewing students' level of success after being placed.

As it is important for test score users to consider local contexts in their use of recommended cut scores, a further limitation of the present study is that the panel of teachers included representatives from middle schools and high schools, but they developed one general set of recommendations rather than two sets (one for middle and one for high school). We are aware that cut scores may be used by a diverse population of schools and we recommend that schools collect local use data, such as test scores, teacher ratings, and, where possible, additional data representing factors that may be related to students' language skills. These data should be used in periodic reviews of cut scores to validate their appropriateness

in a given context. Following this recommendation, we collected additional data from the teachers during the current study, as well as institutional characteristics, for the purpose of future research and follow-up analyses. These data include teacher classifications of students in their classrooms and student test scores on TOEFL Junior Standard. Our hope is that by analyzing these data we will demonstrate how such data can be used to validate cut scores in local contexts.

Acknowledgments

We extend our sincere appreciation to Aina Daud for organizing the logistics and materials for the standard-setting workshop and managing the general well-being of the group. We also thank the following colleagues from the ETS Princeton office: Valerie Becker, who provided assessment development expertise prior to and during the workshop; Emily Leibowitz, for assistance in recruiting and communicating with the panelists, on-site assistance during the standard-setting workshop, and for editorial work on the manuscript; and Craig Stief, for his work on the rating forms, analysis programs, and on-site scanning.

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, 3*(1), 3–30.

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

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

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

Bridgeman, B., Cho, Y., & DiPietro, S. (2015). Predicting grades from an English language assessment: The Importance of peeling the onion. *Language Testing*. Advance online publication. doi:10.1177/0265532215583066

Cho, Y., & Bridgeman, B. (2012). Relationship of TOEFL IBT® scores to academic performance: Some evidence from American universities. *Language Testing*, *29*(3), 421–442.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

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.

Educational Testing Service. (2012). *Mapping the TOEFL® Junior™ Standard Test onto the Common European Framework of Reference: Executive Summary*. Retrieved from http://www.ets.org/s/toefl_junior/pdf/mapping_toefl_junior.pdf

Educational Testing Service. (2008). *Technical report on the standard setting workshop for the California Alternate Performance Assessment* (California Department of Education Contract Number 5417). Princeton, NJ: Author. Retrieved from www.cde.ca.gov/ta/tg/sr/documents/caparptstndstng.doc

Ferdous, A. A., & Plake, B. S. (2005). Understanding the factors that influence decisions of panelists in a standard setting study. *Applied Measurement in Education*, *18*(3), 257–267.

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., Jaeger, R. M., Plake, B. S., & Mills, C. (2000). Setting performance standards on complex educational assessments. *Applied Psychological Measurement*, *24*(4), 355–366.

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.

Hambleton, R. K., Pitoniak, M. J., & Copella, J. M. (2012). Essential steps in setting performance standards on educational tests and strategies for assessing reliability of results. In G. J. Cizek (Ed.), *Setting Performance Standards: Foundations, Methods, and Innovations* (2nd ed., pp. 47–76). New York, NY: Routledge.

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

Papageorgiou, S. (2010). Investigating the decision-making process of standard setting participants. *Language Testing*, *27*(2), 261–282.

Papageorgiou, S., Baron, P. A., & Tannenbaum, R. J. (2015). *Developing and validating a tool for setting minimum score requirements on a young learners assessment*. Paper presented at the 12th Annual Conference of EALTA, Copenhagen, Denmark.

Papageorgiou, S., & Cho, Y. (2014). An investigation of the use of TOEFL® Junior™ Standard scores for ESL placement decisions in secondary education. *Language Testing*, *31*(2), 223–239.

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.

Raymond, M. R., & Reid, J. B. (2001). Who made thee a judge? Selecting and training participants for standard setting. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 119–157). Mahwah, NJ: Erlbaum.

Skorupski, W. P. (2012). Understanding the cognitive processes of standard setting panelists. In G. Cizek, J. (Ed.), *Setting performance standards: Foundations, methods, and innovations* (pp. 135–147). New York, NY: Routledge.

Skorupski, W. P., & Hambleton, R. K. (2005). What are panelists thinking when they participate in standard setting studies? *Applied Measurement in Education*, *18*(3), 233–256.

So, Y. (2014). Are teacher perspectives useful? Incorporating EFL teacher feedback in the development of a large-scale international English test. *Language Assessment Quarterly*, *11*(3), 283–303.

So, Y., Wolf, M. K., Hauck, M. C., Mollaun, P., Rybinski, P., Tumposky, D., & Wang, L. (2015). *TOEFL Junior® Design Framework* (TOEFL Junior® Research Report TOEFL JR–02). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12058

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: Educational Testing Service.

Tannenbaum, R. J., & Cho, Y. (2014). Critical factors to consider in evaluating standard-setting studies to map language test scores to frameworks of language proficiency. *Language Assessment Quarterly*, *11*(3), 233–249.

Tannenbaum, R. J., & Kannan, P. (2015). Consistency of angoff-based standard-setting judgments: Are item judgments and passing scores replicable across different panels of experts? *Educational Assessment*, *20*(1), 66–78.

Tannenbaum, R. J., & Katz, I. R. (2013). Standard setting. In K.F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology*. Washington, DC: American Psychological Association.

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

Appendix A Panelists' Affiliation

Name	Affiliation	Location
Daiva Berzinskas	The Webb School	Tennessee
Lisa Boulestreau	Foxcroft School	Virginia
Susan Conklin	Buxton School	Massachusetts
Brenda Epifani	The Rectory School	Connecticut
Karina Escajeda	Kent Hill School	Maine
Maj. Elaine Espinosa-Sims	New Mexico Military Institute	New Mexico
Meredith Hanson	North Country School	New York
Eden Kaiser	The MacDuffie School	Massachusetts
Mary Christine Leslie	Besant Hill School	California
Johanna Maranto	Garrison Forest School	Maryland
Frank Massey	The Bement School	Massachusetts
Maya Ramírez	Army and Navy Academy	California
Amy Richardson	Villanova Preparatory School	California
Marvine Stamatakis	Interlochen Arts Academy	Michigan
Jeff Thompson	Wayland Academy	Wisconsin
Cinnie Wappel	Solebury School	Pennsylvania

Note. Panelists' affiliations are listed as requested.

Appendix B
Panelist Demographics

Table B1 Total Years' Experience Teaching All Students Ages 10–16

Years	Number	Percent
0–5 Years	0	0
6–10 Years	6	38
11–20 Years	7	44
>20 Years	3	19

Table B2 Total Years' Experience Teaching ESL Students Ages 10–16

Years	Number	Percent
0–5 Years	4	25
6–10 Years	4	25
11–20 Years	5	31
>20 Years	3	19

Table B3 Total Years' Experience Teaching Non-ESL Students Ages 10–16

Years	Number	Percent
0–5 Years	9	56
6–10 Years	3	19
11–20 Years	3	19
>20 Years	1	6

Table B4 Grades of the ESL Students You Are Currently Teaching

Grade	Number
Grade 5	1
Grade 6	4
Grade 7	6
Grade 8	6
Grade 9	15
Grade 10	13
Grade 11	10
Grade 12	7

Note. Overlap in number of total panelists per category is due to panelists selecting more than one response. Data reflect responses collected at the end of the standard-setting study. A separate institutional survey was provided to the school administrator.

Table B5 Grades of the Non-ESL Students You Are Currently Teaching

Grade	Number
Grade 5	0
Grade 6	0
Grade 7	0
Grade 8	0
Grade 9	1
Grade 10	2
Grade 11	2
Grade 12	3
Does not apply	11

Note. Overlap in number of total panelists per category is due to panelists selecting more than one response. Data reflect responses collected at the end of the standard-setting study. A separate institutional survey was provided to the school administrator.

Table B6 For Which of the Following Does Your School Use Any TOEFL Test(S)?

Answer	Number
Currently do not use TOEFL	3
Placement into English-learning language (ELL) program	8
Monitoring progress in ELL program	5
Exit criterion from ELL program	4
Admissions into school	11
Other	1

Note. Overlap in number of total panelists per category is due to panelists selecting more than one response. Data reflect responses collected at the end of the standard-setting study. A separate institutional survey was provided to the school administrator. In the second survey, only 1 school indicated they did not use any TOEFL tests.

Appendix C
Agenda

October 20–24, 2014

Day One: Monday, October 20 **8:30 a.m. – 3:30 p.m.**

Registration, Receive materials
Welcome and Overview
Break; move to Landgraf Hall
Take Speaking and Writing tests
Lunch at Chauncey Conference Center
Take Listening Comprehension, Language Form and Meaning, Reading Tests
Self-score and review rubrics
End of Day 1

Day Two: Tuesday, October 21 **8:30 a.m. – 5:00 p.m.**

Sign in and receive materials
Reading Comprehension: Review and discuss
Review Training on Borderline Student definitions
Develop Borderline Student definitions for Reading Comprehension
Lunch
Training and practice on Angoff Standard-Setting method
Round 1 judgments
Break/Data scanning
Round 1feeback and discussion; Round 2 Judgments
Break/Data scanning
Round 2 feedback and discussion; Round 3 judgments
End of Day 2

Day Three: Wednesday, October 22 **8:30 a.m. – 5:00 p.m.**

Sign in and receive materials
Language Form and Meaning: Review and discuss
Develop Borderline Student definitions for Language Form and Meaning
Round 1 judgments
Break/Data scanning
Round 1 feedback and discussion; Round 2 judgments
Break/Data scanning
Round 2 feedback and discussion; Round 3 judgments
Lunch
Listening Comprehension: Review and discuss
Develop Borderline Student definitions for Listening Comprehension
Round 1 judgments for Listening Comprehension
Break/Data scanning
Round 1 feedback and discussion; Round 2 judgments
Break/Data scanning
Round 2 feedback and discussion; Round 3 judgments
End of Day 3

Day Four: Thursday, October 23 **8:30 a.m. – 5:00 p.m.**

Sign in and receive materials
Speaking: Review and discuss test and rubrics
Develop Borderline Student definitions for Speaking
Training on Performance Profile Method
Lunch
Round 1 judgments
Break/Data scanning
Round 1 feedback and discussion; Round 2 judgments
Break/Data scanning
Writing: Review and discuss
Develop Borderline Student definitions for Writing
End of Day 4

Day Five: Friday, October 24 **8:30 a.m. – 5:00 p.m.**

Sign in and receive materials
Round 1 judgments for Writing
Break/Data scanning
Round 1 feedback and discussion; Round 2 judgments
Final evaluations of process
Lunch
Introduction to Phase II Standard-Setting Study
Introduction to **Standard-Setting TOOL**
Focus Group Discussion
Closing comments
End of Study

Appendix D
Aspects of Language Proficiency for Different Proficiency Levels

Operations (what learners need to do)	Text/input types	Strategy	Conditions and limitations	Domain of language use	Topics	Vocabulary	Grammar
Locate	Leaflets and brochures	Ask for repetition	Dictionary required for more specialized or unfamiliar texts	Educational	Mostly concrete topics	Frequent vocabulary	Simple structures
Identify	Lectures	Reread difficult sections	Difficulty with less common phrases and idioms and with terminology	Occupational	Mostly abstract topics	Limited range of extended vocabulary	Limited range of complex structures
	Letters	Read at different speed	Only familiar text/input/topic	Personal		Wide range of extended vocabulary	Wide range of complex structures
Obtain	Literary texts		Speakers need to talk slowly and clearly	Public			
Select	Signs		With a large degree of independence				
Understand	Correspondence						
Present	Announcements						
Evaluate	Labels						
Monitor	Instructions						
Scan	Articles						

Note. The table lists several aspects of language proficiency to consider. This is not an exhaustive list. It is provided as guidance only. [a]Use of language in relation to learning, e.g. in educational institutions. [b]Use of language in relation to profession. [c]Use of language as a private individual (home life, family and friends, reading for pleasure, etc.). [d]Use of language as a member of the general public and in order to complete transactions of various kinds for a variety of purposes.

Appendix E
Borderline Student Definitions

Borderline students at cut score 1 are considered ready for conditional admission, and borderline students at cut score 2 are considered prepared for unconditional admission.

Listening Comprehension

Listening Borderline Cut Score 1

1. Can understand conversations spoken at a slow rate on familiar topics with limited vocabulary and repetition.
2. Can identify key information in short and simple spoken texts.
3. Can follow simple instructions and directions.

Listening Borderline Cut Score 2

1. Can understand main ideas and supporting details in academic and social contexts at conversational speed.
2. Can demonstrate understanding of strategic listening skills in different academic and social situations, such as, predicting, inferring, summarizing, and interpreting information.
3. Can understand a wide range of vocabulary including common idioms, and infer meaning of unfamiliar words based on context.
4. Can comprehend longer spoken texts and recall information presented in multiple steps.

Language Form and Meaning

Language Form and Meaning Borderline Cut Score 1

1. Can demonstrate some knowledge of simple verb tenses and basic sentence structure.
2. Can understand high-frequency vocabulary related to familiar topics.

Language Form and Meaning Borderline Cut Score 2

1. Can demonstrate knowledge of all verb tenses within context.
2. Understands a wide range of vocabulary and word families appropriate to register and context.
3. Understand complex clause structures and phrases in a variety of texts.
4. Can demonstrate knowledge of appropriate noun structures (e.g., gerunds and infinitives, count and non-count nouns, pronouns, and quantifiers).

Reading Comprehension

Reading Borderline Cut Score 1

1. Can identify the topic of a simple passage on a familiar subject.
2. Can recognize a limited range of vocabulary, mostly consisting of high-frequency words.
3. Can identify a sequence of events based on simple lexical cues or signal words.

Reading Borderline Cut Score 2

1. Can understand main ideas and supporting details when reading informational and literary texts.
2. Can understand a wide range of vocabulary and determine unfamiliar words based on context.
3. Can understand organizational patterns, such as sequence of events, in a variety of texts in order to infer meaning and predict outcomes.
4. Can interpret and analyze information in a text in order to summarize and draw conclusions.

Speaking

Speaking Borderline Cut Score 1

1. Can speak clearly enough to convey simple messages in short chunks.
2. Can use some vocabulary to talk about daily life and survival needs in a limited fashion.
3. Can use simple tenses but frequently makes errors that don't impede communication.

Speaking Borderline Cut Score 2

1. Can speak with clarity and fluidity, using stress and intonation effectively, with occasional imperfections.
2. Can vary grammar and word choice to convey meaning appropriate to the task, with minimal error.
3. Can use a range of vocabulary appropriately to convey meaning clearly on a variety of topics, although inaccuracies occasionally occur.
4. Can pronounce most words accurately, with occasional mispronunciations which do not impede communication.
5. Can independently retell a story, in his/her own words, in order to communicate the gist with key information.

Writing

Writing Borderline Cut Score 1

1. Can respond to the topic, but not always in a clear and logical manner.
2. Can present some supporting details and relevant examples, although grammatical and mechanical errors are common and may interfere with meaning.
3. Can use a limited range of vocabulary and word choice, not always appropriate to the task.

Writing Borderline Cut Score 2

1. Can clearly and logically respond to the topic and present a relevant position with supporting details and examples to form paragraphs, with some mechanical errors, which do not interfere with meaning.
2. Can display some variation of sentence structure, edit, and use complex grammatical forms with a certain degree of accuracy.
3. Can use a range of vocabulary and word choice appropriate to the writing task, including common transitional words and phrases.
4. Can use simple tenses but frequently makes errors that don't impede communication.

Appendix F
School Characteristics

School name	Grades represented	Total number of students enrolled in Grades 6–12	Total number of applicants per year	Approximate acceptance rate[a]
Army and Navy Academy	7–12	299	506	80%
Besant Hill School	9–12	98	200	50%
Buxton School	9–12	100	70	
Foxcroft School	9–12	160	200	86%
Garrison Forest School	K–12	412	370	58%
Interlochen Arts Academy	9–12 + PG	483	420	58%
Kents Hill School	9–12 + PG	230	700	70%
MacDuffie School	6–12	268	203	58%
New Mexico Military Institute	9–12 + PG	358	2000	60%
North Country School	4–9	76	100	75%
Solebury School	7–12	218	300	75%
The Bement School	K–9	120	168	58%
The Rectory School	K–9	203	170	82%
The Webb School	6–12	296	80	64%
Villanova Preparatory School	9–12	261	250	60%
Wayland Academy	9–12	185	230	72%

Note. PG = students taking a post graduate year of college-prep courses

[a]Approximate acceptance rate percentages are based on the rates listed on the website http://www.admissionsquest.com/

Appendix G
Institutional Decision Process Information

Table G1 What Is the Admission Policy Based on English-language Requirements for Your School?

Policy	Number of Schools	%[a]
Denied Admission if requirements are not met	8	50
Conditionally admitted, and after taking English classes may apply for admission at a later time.	2	13
Admitted and placed into ESL support class	12	75
English-language requirements not part of admission decisions. Once admitted based on other criteria, students are tested to evaluate need for ESL classes.	1	6

[a]Percentages that total over 100% are due to participants being able to select more than one response.

Table G2 What Is the Process for Placing Students into English Support Classes?

Process	Number of Schools	% [a]
Students tested using an "off the shelf" test	8	50
Students tested using a custom made test for the specific institution	5	31
Students ability to read in English is tested	9	56
Students supply a writing sample	13	81
Students are interviewed face-to-face at our institution	11	69
Students are interviewed in a distance-based interview (i.e., Skype)	13	81
Other	1	6

[a]Percentages that total over 100% are due to participants being able to select more than one response.

Appendix H
Institutional *TOEFL* Product Experience

Product experience	Number	%[a]
Experience using TOEFL product		
Yes	15	94
No	1	6
Years' Experience using TOEFL product		
1 – 5 years	5	33
6 – 10 years	2	13
11 – 15 years	2	13
16 – 20 years	3	20
20+ years	1	7
Did not respond	2	13
TOEFL products used		
TOEFL Junior Standard (paper-based)	8	53
TOEFL Junior Comprehensive (computer delivered)	2	13
TOEFL® Primary™	1	7
TOEFL iBT®	11	73
TOEFL ITP®	5	33

[a]Percentages are based on the number of participants who indicated they had previous experience with any TOEFL product(s). Percentages that total over 100% are due to participants selecting more than one response.

Suggested citation:

Baron, P. A., & Papageorgiou, S. (2016). *Setting language proficiency score requirements for English-as-a-second-language placement decisions in secondary education* (Research Report No. RR-16-17). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12102

Action Editor: Donald Powers

Reviewers: Yeonsuk Cho and Veronika Timpe Laughlin