Measuring the Power of Learning.™

Research Memorandum
ETS RM–15-08

Mapping *TOEIC*® Scores to the Vietnamese National Standard: A Study to Recommend English Language Requirements for Admissions Into and Graduation From Vietnamese Universities

Richard J. Tannenbaum

Patricia A. Baron

September 2015

ETS Research Memorandum 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 Memorandum 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 Memorandum 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.

Mapping *TOEIC*® Scores to the Vietnamese National Standard: A Study to Recommend English Language Requirements for Admissions Into and Graduation From Vietnamese Universities

Richard J. Tannenbaum and Patricia A. Baron
Educational Testing Service, Princeton, New Jersey

September 2015

Corresponding author: R. J. Tannenbaum, E-mail: rtannenbaum@ets.org

Suggested citation: Tannenbaum, R. J., & Baron, P. A. (2015). M*apping TOEIC® scores to the Vietnamese National Standard: A study to recommend English Language requirements for admissions into and graduation from Vietnamese universities* (Research Memorandum No. RM-15-08). Princeton, NJ: Educational Testing Service.

Action Editor: Donald Powers

Reviewers: Priya Kannan and Spiros Papageorgiou

Abstract

The purpose of this study is to provide recommendations about the levels of English-language proficiency considered necessary for admissions into and graduation from Vietnamese universities where English instruction plays a significant role. In this regard, 13 Vietnamese educators from a range of universities participated both in alignment and standard-setting activities to identify the minimum *TOEIC*® test scores (listening, reading, speaking, and writing) corresponding to targeted levels of the Vietnamese National Standard (VNS), which defines 6 levels of English-language proficiency.

Key words: *TOEIC*®, standard setting, mapping

Table of Contents

Page

Scores on English-language proficiency tests for speakers of English as a foreign or second language (EFL/ESL) are often used to inform a variety of decisions, such as admission into a university program where instruction in English may play a significant role (Tannenbaum & Cho, 2014). In such contexts, the decision is whether a prospective student has sufficient English skills (listening, speaking, reading, and writing) to cope with the curriculum in a satisfactory manner. However, a test score, by itself, does not provide decision makers with enough information to make this decision (Tannenbaum & Cho, 2014). Knowing that a student has earned a score of 150 out of a maximum of 200 on an English listening test, for example, does not convey what that student may be expected to do outside of that test (Tannenbaum & Cho, 2014, p. 234). Tannenbaum and Cho (2014) suggested two additional criteria: an operational definition of the English skills needed for an intended purpose, such as admission into a university program, and a way to associate a test score with that definition; in this way, a test score "takes on the meaning" (p. 234) of the operational definition. This point was similarly reinforced by Kane (2012), "We can add meaning to the scores by referencing them to norms for different groups or to performance levels, benchmark performance levels, or achievement levels (e.g., as in NAEP [National Assessment of Educational Progress] or CEFR [Common European Framework of Reference])" (p. 8).

In our current study, the Vietnamese National Standard (VNS; Ministry of Education and Training, 2014) served as the external English-language framework used to assign meaning to *TOEIC*® test scores specifically for purposes of informing Vietnamese university decisions about program admission and graduation requirements. The VNS is a close adaption of the CEFR levels and level descriptors (Council of Europe, 2001). The programs of interest were those where English instruction plays a significant role. However, these were not EFL-focused programs; that is, these were not programs where Vietnamese students majored in EFL. The aim was to consider a broad variety of programs. It is important to note that the current study and its results are highly localized to the needs, expectations, and contexts of Vietnamese universities. In this regard, the results from previous studies that may have mapped TOEIC test scores to external frameworks such as the CEFR (e.g., Tannenbaum & Wylie, 2008) are not relevant. The policy-based nature of standard setting (Kane 2001; Tannenbaum & Katz, 2013) means that the local needs of decision makers, including those reflected by the standard-setting panel, likely moderate the cut-score recommendations; and so, results obtained under different contexts are not applicable.

TOEIC test scores, although more commonly used by businesses, may be used by academic institutions as they prepare students to compete in the international workplace where English is the primary language of communication (https://www.ets.org/toeic/succeed). This is the case in Vietnam, where universities that use TOEIC do so to prepare the majority of their students to enter the workforce. This report documents the methods, procedures, and results of the research study conducted to map TOEIC test scores to the VNS.

Methods Overview

The specific procedures we followed will be described in detail in subsequent sections of this report. We believe that providing a brief account, at this stage, will help place other facets of the study into proper context, such as selecting Vietnamese educators to participate in the research study. The two general approaches to mapping test scores were alignment and standard setting.

Alignment, in the present case, refers to the judged overlap between what the TOEIC test measures and the levels of the VNS, which serve as the external benchmark (domain of interest). Alignment information is an important source of content-based validity evidence—evidence that the test is a reasonable measure of the domain of interest (Bhola, Impara, & Buckendahl, 2003; Davis-Becker & Buckendahl, 2013; Martone & Sireci, 2009). Evidence supporting alignment is a necessary precondition for conducting a standard-setting study (Council of Europe, 2009). If the tested content does not reasonably overlap with the domain of interest, little justification exists for proceeding to a standard-setting study that is intended to associate scores from the test to the domain of interest (Tannenbaum & Cho, 2014, p. 237). After verifying that the test content and specified levels of the domain of interest are aligned, it is reasonable to consider setting cut scores for those levels.

Setting a standard, in the current context, refers to specifying the minimum TOEIC test scores needed to reach defined levels of the VNS. Each minimum score is referred to as a cut score; and it is through the process of standard setting that cut scores are constructed to make classifications into different levels (Tannenbaum & Kannan, 2015). A variety of specific standard-setting methods are available (see Kaftandjieva, 2010), but like alignment, standard setting is primarily based on informed expert judgment (Hambleton, Pitoniak, & Copella, 2012;

Tannenbaum & Cho, 2014) and involves a comparatively small number of experts; both alignment and standard setting often include fewer than 15 experts (Herman, Webb, & Zuniga, 2007; Katz & Tannenbaum 2014). Both alignment and standard setting are part of the larger validity argument supporting test score interpretation and use (Bejar, Braun, & Tannenbaum, 2007; Kane, 2006; Pant, Rupp, Tiffin-Richards, & Köller, 2009; Papageorgiou & Tannenbaum, in press).

Research Questions

This study focused on addressing three related research questions. What levels of the VNS[1] are covered by the TOEIC test? This question was posed for each of the four tests that form the TOEIC test (listening, reading, speaking, and writing). What levels of the VNS are needed for admission into a university program and for graduation from a university program? This defined the range of levels needed by programs. For which levels of the VNS does TOEIC overlap with program needs?

Methods

Educator Panel

The selection of experts to participate in these types of judgment-driven studies is an important source of procedural validity evidence (Kane, 1994; Tannenbaum & Katz, 2013). Thirteen educators from a range of Vietnamese university programs participated in the mapping study. The educators were selected by IIG Vietnam, (the in-country TOEIC service provider). Each educator was fluent in English, was familiar with the VNS and with the Vietnamese Ministry of Education's basic goals regarding English-language instruction, and was an experienced faculty member from a program that conducted a significant amount of instruction in English. One of the 13 educators was a senior vice president of IIG, with prior experience as a faculty member. Appendix A lists the participants and their affiliations.

Premeeting Assignment

In the first premeeting assignment, each educator was asked to review the six levels of the VNS to understand the skills expected at each level and how skill expectations changed as one progressed from one level to the next higher level. In the VNS, positive statements about what an English-language learner can do describe the expectations at each level. Educators conducted the review separately for listening, reading, speaking, and writing by evaluating tables of statements for the different English skills. This assignment was part of the process of calibrating the educators to a shared understanding of the levels; this process was continued during the mapping study.

The second premeeting assignment was for the educators to take the TOEIC test (listening, reading, speaking, and writing). Taking the test makes explicit what the test does and does not measure, and it offers the educators insights into the difficulty of the tested content. Both reviewing the levels of the framework of interest (i.e., the VNS) and taking the test are part of recommended practice (Council of Europe, 2009; Tannenbaum & Cho, 2014). In addition, as part of the alignment process, after the educators took each test (listening, reading, speaking, and writing), they were asked to complete a variation of a CEFR-recommended specification form (Council of Europe, 2009). The prompts included on the specification forms are intended to guide the educators into a reasoned understanding of the challenges posed by the test and ultimately to enable the educators to come to an informed preliminary alignment decision regarding the levels of the framework covered by the test. The Council of Europe specification form focuses attention on more macro content patterns and themes rather than on what occurs in more traditional alignment approaches (e.g., Webb, 1999, 2007) that are very item or question specific. In such traditional alignment approaches, the objective is to obtain judgments about the how well a content standard is covered by each and every test item; often judgments include both content overlap and cognitive complexity overlap (Davis-Becker & Buckendahl, 2013). This was not the focus of the current alignment; our intention was to consider more broad-based alignment concerns, appropriate and sufficient for the panelists to make reasoned judgments about the applicable VNS levels. The specification form included, for example, the following prompts: In what contexts does the test require test takers to demonstrate the English skill? What types of tasks and activities are test takers expected to be able to handle? What text-types and what length of text are test takers expected to be able to handle? The final evaluation decision was identifying the lowest and highest VNS levels covered by the test. Appendix B provides the prompts asked on the specification form used for the listening test. The same questions were asked for each of the other three tests.

Alignment

The purpose of conducting the alignment judgments was to identify the levels of the VNS needed for admission and graduation that were addressed by each of the four tests that form the TOEIC test. The same general process applies to all four tests. First, the educators discussed what the particular test (e.g., listening) measured. They were asked to comment on many of the questions that were included in the specification form from the premeeting assignment; but they were also asked to consider the challenges that the tested content likely posed for their students and other students in Vietnamese university programs where English instruction plays a significant role. Each educator was then asked to state the range of VNS levels that the test covered and a simple tally was computed. The modal lowest and modal highest values defined the range. The results from this activity addressed the breadth of the VNS levels covered by the test.

In the next activity, educators engaged in a discussion of the lowest and highest levels of the VNS needed for admission (lowest level) and for graduation (highest level). The discussion about admissions focused on the English skills (not course-specific content knowledge) needed for a student to have a reasonable opportunity of being successful in the program. Naturally, because the educators represented different universities and programs, some variability was found in the expressed expectations. The modal response was accepted as the consensus. The educators also discussed the level expected for graduation from the program. Here the focus was on what level of English skills students need to be adequately prepared for careers where English plays a role. The discussion was also framed around expectations for student improvement, that is, how many levels beyond that needed for admission are reasonable to expect students to progress. The modal response was accepted as the consensus. The results from this activity defined the range of the VNS levels needed for admission and graduation.

This range was then compared to the range of the VNS levels covered by the TOEIC test to determine where the overlap occurred. In some cases, a test may have covered levels not needed by programs. For example, the listening test was judged to cover up to Level 5, but that level was not judged to be needed by programs, as Level 4 was the highest one needed for graduation. This was also the case for the speaking and writing tests, where both tests covered levels exceeding those needed by programs. In other cases, the test did not go low enough to accommodate admission needs. For example, a majority of educators indicated that Level 1 listening skills were needed for admissions, but the test was not judged to cover a level that low.

The next most frequently cited level of listening needed for admissions (Level 2) did overlap with the test, and so that level was selected as the admission focus. Table 1 indicates the VNS levels where the TOEIC tests and the program needs overlapped. In all but one instance, reading, the overlap included Levels 2–4 of the VNS.

In all cases, except for listening (noted above) and reading, a simple majority (at least six out of 11 panelists) was the criterion for overlap. For the reading test, most educators (seven) judged the graduation needed to be at Level 4, but five educators indicated Level 5; because the reading test was also judged to cover Level 5, that level was selected as the graduation focus for this study. This selection was considered acceptable, as programs could, at their discretion, use the Level 4 or Level 5 standard as the graduation requirement. In fact, as will be clarified below, cut scores were recommended for each of the overlapping levels noted in Table 1. This gives a program the latitude to make use of any of the points of overlap for its admission and graduation requirements.

Table 1. Overlap Between Each TOEIC Test and the Needed Levels of the Vietnamese National Standard (VNS)

TOEIC test	Level 2	Level 3	Level 4	Level 5
Listening	X	X	X	
Reading	X	X	X	X
Speaking	X	X	X	
Writing	X	X	X	

Note. An *X* indicates alignment. Level 1 was not aligned with any of the tests and is not included here.

Standard Setting

Method

Listening and reading tests. These tests consist of multiple-choice items; for these tests we applied a Yes/No Angoff approach (Cizek & Bunch, 2007; Impara & Plake, 1997). This approach requires standard-setting panelists (educators) to decide if a test taker with a defined set of skills would answer an item correctly (yes) or not (no). This reduces the complexity of the standard-setting process for panelists (Impara & Plake, 1997) and is efficient when recommending cut scores for multiple levels (Plake & Cizek, 2012), as in the present case. The basic approach, described below, has been applied in the context of English-language assessment (e.g., Kantarcioğlu, Thomas, O’Dwyer, & O’Sullivan, 2010; O’Neill, Buckendahl, Plake, & Taylor, 2007).

Listening and reading tests. Educators worked in small groups and then as a whole panel to define the minimal English-language skills required to reach each of the VNS levels for which a cut score was to be recommended. A student meeting the defined minimal requirements is referred to as a just-qualified candidate (JQC). Three JQCs were defined for listening, one each for Levels 2 through 4; four JQCs were defined for reading, one each for Levels 2 through 5. The expectations of what a JQC can do increase as one progresses from one level to the next higher adjacent level. In other words, the JQC for Level 4, for example, is more able than the JQC at Level 3, who is, in turn, more able than the JQC at Level 2. Appendix C provides the JQC definitions for the four tests.

The educators used the definitions of the JQCs to guide their standard-setting judgments; more specifically, for each multiple-choice item, each educator independently decided if a JQC for a level would (yes) or would not (no) know the correct answer. Further, because each successive JQC is more able than the preceding JQC, if the JQC at Level 2, for example, would know the correct answer, then the JQCs at Levels 3 and 4 (and 5 for reading) would, without further deliberation, also know the correct answer. If a Level 2 JQC would not know the correct answer, an educator must then consider if the JQC at Level 3 would; if so (yes), then automatically a Level 4 JQC would as well. The logic of the process is the foundation of its efficiency for use with multiple proficiency levels.

Standard setting was completed for the listening test and then for the reading test. The general process was the same in application. The educators were trained in the yes/no approach and given an opportunity to practice applying the approach. As part of the training process, the educators shared the rationales for their judgments; this helped to clarify how the educators were interpreting the JQC for each level when they considered the test items. The educators completed a training evaluation form, which documented the adequacy of the training received and their readiness to begin their first round of standard-setting judgments. All educators affirmed that they were ready to proceed. The training was implemented in full during the standard-setting for the listening test; a refresher was conducted for the reading test.

In both applications, two rounds of judgments occurred. In the first round, each educator made independent item-level judgments for each of the relevant VNS levels. There are 100 items on the listening test and 100 items on the reading test. For each educator, for each VNS level, the sum of *yes* judgments across items is the educator's recommended cut score for that level. Within each VNS level, the item judgments were summed across the educators and then the average of this value was computed across the educators, resulting in the panel's recommendation of a cut score for that level. For example, if for Level 2 listening, each and every educator indicated that the JQC at this level would know the correct answer to any 40 of the 100 items, each educator's recommendation of a cut score would be 40 (i.e., the sum of 40 *yes* judgments or 1s) and the mean (computed average) across the educators would also be 40. This is a hypothetical example, as variability is always present in the recommendations of the individual educators; however, the computations applied operationally are the same as in the example.

The results of the first round were shared with the educators. This presentation consisted of each individual educator's recommended cut score for each level; summary information that represented the panel's recommendations by level: mean, lowest and highest values and standard deviation; and the percentage of test takers who would be classified into each of the levels, if the mean cut score were applied. The latter (impact data) included roughly 4,000 students who had recently tested in Vietnam. The impact data provide the educators with a better sense of the likely consequences of the Round 1 recommended cut score (Reckase, 2001). These multiple sources of data (educator, panel, and impact) were used to facilitate discussion among the educators about the reasonableness of the Round 1 recommended cut scores.

A second presentation provided the educators with item-level information, specifically, for each item, the number of yeses by each level were displayed. This information enabled the educators to see where they were more or less in agreement in their item judgments. Additionally, we presented *p*-values (the percentage of the 4,000 test takers answering an item correctly) for selected items to help focus and inform discussions. These values were useful to show how difficult items are for test takers, in general, and to show the relative difficulty of the items, which items are more or less difficult than other items. However, mindful of research that indicates standard-setting panelists may be unduly influenced by such data (e.g., Clauser, Mee, Baldwin, Margolis, & Dillion, 2009), we elected to show *p*-values only for a small number of items, where we thought the information would be useful. These were instances, for example, where panelists' ratings for adjacent test items may have been dissimilar but the *p*-values were similar, and vice versa.

The feedback and discussion of the Round 1 results informed the educators' second round of standard-setting judgments. Here, however, the judgment was not on the individual items, as in the first round, but on the recommended cut scores for each level. Educators had the opportunity, but were not required, to change their individual recommendations. The Round 2 mean recommended cut scores for each level represent the final set of recommendations.

Speaking and writing tests. These tests require students (test takers) to produce responses to several tasks (prompts); for these we applied a variation of the Performance Profile approach (Perie & Thurlow, 2012; Tannenbaum & Baron, 2010; Tannenbaum & Cho, 2014; Tannenbaum & Wylie, 2008; Zieky, Perie, & Livingston, 2008). One clear advantage of this approach is that the educators engage with direct evidence of students' English-language skills, which is something the educators do on a regular basis in their classrooms. A profile is a student's response to each of the test tasks that results in a total test score. For example, there are eight TOEIC Writing tasks, so a student's response to each of the eight tasks forms that student's performance profile. Profiles are presented to the educators in increasing total test score order, lowest to highest; for the writing test, the written responses are provided to the educators in a binder, and for the speaking test, the responses are played aloud. The educators consider the presented evidence and decide on the total test score a JQC at each of the targeted levels would most likely earn.

Standard setting was completed for the speaking test and then for the writing test. The general process was the same in application. Working in small and whole groups, the educators constructed the JQC definitions for each of the targeted levels. They were trained in the Performance Profile approach, and given an opportunity to practice applying the approach. The educators shared the rationales for their judgments and completed a training evaluation form. All educators affirmed that they were ready to proceed. The training was implemented in full during the standard-setting for the speaking test; a refresher was conducted for the writing test.

Two rounds of judgments occurred. In the first round each educator made independent judgments for each of the relevant VNS levels. Recall that when applying the Performance Profile approach, decisions are made at the total score level, not at an individual task level. The raw score range is 0 to 11 points for the speaking test and 0 to 26 points for the writing test.

Although the focus is on the total score, the educators were provided with the tasks and scoring rubrics to reinforce what performance was expected from the students (test takers). The educators were presented with 24 response profiles for the writing test (ranging from 7.8 to 26 points) and 21 for the speaking test (ranging from 3.8 to 11 points). Because the same total score can be obtained through different arrays of task scores we presented more than one response sample for some of the same total scores, where the frequency of multiple profiles was more prevalent. The samples serve as illustrative benchmarks. They are not exhaustive and they do not illustrate every possible legitimate total score.

In Round 1, the educators independently considered the evidence and noted the total scores that each JQC (by level) would most likely earn. The educators were able to select any total score in half-point increments within the legitimate range for that test, even if a performance sample was not available for that total score. Each individual educator's recommended cut scores were averaged to arrive at the panel recommendations.

The results of the first round were shared with the educators. This presentation consisted of each individual educator's recommended cut score for each level; summary information that represented the panel's recommendations by level: mean, lowest and highest values, and standard deviation; and impact data (percentage of test takers classified into each level). Unlike the listening and reading tests, the impact data applied here were based on 850 test takers from a range of international countries because the speaking and the writing tests, currently, are not widely used in Vietnam. The multiple sources of data (educator, panel, and impact) were used to facilitate discussion among the educators about the reasonableness of the Round 1 recommended cut scores and to inform their Round 2 judgments.

Results

Listening and reading tests. Tables 2 and 3 present the standard-setting results for the listening test and reading test, respectively. The Round 2 means represent the panel-recommended cut scores. A cut score reflects the minimum required (lowest acceptable) score; for example, for the listening test, the Level 2 recommended cut score is 22.6 points out of 100. This recommendation indicates that, on average, the panel believes that a score higher than 22 points is needed; we therefore rounded all nonwhole recommended cut scores to the next highest whole number. The recommended cut scores for the listening test are 23 points for Level 2, 59 for Level 3, and 87 for Level 4. The recommendations for the reading test are 20 for Level 2, 54 for Level 3, 86 for Level 4, and 96 for level 5.

The variability among the educator's recommendations is reflected by the range between the lowest and highest cut scores and the standard deviation. These values, for both tests, indicate that there were differences among the educators. This finding is to be expected, given the policy-based nature of setting standards, the differing program perspectives represented by the educators, and the relatively large number of items on each test. The most variability occurred at Level 3. Level 3 reflects the transition point from basic proficiency to the first stage of intermediate proficiency, and so inferences related to this transition stage may have added to the other likely sources of variability. The relatively small standard deviations at Level 5 for the reading test are due to a ceiling effect; that is, all the recommended cut scores needed to reach that level were high, 90 points or more out of 100.

Table 2. Recommended Cut Scores by Vietnam National Standard (VNS) Level and Standard-Setting Round for Listening

Statistic	Round 1			Round 2		
	L2	L3	L4	L2	L3	L4
Mean	22.2	60.2	89.6	22.6	58.4	86.4
Minimum	3.0	35.0	78.0	6.0	34.0	76.0
Maximum	36.0	77.0	96.0	33.0	78.0	96.0
SD	10.5	12.6	5.6	9.3	13.9	6.8

Note. Level 1 was not aligned with any of the tests and is not included here.

Table 3. Recommended Cut Scores by Vietnam National Standard (VNS) Level and Standard-Setting Round for Reading

Statistic	Round 1				Round 2			
	L2	L3	L4	L5	L2	L3	L4	L5
Mean	18.8	54.9	88.2	98.3	19.8	54.0	86.0	95.8
Minimum	6.0	31.0	67.0	90.0	6.0	33.0	70.0	90.0
Maximum	40.0	77.0	95.0	100.0	37.0	80.0	95.0	99.0
SD	10.0	13.3	8.3	2.9	9.7	13.1	7.1	2.9

Note. Level 1 was not aligned with any of the tests and is not included here.

Speaking and writing tests.[2] Tables 4 and 5 present the standard-setting results for the speaking test and writing test, respectively. Legitimate raw scores progress in 0.1 increments for the speaking test and in 0.2 increments for the writing test. This means that the recommended cut scores for speaking were not rounded up; however, the mean cut scores for writing, Levels 3 and

4, were rounded to the next highest legitimate value. The recommended cut scores for the speaking test are 4.4 points for Level 2, 6.5 for Level 3, and 7.8 for Level 4. The recommendations for the writing test are 9.8 for Level 2, 14.6 for Level 3, and 20.2 for Level 4.

The variability among the educator's recommendations is reflected by the range between the lowest and highest cut scores and the standard deviation. These values, for both tests, are considerably smaller than those for the listening and reading tests. This finding indicates that the educators were more in agreement on their recommendations for the speaking and writing tests than for the other two tests. This agreement, too, is to be expected, given the comparatively small number of points available for the speaking and writing tests, 11 and 26, respectively. The closer agreement is also likely due to the fact that here the educators were engaging with direct evidence of students' performance (the response samples), which helps to reduce some of the inferences that are more prevalent when setting standards on multiple-choice items.

Table 4. Recommended Cut Scores by Vietnam National Standard (VNS) Level and Standard-Setting Round for Speaking

Statistic	Round 1			Round 2		
	L2	L3	L4	L2	L3	L4
Mean	4.2	6.4	7.8	4.4	6.5	7.8
Minimum	3.5	5.0	7.0	4.0	5.5	7.0
Maximum	5.5	7.5	8.5	5.0	7.5	8.5
SD	0.6	0.7	0.5	0.5	0.6	0.6

Note. Level 1 was not aligned with any of the tests and is not included here.

Table 5. Recommended Cut Scores by Vietnam National Standard (VNS) Level and Standard-Setting Round for Writing

Statistic	Round 1			Round 2		
	L2	L3	L4	L2	L3	L4
Mean	9.6	14.1	19.0	9.8	14.5	20.1
Minimum	8.0	12.0	16.0	8.0	12.5	18.0
Maximum	13.0	16.0	23.0	13.5	16.5	22.5
SD	1.4	1.2	2.2	1.6	1.3	1.5

Note. Level 1 was not aligned with any of the tests and is not included here.

Panelist Evaluations

One source of validity evidence when setting standards comes directly from feedback provided by the educators' (Cizek, Bunch, & Koons, 2004) so-called procedural validity evidence (Kane, 1994, 2001). Cizek (2012) notes that feedback from panelists about the standard-setting process and results is a critical source of information for decision makers to consider when either adopting or adjusting recommended cut scores. Each educator completed an end-of-study evaluation form addressing the standard-setting process and the reasonableness of the Round 2 mean recommended cut scores (see Tables 6 and 7).[3]

All educators either strongly agreed or agreed with each of the evaluation questions, with the majority of educators replying strongly agree (Table 6). Although the response scale included options for the educators to disagree or strongly disagree, no educator chose those options, and so those options are not included in the table.

Table 6. Panelist Evaluations: Standard-Setting Process

Statement	Strongly agree	Agree
	N	*N*
The premeeting assignment was useful preparation for the study.	10	1
I understood the purpose of this study.	9	2
The instructions and explanations provided by the facilitators were clear.	11	0
The standard-setting training gave me the information I needed to complete my judgment task:		
• Y/N Angoff method	10	0
• Performance profile method	10	1
The explanation of how the recommended cut scores are computed was clear.	7	4
The opportunity for feedback and discussion between rounds was helpful.	9	2

Note. One panelist omitted the question about training for the Y/N Angoff method.

Further, as reported in Table 7, the majority of the educators reported being either very comfortable or somewhat comfortable with the set of recommended cut scores. There was one instance, for the writing test, where an educator indicated being very uncomfortable, and four other instances where educators were somewhat uncomfortable with the recommended cut scores. The educators were asked to offer a reason for any response indicating they were uncomfortable with the cut scores; however, not all did. Those who did commented that the cut score for listening, Level 4, was too high; that the cut scores for speaking, at each level, were too high; and that the cut scores for writing, Levels 2 and 4, were too high. Collectively, the results from Tables 6 and 7 provide positive evidence of the procedural validity of the standard-setting process and results.

Table 7. Panelist Evaluations: Comfort Level With the Recommended Cut Scores

Test	Very comfortable	Somewhat comfortable	Somewhat uncomfortable	Very uncomfortable
	N	*N*	*N*	*N*
Listening	4	5	2	0
Reading	6	5	0	0
Speaking	8	2	1	0
Writing	3	6	1	1

Conclusions

This study focused on constructing minimum TOEIC test scores (cut scores) that correspond to targeted levels of the VNS. The cut scores will be used to inform decisions regarding English-language requirements for admission into and graduation from university programs in Vietnam. Through alignment and standard-setting procedures, educators identified the targeted levels of the VNS and then recommended corresponding cut scores for each of the four TOEIC tests.

Throughout the standard-setting study, educators worked with test information in raw-score units; for the listening and reading test, this is 100 points, and for the speaking and writing tests, this is 11 points and 26 points, respectively. These raw scores are based directly on the number of multiple-choice items answered correctly and on the number of rubric points earned across tasks. The direct relationship between student performance (on items and tasks) and total points earned makes it is easier and more understandable for educators to recommend cut scores in terms of raw scores. TOEIC scores, however, are reported to test takers and university decision makers as scale scores. Scale scores represent the raw scores on a new scale, which makes comparing test takers’ performance on different forms of the same test simpler. The scale-score range for the listening test and the reading test is 5 to 495; it is 0 to 200 for the speaking test and the writing test. Table 8 presents the recommended cut scores both as raw scores (rounded, if appropriate, as described previously) and scale scores. It is important to note that test takers’ performance on two tests measuring different skills can result in the same reported scale score, but that does not necessarily mean that identical levels of proficiency were demonstrated on the two different tests. For example, if a test taker earned 6.5 raw points on the speaking test and 14.6 points on the writing test, these scores would each be reported as 110 points, as shown in Table 8.

Table 8. Recommended Cut Scores as Raw Scores and Scale Scores

Test	Raw score				Scale score			
	L2	L3	L4	L5	L2	L3	L4	L5
Listening	23	59	87		85	280	445	
Reading	20	54	86	96	65	225	420	485
Speaking	4.4	6.5	7.8		60	110	140	
Writing	9.8	14.6	20.2		70	110	170	

Note. Level 1 was not aligned with any of the tests and is not included here.

Poststudy Adjustments

Setting standards is very similar to forming a policy (Cizek & Bunch, 2007; Kane, 2001; Tannenbaum & Katz, 2013) in that, in each instance, the goal is to construct and apply decision rules that are reasonable and appropriate for their intended use. In the present case, the rules relate to what cut scores are reasonable requirements for admissions and graduation decisions. Geisinger and McCormick (2010) correctly noted that the process of setting a standard begins with the recommendations from the assembled panel of educators, but it does not end there. Decision makers who interpret and apply cut scores operationally (for consequence) are encouraged to evaluate them, and to adjust them, if doing so better meets their local needs and values (Geisinger & McCormick, 2010).

One source of information to use when considering adjusting a cut score is the standard error of measurement of the test (Cizek, 1996; Geisinger & McCormick, 2010). Test scores are not free of error. A band of uncertainty, which is represented by the SEM, is always associated with test scores. Adjusting cut scores by a standard error of measurement, for example, is considered common practice (Cizek & Bunch, 2007) when doing so offers a more reasonable fit with the intended use. The standard errors of measurement[4] for TOEIC tests are reported in scale score units. They are 25 for the listening test, 25 for the reading test, 15 for the speaking test, and 20 for the writing test. Decision makers may, therefore, consider raising or lowering one or more of the recommended cut scores (in scale score units) using the standard error of measurement, again, if doing so will likely better support their local needs. One reason for lowering a cut score may be that a particular university program offers considerable English-language support to those students in need, and therefore, it may be in a position to accept lower English-language scores for students admitted into the program. So, for example, if a program offers English writing development resources, it could decide to lower the Level 3 writing cut score by a standard error of measurement, the adjusted score would be 90 (110−20). If decision makers do adjust recommended cut scores, they are advised to document the rationales for those changes, as the rationales become part of the overall justification for the operational cut scores.

References

Bejar, I. I., Braun, H. I., & Tannenbaum R. J. (2007). A prospective, progressive, and predictive approach to standard setting. In R. W. Lissitz (Ed.), *Assessing and modeling cognitive development in school* (pp. 1−30). Maple Grove, MN: JAM Press.

Bhola, D., Impara, J. C., & Buckendahl, C. W. (2003). Aligning tests with states' content standards: Methods and issues. *Educational Measurement: Issues and Practice, 22,* 21−29. http://dx.doi.org/10.1111/j.1745-3992.2003.tb00134.x

Cizek, G. J. (1996). Standard setting guidelines. *Educational Measurement: Issues and Practice, 15*, 13–21. http://dx.doi.org/10.1111/j.1745-3992.1996.tb00802.x

Cizek, G. J. (2012). An introduction to contemporary standard setting: Concepts, characteristics and contexts. In G. J. Cizek (Ed.), *Setting performance standards*: *Foundations, methods, and innovations* (2nd ed., pp. 3−14). New York, NY: Routledge.

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

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice, 23*, 31−50. http://dx.doi.org/10.1111/j.1745-3992.2004.tb00166.x

Clauser, B. E., Mee, J., Baldwin, S. G., Margolis, M. J., & Dillon, G. F. (2009). Judge's use of examinee performance data in an Angoff standard-setting exercise for a medical licensing examination: An experimental study. *Journal of Educational Measurement, 46*, 390−407. http://dx.doi.org/10.1111/j.1745-3984.2009.00089.x

Council of Europe. (2001). *Common European framework of reference for language: Learning, teaching, assessment*. Cambridge, UK: Cambridge University Press.

Council of Europe. (2009). *Relating language examinations to the Common European framework of reference for languages: Learning, teaching, assessment.* Strasbourg, France: Author.

Davis-Becker, S. L., & Buckendahl, C. W. (2013). A proposed framework for evaluating alignment studies. *Educational Measurement*: *Issues and Practice*, *32,* 23−33. http://dx.doi.org/10.1111/emip.12002

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. http://dx.doi.org/10.1111/j.1745-3992.2009.00168.x

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

Herman, J. L., Webb, N. M., & Zuniga, S. A. (2007). Measurement issues in the alignment of standards and assessments: A case study. *Applied Measurement in Education, 20,* 101−126.

Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement, 34,* 353−366. http://dx.doi.org/10.1111/j.1745-3984.1997.tb00523.x

Kaftandjieva, F. (2010). *Methods for setting cut scores in criterion-referenced achievement tests: A comparative analysis of six methods with an application to tests of reading in EFL*. Arnhem, The Netherlands: CITO.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64,* 425−461. http://dx.doi.org/10.3102/00346543064003425

Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 53−88). Mahwah, NJ: Lawrence Erlbaum.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17−64). Westport, CT: Praeger.

Kane, M. (2012). Validating score interpretations and uses. *Language Testing*, *29,* 3−17. http://dx.doi.org/10.1177/0265532211417210

Kantarcioğlu, E., Thomas, C., O'Dwyer, J., & O'Sullivan, B. (2010). Benchmarking a high-stakes proficiency exam: The COPE linking project. In W. Martyniuk (Ed.), *Studies in language testing: Aligning tests with the CEFR* (pp. 102−118). Cambridge, UK: Cambridge University Press.

Katz, I. R., & Tannenbaum, R. J. (2014). Comparision of web-based and face-to-face standard setting using the Angoff method. *Journal of Applied Testing Technology, 15,* 1−17.

Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessment, and instruction. *Review of Education Research, 79,* 1332–1361. http://dx.doi.org/10.3102/0034654309341375

Ministry of Education and Training. (2014). *Promulgating the Vietnam's 6-level framework of reference for foreign language competence.* Retrieved from http://www.moet.gov.vn/?page=6.10&view=5552

O'Neill, T. R., Buckendahl, C. W., Plake, B. S., & Taylor, L. (2007). Recommending a nursing-specific passing standard for the IELTS examination. *Language Assessment Quarterly, 4,* 295−317. http://dx.doi.org/10.1080/15434300701533562

Pant, H. A., Rupp, A. A., Tiffin-Richards, S. P., & Köller, O. (2009). Validity issues in standard-setting studies. *Studies in Educational Evaluation, 35,* 95−101. http://dx.doi.org/10.1016/j.stueduc.2009.10.008

Papageorgiou, S., & Tannenbaum, R. J. (in press). Situating standard setting within argument-based validity. *Language Assessment Quarterly*.

Perie, M., & Thurlow, M. (2012). Setting achievement standards on assessments for students with disabilities. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 347−377). New York, NY: Routledge.

Plake, B. S., & Cizek, G. J. (2012). Variations on a theme: The modified Angoff, extended Angoff, and yes/no standard setting methods. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 181−199). New York, NY: Routledge.

Reckase, M. D. (2001). Innovative methods for helping standard-setting participants to perform their task: The role of feedback regarding consistency, accuracy, and impact. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 159−173). Mahwah, NJ: Lawrence Erlbaum.

Tannenbaum, R. J., & Baron, P. A. (2010). *Mapping TOEIC test scores to the STANAG 6001 language proficiency levels* (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,* 233−249. http://dx.doi.org/10.1080/15434303.2013.869815

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,* 66−78. http://dx.doi.org/10.1080/10627197.2015.997619

Tannenbaum, R. J., & Katz, I. R. (2013). Standard setting. In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology: Vol 3. Testing and assessment in school psychology and education* (pp. 455−477). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/14049-022

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 Research Report No. TOEFLiBT-06). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2008.tb02120.x

Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states*. Council of Chief State School Officers & National Institute for Science Education. Madison, WI: Wisconsin Center for Education Research, University of Wisconsin.

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education, 20,* 7−25. http://dx.doi.org/10.1080/08957340709336728

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. Educator Panel

Name	Affiliation
Từ Thị Minh Thúy	University of Foreign Languages and International Studies, Vietnam National University, Hanoi
Nguyễn Thị Nhân Hòa	International School, Vietnam National University, Hanoi
Pham Thị Diệu Linh	Academy of Policy and Development
Nguyễn Vân Thịnh	Thai Nguyen University of Economics and Business Administration
Nguyễn Kiều Oanh	University of Foreign Languages and International Studies, Vietnam National University, Hanoi
Nguyễn Phương Sửu	IIG Vietnam
Trần Thị Minh Thảo	People’s Security Academy
Tân Anh	Academy of Policy and Development
Đỗ Thi Xuân Hoa	University of Foreign Languages and International Studies, Vietnam National University, Hanoi
Bùi Mỹ Ngọc	University of Economics Ho Chi Minh City
Hoàng Thị Ngọc Diệp	Vietnam Maritime University
Trần Thị Thu Giang	National Economics University

Note. Panelists’ affiliations are listed as requested.

Appendix B. Specification Form for Listening

Listening	Provide a brief rationale
What contexts (domains, situations) does the test require test takers to demonstrate listening skills?	Brief rationale:
Given these contexts, what level of challenge will this present to test takers?	Degree of challenge (Check one): __ Small __ Moderate __High
2. Which communication themes are test takers expected to be able to handle?	Brief rationale:
Given these themes, what level of challenge will this present to test takers?	Degree of challenge (Check one): __ Small __ Moderate __High
3. What types of tasks and activities are test takers expected to be able to handle?	Brief rationale:
Given these tasks and activities, what level of challenge will this present to test takers?	Degree of challenge (Check one): __ Small __ Moderate __High
4. What text-types and what length of text are test takers expected to be able to handle?	Brief rationale:
Given these text types and lengths, what level of challenge will this present to test takers?	Degree of challenge (Check one): __ Small __ Moderate __High

5. Given your responses to the four questions above, and after reading the Overall Listening level descriptors, indicate the highest and lowest levels of the Vietnamese National Standard covered by the TOEIC Listening test?	Highest (Check one):
	__ 6 __ 5 __4 __3 __2 __1
	Lowest (Check one):
	__ 6 __ 5 __4 __3 __2 __1

Appendix C. Just Qualified Candidate Definitions

Listening

Listening Level 4 (JQ4)

1. Can understand unfamiliar and abstract topics beyond their interests and specialization, provided the speech is standard spoken language
2. Can understand live or broadcast standard spoken language at normal speed
3. Can follow extended speech when some background noise is present
4. Can infer the meaning based on explicit transition
5. Can follow extended and complex lines of argument
6. Can recognize common idiomatic expressions

Listening Level 3 (JQ3)

1. Can understand straightforward factual information, when the speech is clearly articulated
2. Can identify general idea and some specific details on familiar topics
3. Can follow a series of simple instructions to operate common, everyday equipment or technology

Listening Level 2 (JQ2)

1. Can follow simple instructions, when the speech is slowly and clearly articulated
2. Can understand phrases and expressions on basic topics of priority needs, when the speech is slowly and clearly articulated
3. Can understand only the main ideas, when the speech is slowly and clearly articulated

Reading

Reading Level 5 (JQ5)

1. Can understand in detail lengthy, complex texts provided he/she can reread difficult sections
2. Can identify finer points of details including attitudes and implied and stated opinions
3. Can use advanced grammatical structures to facilitate comprehension, e.g., inversion; mixed conditional sentences; speculation
4. Can infer the meaning by combining information from different parts of a text or from different texts on the same topic
5. Can understand some low-frequency idioms

Reading Level 4 (JQ4)

1. Can quickly scan and locate relevant information in long and complex texts
2. Can obtain information, ideas and opinions from highly specialized sources within his/her field
3. Can understand some high-frequency idioms
4. Can understand articles and reports concerned with contemporary problems in which writer adopts particular stances or viewpoints
5. Can use some advanced grammatical structures to facilitate comprehension, e.g., noun clauses; adjective phrases; third conditional sentences
6. Can make simple inferences

Reading Level 3 (JQ3)

1. Can understand straightforward factual text related to his/her field of interest
2. Can find and understand relevant information in everyday materials
3. Can recognize lines of argument with explicit markers
4. Can use intermediate grammatical structures to facilitate comprehension, e.g., second conditional sentences; reported speech; passive voice

Reading Level 2 (JQ2)

1. Can understand short and simple texts on familiar matters
2. Can locate specific and predictable information in everyday materials
3. Can understand high frequency vocabulary in everyday and job-related materials
4. Can understand short simple description with visual aids
5. Can use elementary grammatical structures to facilitate comprehension, e.g., simple present, simple comparative, and common modal verbs

Speaking

Speaking Level 4 (JQ4)

1. Can present personal viewpoints with relevant and specific explanations and reasons
2. Can present on a wide range of subjects related to his/her field of interests
3. Can communicate confidently, clearly and politely in everyday interactions
4. Can speak with clear pronunciation and intonation
5. Can speak with a certain degree of fluency and accuracy without causing difficulties for listeners to understand

Speaking Level 3 (JQ3)

1. Can speak at length on everyday topics with reasonable intelligibility even with some mispronunciation
2. Can interact appropriately with a certain degree of fluency but with frequent hesitations
3. Can express personal viewpoints and exchange information on familiar topics
4. Can narrate short stories and give straightforward descriptions
5. Can produce intelligible pronunciation even with heavy non-native stress and intonation
6. Can use a wide range of basic vocabulary/grammar with some difficulties in expression

Speaking Level 2 (JQ2)

1. Can communicate simple and routine activities using basic structures
2. Can handle very short social exchanges but cannot maintain conversation
3. Can communicate but pronunciation may require a lot of effort from listener to understand
4. Can give a simple description on everyday topics

Writing

Writing Level 4 (JQ4)

1. Can write clear detailed texts on a variety of subjects related to his/her field of interests
2. Can write an essay or a report that develops simple arguments
3. Can express most common contemporary news and views
4. Can use some common idiomatic expressions, appropriate word choice with minor errors
5. Can use appropriate format and formality in personal and business correspondence
6. Can cover good grammatical control; occasional grammatical errors don't lead to misunderstanding.

Writing Level 3 (JQ3)

1. Can write straightforward connecting texts in form of a standard paragraph on a range of familiar topics within his/her field of interest
2. Can write personal correspondence, notes and simple instructions
3. Can use sufficient vocabulary on most topics pertinent to his/her everyday life but major errors still occur when expressing complex thoughts or handling unfamiliar topics or situations.
4. Can use basic grammatical structures with reasonable accuracy in familiar contexts with noticeable mother tongue influence

Writing Level 2 (JQ2)

1. Can write short, simple, formulaic notes and messages on matters of immediate needs
2. Can write simple phrases and some standard sentence structure with simple connectors (e.g., and, but)
3. Can write with reasonable mechanic accuracy (e.g., punctuation, spelling)

Notes

[1] The VNS closely parallels the CEFR. In this study, we worked with the English version of the CEFR, which was readily available, and not an English translation of the VNS. A cross-walk of the two documents by members of the expert panel affirmed their consistency and the reasonableness of working from the CEFR. We will, nonetheless, refer to the VNS in this report.

[2] One educator did not participate in the standard setting for the speaking test and one did not participate for writing test; both absences were due to unrelated personal commitments.

[3] The one educator who did not participate in the standard setting for writing did not complete a final evaluation.

[4] The standard errors of measurement reflect the amount of test score uncertainty, on average. Each individual test score may have more or less associated uncertainty than the average value. But the use of the average offers a reasonable approximation.