

Mapping TOEIC® Test Scores to the STANAG 6001 Language Proficiency Levels

Richard J. Tannenbaum

Patricia A. Baron

September 2010

ETS RM-10-11



Mapping TOEIC® Test Scores to the STANAG 6001 Language Proficiency Levels

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

September 2010

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

<http://www.ets.org/research/contact.html>

Technical Review Editor: Daniel Eignor
Technical Reviewers: Michael Zieky and Donald Powers

Copyright © 2010 by Educational Testing Service. All rights reserved.

ETS, the ETS logo, LISTENING. LEARNING. LEADING, *Test of English for International Communication*, and TOEIC are registered trademarks of Educational Testing Service (ETS).



Abstract

The Standardization Agreement (STANAG 6001) describes language proficiency levels associated with listening, speaking, reading, and writing. Six levels define each skill area: 0 (*no proficiency*), 1 (*survival*), 2 (*functional*), 3 (*professional*), 4 (*expert*), and 5 (*highly-articulate native*). The North Atlantic Treaty Organization (NATO) developed these levels to define the general English proficiency (non-military specific) of military personnel. This study linked TOEIC® scores to the STANAG levels based on expert judgments. A panel of 15 experts in English language instruction and assessment from seven NATO countries participated. An alignment process was first conducted to determine which STANAG levels were sufficiently addressed by the TOEIC assessment to merit recommending corresponding cut scores. Standard setting methodologies were implemented to construct the cut scores.

Key words: STANAG, TOEIC, proficiency levels, standard setting, cut scores

Acknowledgments

We extend our sincere appreciation to Catherine Meyer, our colleague from the ETS Global office in Paris, for her recruitment of the standard setting panelists and her overall assistance with planning the standard setting study. We also offer our gratitude to our other ETS Global colleague, Françoise Azak, for organizing the logistics of the meeting. We also thank our colleague from ETS Princeton, Craig Stief, for his work on the various rating forms and onsite scanning0

Table of Contents

	Page
Panelists	2
Pre-meeting Assignment: Familiarization with STANAG and TOEIC	2
Alignment	4
Method	5
Results	6
Standard Setting	7
Method	7
Listening and Reading sections..	8
Writing and Speaking sections.....	10
Results	11
Recommended cut scores.....	11
TOEIC Listening section.	12
TOEIC Reading section.	12
TOEIC Writing section.	13
TOEIC Speaking section.....	14
End-of-Study evaluation survey.	14
Conclusions	16
Setting Final Cut Scores.....	18
References	21
Appendix	23

List of Tables

	Page
Table 1 Panelist Demographics	3
Table 2 Round 1 Alignment Judgments.....	6
Table 3 Round 2 Alignment Judgments.....	7
Table 4 Listening: Standard Setting Results	12
Table 5 Reading: Standard Setting Results.....	13
Table 6 Writing: Standard Setting Results.....	14
Table 7 Speaking: Standard Setting Results	14
Table 8 Feedback on Standard Setting Process	15
Table 9 Comfort Level With Final Recommended Cut Scores	16
Table 10 Final Cut Score Recommendations.....	17

The Standardization Agreement (STANAG 6001) describes language proficiency levels associated with listening, speaking, reading, and writing. Six levels define each skill area: 0 (*no proficiency*), 1 (*survival*), 2 (*functional*), 3 (*professional*), 4 (*expert*), and 5 (*highly-articulate native*). The North Atlantic Treaty Organization (NATO) developed these descriptors to define the general English proficiency (non-military specific) of military personnel. NATO member countries may use the STANAG levels as a means of communicating English language requirements and expectations across a range of assignments and postings within the military.

There is no single official assessment to measure an individual's English language proficiency relative to the STANAG levels. Different assessments potentially may be used to mark an individual's existing proficiency and track an individual's progress across proficiency levels. An assessment initially designed and constructed to measure the breadth and depth of English language proficiency defined by the STANAG levels maximizes the alignment of the tested content to the levels, and, as such, increases the likelihood that cut scores may be identified to differentiate between adjacent proficiency levels. Some type of standard setting process is still needed to construct the cut scores, but justification for conducting the standard setting is clear: The tested English language abilities and the English language expectations defined by the STANAG descriptors are aligned.

The *Test of English for International Communication*[™] (TOEIC®) is a measure of general English language ability. It is proffered as one assessment that may be used to gauge an individual's standing with respect to the STANAG levels. The assessment consists of 100 selected-response reading questions; 100 selected-response listening questions; eight writing tasks, scored using rubrics; and 11 speaking tasks, scored using rubrics. The purpose of this study was to construct minimum test scores (cut scores) for each of the skill areas assessed by TOEIC that correspond to the different STANAG proficiency levels. Two different standard setting procedures (described below) were implemented to construct cut scores. A cut score differentiates adjacent proficiency levels; it separates the highest performance in the lower proficiency level (e.g., *survival*) from the lowest performance in the next higher proficiency level (e.g., *functional*). A separate set of cut scores is constructed for each of the four skill areas tested.

Although the TOEIC assessment measures the same four skill areas as the STANAG, listening, speaking, reading, and writing, the assessment was not specifically designed to

measure these skills in the same way or to the same degree as described by the STANAG descriptors. As a consequence, before engaging in standard setting, we conducted an alignment process to determine which STANAG proficiency levels — separately for each of the four skill areas — were addressed by TOEIC. It was only for those *matches* between tested skill area and STANAG level that we proceeded to construct cut scores. In this report we describe the methods, procedures, and results of the alignment and standard setting processes. The methods and results specific to each process are presented in separate sections. The implications of the results are presented in a general conclusion section.

Panelists

Fifteen individuals representing seven NATO countries served on the panel that participated in the alignment and standard setting processes. They all had expertise in English language instruction, development, and assessment, with 8 members reporting more than 10 years of such experience; no member had fewer than 5 years of experience. Most were active members of a branch of the military (e.g., navy, air force, ground forces), and several served one or more branches of the military as a specialist, director, or professor in a language training center. Because the study was conducted at the request of the French military, the largest number of experts was from France. Table 1 provides a description of the self-reported demographics of the experts. (See the Appendix for panelist affiliations.)

Pre-meeting Assignment: Familiarization with STANAG and TOEIC

Before the experts assembled for the alignment and standard setting meeting, they were asked to review the STANAG 6001 Proficiency Levels and to take the TOEIC assessment. The goals were for the experts to come to the meeting with a sound understanding of (a) the STANAG descriptors, how each level was defined and what differentiated the levels, and (b) the content of the TOEIC assessment, how each skill was measured and the difficulty of each of the assessment questions that comprised each skill area. Each expert was e-mailed a copy of the STANAG descriptors and asked to write down two or three summary statements that clearly defined the transition from one level to the next. They were prompted to consider how they would know when an individual has progressed, for example, from Level 1 to Level 2, from Level 2 to Level 3. What can an individual who is a low performing Level 3 do that a high performing Level 2 cannot do? Each expert was asked to write down summary statements for

each of the four skill areas across Level 1 (*survival*) through Level 4 (*expert*). The extremes of the STANAG Levels, 0 (*no proficiency*) and 5 (*highly-articulate native*), were not part of the assignment, as they were succinctly defined by STANAG. The experts were asked to bring their notes to the meeting.

Each expert was also provided with an opportunity to take all four sections of the TOEIC assessment. The assessment was accessed through a secure website so that the experts could take the assessment at a location and time of their choosing. Before gaining access to the assessment, each expert signed and returned a non-disclosure/confidentiality form. Direct experience with the assessment is necessary for the experts to understand not only the scope of the assessment, what is and is not being measured, but also the difficulty of the questions.

Table 1

Panelist Demographics

Variable		<i>N</i>
Gender	Female	7
	Male	8
Position	Head of National Language Testing Center	1
	Head of Language Department	1
	Head of Language and Learning Center	2
	Senior Tester	2
	Language Specialist (Instructor)	6
	Language Professor	3
Years of Experience	5–10 years	7
	11–15 years	3
	16–25 years	2
	26–35 years	3

Table continues

Table 1 (continued)

Variable		<i>N</i>
Country	Belgium	1
	Denmark	1
	France	7
	Italy	1
	Poland	3
	Romania	1
	United States of America	1

Alignment

If the TOEIC assessment is to be used as a measure of the STANAG proficiency levels, indicating through test scores an individual's current level and progress across levels over time, then the tested content and the proficiency level descriptors need to be aligned. Alignment in this case may be considered a process of gathering evidence of the extent to which the assessment measures each skill area (listening, speaking, reading, and writing) in a way that is consistent with how these skill areas are defined by the STANAG. It is not appropriate to assume that because an assessment measures reading, for example, that how the assessment defines reading agrees with how reading is defined in an external framework. An assessment of reading may focus on decoding and simple comprehension skills, while a framework may place most value on complex comprehension skills and fluency. The assessment and framework each may legitimately claim to reflect reading skills, but it would be inappropriate to claim they measure the same aspects of reading.

Evidence of alignment justifies the use of assessment scores as a means of making interpretations relative to an external framework such as the STANAG; such evidence, therefore, is a critical component of validity (Bhola, Impara, Buckendahl, 2003; D'Agostino, Welsh, Cimetta, Falco, Smith, VanWinkle, & Powers, 2008). Our alignment effort focused on identifying for each skill area which particular levels of STANAG were adequately addressed by the corresponding section of the TOEIC assessment.

Method

At the start of the meeting, we provided the experts with an overview of the general purpose of the meeting, which was to identify scores of the TOEIC assessment that corresponded to the STANAG levels, and explained the experts' role in the process of identifying the scores. The first day of the meeting, however, was devoted to alignment. We explained that because the assessment was not specifically designed to measure the four skill areas as defined by the STANAG, it was necessary to consider the extent to which the two were aligned. Prior to engaging in the actual alignment process, the experts independently reviewed the STANAG level descriptions for each skill and the four TOEIC assessment sections. A whole-panel discussion followed in which the experts were asked to share their perceptions about the meaning of the levels and what differentiated one level from the next higher level, and to share their perceptions of what the assessment measured and what might make some types of questions challenging for English language learners. Although the experts had reviewed the levels and had taken the assessment before the meeting, this exercise reinforced that preparatory experience and helped to make known the different perceptions of the experts.

The alignment judgment was whether the experts perceived that the assessment section (e.g., Listening) met the general expectations delineated by a STANAG level. The alignment judgment was holistic. The experts considered all the questions that formed the Listening section, for example, and considered whether the measured listening skills, collectively, addressed the STANAG listening skills described, collectively, at each level (1, 2, 3, 4, and/or 5). Level 0, *no proficiency*, was not relevant. The form to record the alignment judgments was a simple matrix with the four TOEIC sections listed in rows and the five STANAG levels listed in columns. An expert filled in a cell if perceiving alignment.

We trained the experts in the alignment process, including how to record their judgments. We explained that an alignment between an assessment section and a level would be counted if at least 10 of the 15 experts (67%) filled in that cell; in other words, a *clear majority* needed to indicate the existence of an alignment. Two rounds of alignment judgments occurred. The first round consisted of independent judgments. We tallied the results and presented them to the experts, who were asked to share their judgment rationales and reactions to the Round 1 results. In addition to presenting their Round 1 judgments, we

also shared two performance samples (actual test taker responses) with the experts, one for writing and one for speaking. Each sample had earned the highest total raw score possible for the Writing section and the Speaking section. The samples were intended to help clarify the extent of writing and speaking proficiency these two sections of the assessment could elicit from test takers. (This was not feasible to do for the two selected-response sections of the assessment.) The experts then completed a second round of judgments for the four skill areas. While they were not obligated to change their Round 1 judgments, this was an opportunity for them to do so, if the Round 1 discussion persuaded them to reconsider their initial judgments of alignment.

Results

The Round 1 results are presented in Table 2. The numbers represent the count of experts who judged there to be an alignment between a TOEIC section and a STANAG level. The criterion for a positive alignment was 10 of 15 (67%). The Listening section was judged to be aligned with STANAG Levels 1 (*survival*) and 2 (*functional*). Reading was aligned with Levels 2 and 3 (*professional*). Writing was aligned with Levels 1, 2, and 3; but Speaking, the other productive skill, was aligned only with Levels 1 and 2. Levels 4 (*expert*) and 5 (*highly-articulate native*) were not judged to be addressed by any of the TOEIC sections.

Table 2
Round 1 Alignment Judgments

TOEIC®	STANAG Levels				
	1	2	3	4	5
Listening	11*	15*	8	2	0
Reading	7	13*	10*	3	0
Writing	14*	15*	10*	3	0
Speaking	13*	14*	8	3	2

Note. A positive alignment is indicated by the asterisk (*).

The Round 2 results are presented in Table 3. Although there were slight changes to the counts, the outcomes remained unchanged from Round 1. Listening and Speaking were aligned with Levels 1 and 2; Reading with Levels 2 and 3; and Writing with Levels 1, 2, and 3.

Table 3***Round 2 Alignment Judgments***

TOEIC®	STANAG Levels				
	1	2	3	4	5
Listening	12*	15*	9	0	0
Reading	8	14*	10*	1	0
Writing	14*	15*	10*	2	0
Speaking	14*	15*	7	5	3

Note. A positive alignment is indicated by the asterisk (*).

Standard Setting

The purpose of the standard setting portion of the study was for the panel of experts to construct minimum TOEIC assessment scores (cut scores) corresponding to STANAG proficiency levels identified by the alignment process (see Table 3). Separate cut scores were constructed for each of the four TOEIC sections (Listening, Reading, Writing, and Speaking). The cut scores delineate the lowest score judged necessary to enter a proficiency level.

Method

A modified Angoff method (Brandon, 2004; Cizek & Bunch, 2007), coupled with a holistic judgment, was applied to the Listening and Reading (multiple-choice) sections and a variation of a Performance Profile [Sample] approach (Zieky, Perie, & Livingston, 2008) was applied to the Writing and Speaking (productive-response) sections. The specific implementation of these methods followed the work of Tannenbaum and Wylie (2008). In that study, cut scores were constructed linking TOEIC to the Common European Framework of Reference (CEFR). Recent reviews of research on standard-setting approaches also reinforce a number of core principles for best practice: careful selection of panel members and a sufficient number of panel members to represent varying perspectives, sufficient time devoted to develop a common understanding of the domain under consideration, adequate training of panel members, development of a description of each performance level, multiple rounds of judgments, and the inclusion of data where appropriate to inform judgments (Brandon, 2004; Cizek, 2006; Hambleton & Pitoniak, 2006). The approaches used in this study adhere to all of these principles.

Listening and Reading sections. The results of the alignment process indicated that the TOEIC Listening section addressed STANAG Levels 1 and 2 and that the Reading section addressed Levels 2 and 3. These levels were the focus of the standard setting. Because the same standard setting method was applied to these two sections, we provide a general description of the method.

First, as noted above, before the meeting, the experts took all four sections of the TOEIC and reviewed the STANAG proficiency level descriptors, paying particular attention to those skills that differentiated one level from the next higher level. These familiarization activities reinforced what was covered by the assessment and the difficulty presented by the specific questions, and reinforced the meaning of the proficiency levels. During the meeting, the panelists defined the minimum skills needed to reach each of the relevant (as identified by the alignment process) STANAG proficiency levels. The panelists first worked in two groups, with each group defining the skills that distinguished a person who was just at the beginning of the proficiency level. We referred to this person as a *just qualified candidate* (JQC). Each group independently defined the JQC for each applicable level and then a whole-panel discussion occurred to reach final agreed upon JQC definitions for each applicable proficiency level (e.g., Levels 2 and 3 for Reading). These final definitions were what the panelists used to guide their standard setting judgments; that is the panelists considered the assessment questions in relation to these definitions.

Panelists were trained in the standard setting process and given an opportunity to practice making their judgments. At this point, panelists were asked to sign a training evaluation form confirming their understanding and readiness to proceed, which all panelists did. Then they went through three rounds of operational judgments, with feedback and discussion between rounds. In Round 1, for each question, each panelist judged the percentage of 100 JQCs for the particular STANAG level who would know the correct answer. Panelists used the following judgment scale (expressed as percentages): 0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100. Panelists were instructed to focus only on the skill demanded by the question and the skill possessed by JQCs, and not to factor guessing into their judgments. Panelists made judgments for a question across the appropriate STANAG level before moving onto the next question. For example, panelists completed their

judgments for the first Listening question for Levels 1 and 2, before moving onto the second Listening question.

The sum of each panelist's cross-question judgments (divided by 100) represents his or her recommended cut score. Each panelist's recommended cut score was provided to the panelist. The panel's average (panel's recommended cut score), median, and highest and lowest cut scores (unidentified) were compiled and presented to the panel to foster discussion. Panelists were then asked to share their judgment rationales. As part of the feedback and discussion, we provided the panelists with the percentage of test takers who answered each question correctly (*P+ values*). In addition, *P+* values were calculated for test takers scoring at or above the 75th percentile on that particular section (i.e., the top 25% of test takers) and for candidates at or below the 25th percentile (i.e., the bottom 25% of test takers). Examining question difficulty for the top 25% and bottom 25% of test takers was intended to give panelists a better understanding of the relationship between overall language ability for that TOEIC section and each of the questions. The partitioning, for example, enabled panelists to see any instances where a question was not discriminating, or where a question was found to be particularly challenging or easy for test takers at the different ability levels.

The question-level data were from a recent administration to test takers from Korea and Japan; these are the two countries where the TOEIC assessment is predominantly administered. The panelists were informed of this and discussed how representative they believed the Asian test-taking population was of the European, military test-taking population. (Although both the STANAG proficiency levels and the TOEIC assessment address general English, and not military-specific English, the focal test-taking population for this meeting was military personnel.) The panelists expected that most Asian test takers would have been exposed to many more *drill and practice* opportunities than typical European, military test takers, so they directed most of their consideration and discussion on the panel-focused feedback.

In Round 2, each panelist made a holistic judgment for the TOEIC section. Given the panelist's Round 1 cut score recommendation for a STANAG level, and following the feedback and discussion of the panel's first round of judgments, each panelist indicated a Round 2 cut score for the section for each of the applicable STANAG proficiency levels. The transition to the section introduced a shift from discrete questions judgments (Round 1) to the overall English language skill (e.g., Listening). This holistic approach seemed more relevant

and appropriate to the language construct of interest than did deconstructing the construct through another series of question-level judgments. Panelists reported no difficulty with the holistic approach; this approach had been followed in previous standard setting applications (Tannenbaum & Wylie, 2005, 2008). After making their second round of judgments, feedback similar to that in Round 1 was provided, but in addition, the percentage of test takers (Korean and Japanese) who would be classified into each of the applicable STANAG levels based on the panel's Round 2 recommendations was presented and discussed. The panelists then had a final opportunity (Round 3) to change their section-level recommended cut scores.

Writing and Speaking sections. Panelists were trained in the standard setting process and given an opportunity to practice making their judgments; all panelists signed a training evaluation form confirming their understanding and readiness to proceed. Because the same standard setting method was applied to these two sections, we provide a general description of the method.

In this approach, panelists reviewed the questions and corresponding scoring rubrics. They then reviewed samples of test taker responses to the questions. A test taker's set of responses to the questions formed a profile; the weighted sum of the question scores is that test taker's total (section) score. The Writing section consists of eight questions. A test taker's written response to each of the eight questions is that test taker's response profile. The Speaking section consists of 11 questions, so a test taker's spoken responses to the 11 questions is a response profile. Twenty-four response profiles for Writing were presented to the panelists. The profiles were sampled to represent a range of the most frequently occurring total scores. The samples ranged from a low of 7.8 raw points to a high of 26 (the maximum possible raw score). The profiles (the actual written responses) were presented in order by the total score (the weighted sum of the question scores), from lowest to highest. Twenty-one response profiles for Speaking were presented to the panelists. These profiles also were sampled to represent a range of the most frequently occurring total scores. The responses were audio files that were broadcasted to the panelists in total score order, from lowest to highest. The profiles scores ranged from a low of 6.9 raw points to a high of 24 (the maximum possible raw score). Each panelist was also provided with a printed sheet (one for the Writing section and one for the Speaking section) that showed the question scores and the total score for each profile to facilitate their judgment process.

Each panelist was asked to judge the total score that a JQC would most likely earn for each of the applicable STANAG levels (as identified by the alignment process). The applicable levels for Writing were Levels 1, 2, and 3; for Speaking, they were Levels 1 and 2. The total score for Writing ranges from 0 to 26 raw points; for Speaking the range is 0 to 24 raw points. Panelists were able to select any total score, in half-point increments, within the legitimate range for the section, even if there was not a profile that illustrated that specific score. For example, the total score for one writing profile was 10.9 raw points, the next higher writing profile was 11.2 raw points, and the next higher was 12.6 raw points. Still, a panelist could judge, for example, that a JQC for Level 1 Writing would likely earn a score of 11.0, 11.5, 12, or 12.5 raw points. The sampled profiles were intended to anchor the meaning of the total scores, not to restrict judgments.

Three rounds of judgments took place, with feedback and discussion between rounds, similar to that which occurred for the Angoff method applied to the Listening and Reading sections. After Round 1, each panelist received his or her individual recommendations and a summary of the panel's recommendations was displayed. Panelists shared their judgment rationales and were presented with the average question scores earned by test takers (from Korea and Japan), overall, and by those scoring in the bottom 25% and top 25% for the section. Panelists then made their Round 2 judgments. The Round 2 feedback included the percentage of test takers who would be classified into each of the applicable STANAG levels based on the panel's average Round 2 recommendations. The panelists then had a final opportunity (Round 3) to change their recommended cut scores.

Results

The first set of results summarizes the panel's standard setting judgments by round for each of the TOEIC sections. This is followed by a summary of the panel's responses to the end-of-study evaluation survey.

Recommended cut scores. Tables 4 through 7 summarize the results of the standard setting for each of the three rounds of judgments for each of the TOEIC sections. The results are presented in raw scores. Panelists worked in the raw score metric for the standard setting process. (The scaled score equivalents are provided in the conclusion section). Also included is the standard error of judgment (SEJ), which indicates how close the cut score is likely to be to the current cut score for other panels of experts similar in composition to the current panel and

similarly trained in the same standard setting method. The recommended cut score at the end of each round is the average of the panelists' individual recommendations. The final recommended cut score is the Round 3 average.

TOEIC Listening section. The maximum raw score for the Listening section is 100 points. The panel's average recommendation for Listening Level 1 (Table 4) varied somewhat across the three rounds of judgments, going from 28.9 (Round 1) to 26.7 (Round 2) to 27.2 (Round 3). The panel's average recommendation for Level 2, similarly varied across the rounds: 80.5 (Round 1), 79.2 (Round 2), and 78.5 (Round 3). The most variation (highest standard deviation) among the individual panelists occurred in Round 1, which is often the case, as the first round reflects independent judgments, whereas the second and third rounds include feedback and discussion. The large variance in the Round 1 judgments was due to one of the panelists (who provided the minimum) not understanding that she could use the full range of the rating scale; we clarified this during the discussion of the Round 1 judgments. The SEJs for Listening at Round 3 are less than 2 raw score points and provides some confidence that the recommended cut score would be similar were a panel with similar characteristics convened.

Table 4
Listening: Standard Setting Results

	Round 1		Round 2		Round 3	
Levels	1	2	1	2	1	2
Average	28.9	80.5	26.7	79.2	27.2	78.5
Median	28.7	80.8	25.0	80.0	25.0	80.0
Minimum	9.3	64.6	16.8	64.6	18.7	64.6
Maximum	53.3	89.9	40.0	89.1	40.0	89.1
SD	13.6	6.8	6.4	6.1	5.7	6.6
SEJ	3.5	1.8	1.6	1.6	1.5	1.7

TOEIC Reading section. The maximum raw score for the Reading section is 100 points. The panel's average recommendation for Reading Level 2 (Table 5) varied somewhat across the three rounds of judgments, going from 69.9 (Round 1) to 72.6 (Round 2) to 74.4 (Round 3). The panel's average recommendation for Level 3, similarly varied across the

rounds: 94.0 (Round 1), 93.3 (Round 2), and 95.1 (Round 3). The most variation among the panelists' judgments for Level 2 occurred in the first round; the variation for Level 3 tended to be more consistent across rounds, but did increase in Round 2. However, one panelist elected not to provide a second round judgment, out of concern that, although the Reading section was judged overall to be aligned with STANAG Level 3, the panelist did not support that alignment. This panelist did, however, elect to provide a Round 3 judgment following the Round 2 discussion. The SEJ for Reading is less than 4 raw score points for Level 2 and slightly more than 2 points for Level 3.

Table 5

Reading: Standard Setting Results

	Round 1		Round 2		Round 3	
Levels	2	3	2	3	2	3
Average	69.9	94.0	72.6	93.3	74.4	95.1
Median	70.7	97.1	70.0	96.5	70.0	98.0
Minimum	29.0	68.1	35.0	68.1	40.0	68.1
Maximum	93.1	99.5	99.5	100.0	100.0	100.0
SD	18.1	8.4	14.9	9.4	15.3	8.5
SEJ	4.7	2.2	3.8	2.5	3.9	2.2

TOEIC Writing section. The maximum raw score for the Writing section is 26 points. The panel's average recommendation for Writing Level 1 (Table 6) was consistent across the three rounds: 10.6 (Round 1), 10.7 (Round 2), and 10.8 (Round 3). This consistency also was evident for Levels 2 and 3. The averages for Level 2 ranged from 17.8 to 18.5; the range for Level 3 was 24.1 to 24.4. The variation in panelists' judgments decreased slightly across the rounds for Level 1, but increased slightly for Levels 2 and 3. There was, however, a more notable increase in the variability of judgments in the third round for Level 2. The introduction of the classification data at the end of Round 2 resulted in a significant increase in one panelist's recommended cut score, going from 18.0 (Round 2) to 26.0 (Round 3), causing the increased variability. The SEJs across all three rounds were fairly stable, and at Round 3 about 1 point or less.

Table 6***Writing: Standard Setting Results***

	Round 1			Round 2			Round 3		
Levels	1	2	3	1	2	3	1	2	3
Average	10.6	17.8	24.1	10.7	17.9	24.4	10.8	18.5	24.4
Median	10.3	17.8	24.5	10.5	18.0	26.0	10.5	18.0	26.0
Minimum	7.0	12.0	18.5	7.0	12.0	18.5	7.0	12.0	18.5
Maximum	19.0	26.0	26.0	19.0	26.0	26.0	19.0	26.0	26.0
SD	2.7	3.5	2.1	2.6	3.6	2.3	2.6	4.3	2.3
SEJ	0.7	0.9	0.5	0.7	0.9	0.6	0.7	1.1	0.6

TOEIC Speaking section. The maximum raw score for the Speaking section is 24 points. The panel's average recommendations for Speaking Levels 1 and 2 (Table 7) were consistent across the three rounds. The averages for Level 1 ranged from 14.9 to 15.4; the range for Level 2 was 21.0 to 21.6. The variation among panelists' judgments was the highest in Round 1. The variation decreased in Round 2 and then slightly increased in Round 3. The SEJs decreased at Round 2 and remained low, at .4, through Round 3.

Table 7***Speaking: Standard Setting Results***

	Round 1		Round 2		Round 3	
Levels	1	2	1	2	1	2
Average	14.9	21.0	15.4	21.3	15.4	21.6
Median	15.0	21.0	15.0	21.0	15.0	21.0
Minimum	9.0	12.5	13.0	18.0	13.0	18.0
Maximum	18.5	24.0	18.5	24.0	18.5	24.0
SD	2.3	2.8	1.4	1.5	1.5	1.7
SEJ	0.6	0.7	0.4	0.4	0.4	0.4

Note. One panelist had to leave before providing a Round 3 judgment.

End-of-Study evaluation survey. Evidence of procedural validity (Kane, 1994) was collected from the end-of-study evaluation questions. Table 8 summarizes the panel's feedback regarding the general standard setting process. The majority of panelists *strongly*

agreed or *agreed* that the pre-meeting assignment was useful, that they understood the purpose of the standard setting study, that the explanations and training provided were clear and adequate, that the opportunity for feedback and discussion between rounds was helpful, and that the standard setting process was easy to follow. One or two panelists *disagreed* that the explanations were clear, that the opportunity for feedback and discussion was helpful, and that the standard setting process was easy to follow. Nonetheless, they had previously signed-off on their readiness to carry out the standard setting task.

The panelists also were asked to indicate which of the following four factors most influenced their standard setting judgments: the definition of the JQC, the between-round discussions, the cut scores of the other panelists, or their own professional experience. The two most influential factors (*very influential*) were the definition of the JQC (11 panelists) and their own professional experience (12 panelists). The between-round discussions and the cut scores of other panelists were *somewhat influential*.

Table 8
Feedback on Standard Setting Process

	Strongly Agree		Agree		Disagree		Strongly Disagree	
	N	%	N	%	N	%	N	%
The homework assignment was useful preparation for the study.	5	33%	10	67%	0	0%	0	0%
I understood the purpose of this study.	9	60%	6	40%	0	0%	0	0%
The instructions and explanations provided by the facilitators were clear.	12	80%	2	13%	1	7%	0	0%
The training in the standard setting methods was adequate to give me the information I needed to complete my assignment.	10	67%	5	33%	0	0%	0	0%
The explanation of how the recommended cut scores are computed was clear.	9	60%	5	33%	1	7%	0	0%
The opportunity for feedback and discussion between rounds was helpful.	11	73%	3	20%	1	7%	0	0%
The process of making the standard setting judgments was easy to follow.	9	60%	4	27%	2	13%	0	0%

Table 9 presents the panel’s comfort level with the final recommended cut scores. This may be considered another aspect of procedural validity, addressing more specifically the outcome of the standard setting process. The majority of panelists reported being *very comfortable* or *somewhat comfortable* with the cut scores; the modal response across each of the sections was *somewhat comfortable*. However, two panelists reported that they were *somewhat uncomfortable* with the cut scores for Listening. Three panelists reported being *somewhat uncomfortable* with the Reading cut scores and two reported being *very uncomfortable* with the cut scores. Five panelists reported being *somewhat uncomfortable* with the cut scores for Writing, and one panelist reported being *very uncomfortable* with the cut scores. Two panelists each reported being *somewhat uncomfortable* and *very uncomfortable* with the Speaking cut scores.

Table 9
Comfort Level With Final Recommended Cut Scores

	Very Comfortable		Somewhat Comfortable		Somewhat Uncomfortable		Very Uncomfortable	
Listening	3	21%	9	64%	2	14%	0	0%
Reading	3	21%	6	43%	3	21%	2	14%
Writing	2	14%	6	43%	5	36%	1	7%
Speaking	3	21%	7	50%	2	14%	2	14%

Note. One panelist had to leave before responding.

Conclusions

The purpose of this study was to identify minimum test scores (cut scores) on each of the four sections of the *Test of English for International Communication* (TOEIC) corresponding to the different STANAG proficiency levels. The assessment is a measure of general English language skills, and was not designed to be a specific measure of the STANAG levels. Therefore, the first phase of this study was to conduct an alignment process, whereby the experts on the panel identified which particular STANAG levels were addressed by the TOEIC assessment. Two rounds of alignment judgments occurred, with at least 10 of

the 15 experts (67%) indicating the following alignments: Listening (Levels 1 and 2), Reading (Levels 2 and 3), Writing (Levels 1, 2, and 3), and Speaking (Levels 1 and 2).

Standards were recommended for those levels identified by the alignment process. A modified Angoff standard setting approach with a holistic component was applied to the selected-response questions (Listening and Reading sections), and a variation of a Performance Sample approach was applied to the productive-response questions (Writing and Speaking sections). Three rounds of judgments, with feedback and discussion occurred; the feedback included data on how test takers performed on the questions and the percentage of test takers who would have been classified into each of the proficiency levels. Table 10 presents the final cut score recommendations, both in raw score and scaled score metrics.

Table 10
Final Cut Score Recommendations

TOEIC®	STANAG Levels					
	1		2		3	
	Raw	Scaled	Raw	Scaled	Raw	Scaled
Listening	27.2	95	78.5	400	--	--
Reading	--	--	74.4	350	95.1	475
Writing	10.8	80	18.5	150	24.4	200
Speaking	15.4	130	21.6	190	--	--

Note. Raw scores were rounded to the next highest value before converting to a scaled score.

The responses to the end-of-study survey, overall, support the procedural validity of the standard setting. The majority of the panelists indicated, for example, that the standard setting training had prepared them, that the provided instructions and explanations were clear, and that the process was easy to follow. Additional evidence of procedural validity had been collected immediately following the specific training on the standard setting approaches. All panelists had indicated that they were ready to proceed to make their first round of standard setting judgments. Collectively, these results support the quality of the standard setting implementation.

However, several panelists did express some reservations regarding the final recommended cut scores. Six panelists were not supportive of the Writing cut scores; their

comments suggest that the concern was focused on the Level 3 cut score. Although most panelists had judged a positive alignment between the TOEIC Writing section and STANAG Level 3, these few panelists did not think that there were enough questions that they considered to be measuring Level 3 writing ability. One consequence of this was that the recommended cut score for Writing at Level 3 was very high, 24.4 raw points; this corresponds to a scaled score of 200, the maximum possible. This result suggests that the Writing section is not likely a useful measure of Level 3 proficiency. Panelists also expressed concern regarding the Reading and Speaking cut scores. Similar to Writing, the concern for the Reading cut scores was directed mainly at Level 3. Several panelists did not believe that a sufficient number of the questions addressed Level 3 reading skills. They also were less supportive of the first two parts of the TOEIC Reading section, which they did not believe measured important aspects of reading comprehension. They wanted much more measurement addressing comprehension, as they saw that as a critical skill for military personnel. The concern directed at Speaking was that the assessment was not interactive and did not necessarily indicate the ability of the test taker to sustain conversation.

This last comment was consistent with an overarching concern directed at the overall TOEIC assessment. Panelists thought that the TOEIC assessment lacked *face validity* for military personnel. Even though they knew that the assessment was a measure of general English language skills, they were concerned that there was no military context represented. Several panelists suggested that ETS consider developing an English language assessment specifically designed to measure the STANAG proficiency levels.

Setting Final Cut Scores

The standard setting panel is responsible for *recommending* cut scores. Policymakers consider the recommendation, but are responsible for *setting* the final cut scores (Kane, 2002). In the current context, policymakers are NATO member countries that need decision rules for communicating English language requirements and expectations across a range of assignments and postings within the military.

The full range of policymakers' needs and expectations cannot be represented during the process of recommending cut scores. Policymakers, therefore, have the right and responsibility of considering both the panel's recommended cut scores and other sources to information when setting the final cut scores (Geisinger & McCormick, 2010). The

recommended cut scores may be accepted, adjusted upward to reflect more stringent expectations, or adjusted downward to reflect more lenient expectations. There is no *correct* decision; the appropriateness of any adjustment may only be evaluated in terms of its meeting the policymaker's needs. Two critical sources of information to consider when setting cut scores are the standard error of measurement (SEM) and the standard error of judgment (SEJ). The former addresses the reliability of TOEIC assessment scores and the latter the reliability of panelists' cut score recommendations.

The SEM allows policymakers to recognize that a test score—any test score on any test—is less than perfectly reliable. A test score only approximates what a test taker *truly* knows or *truly* can do on the test. The SEM, therefore, addresses the question: “How close of an approximation is the test score to the *true* score?” A test taker's score likely will be within one SEM of his or her true score 68% of the time and within two SEMs 95% of the time. The scaled score SEM for TOEIC Listening is 25, for Reading it is 25, for Writing it is 20, and for Speaking it is 15.

The SEJ allows policymakers to consider the likelihood that the current recommended cut score would be recommended by other panels of experts similar in composition and experience to the current panel. The smaller the SEJ, the more likely that another panel would recommend a cut score consistent with the current cut score. The larger the SEJ, the less likely the recommended cut score would be reproduced by another panel. The SEJ, therefore, may be considered a measure of credibility, in that a recommendation may be more credible if that recommendation were likely to be offered by another panel of experts. An SEJ no more than one-half the size of the SEM is desirable because the SEJ is small relative to the overall measurement error of the test (Cohen, Kane, & Crook, 1999). The SEJs in this study were in the raw score metric. We approximated the average scaled score change due to the SEJs by applying the raw-to-scale score conversions for each of the TOEIC sections. In all but one instance, the SEJ resulted in an average scaled score change less than one-half that of the scaled SEM. The exception was Reading Level 2, where the average scaled score change was approximately four-fifths of the scaled SEM.

In addition to measurement error metrics (e.g., SEM, SEJ), policymakers should consider the likelihood of classification error. That is, when adjusting a cut score, policymakers should consider whether it is more important to minimize a false positive

decision or to minimize a false negative decision. A false positive decision occurs when a test taker's score suggests one level of ability, but the person's actual level of ability is lower (i.e., the person does not possess the required skills). A false negative occurs when a test taker's score suggests that the person does not possess the required skills, but actually does possess those skills. It may be the case, for example, that a policymaker needs to place enlisted personnel into a position that necessitates they have at least Level 2 English Listening skills. The consequence of a false positive in this instance—assigning personnel to that position, when they may not be at Level 2—might be significant enough that this possibility must be minimized. In this case, the policy maker might consider raising the recommended scaled cut score of 400 (Listening Level 2), further decreasing the likelihood that false positive decisions occur. (However, this also increases the likelihood of false negative outcomes, denying placement into the position of personnel who may be at Level 2 Reading.) There may also be instances where lowering one or more of the recommended cut scores is reasonable; this would be an instance where the consequence of false positive decisions is less significant than the consequence of false negative decisions. For example, there could be a position that may also require Level 2 Reading skills, but there is adequate opportunity on the job to improve these skills without significantly jeopardizing job performance. Policymakers need to consider which decision error to minimize; it is not possible to eliminate both types of decision errors simultaneously.

References

- Bhola, D. S., Impara, J. C., & Buckendahl, C. W. (2003). Aligning tests with state's content standards: methods and issues. *Educational Measurement: Issues and Practice*, 22, 21–29.
- Brandon, P. R. (2004). Conclusions about frequently studies modified Angoff standard-setting topics. *Applied Measurement in Education*, 17, 59–88.
- Cizek, G. C. (2006). Standard setting. In S.M. Downing & T.M. Haladyna (Eds.), *Handbook of Test Development* (pp. 225–258). Mahwah, NJ: Lawrence Erlbaum Publishers.
- Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: SAGE Publications.
- Cohen, A. S., Kane, M. T., & Crooks, T. J. (1999). A generalized examinee-centered method for setting standards on achievement tests. *Applied Measurement in Education*, 12(4), 343–366.
- D'Agostino, J. V., Welsh, M. E., Cimetta, A. D., Falco, L. D., Smith, S., VanWinkle, W. H., & Powers, S. J. (2008). The rating and matching item-objective alignment methods. *Applied Measurement in Education*, 21, 1–21.
- Geisinger, K. F., & McCormick, C. A. (2010). Adopting cut scores: Post-standard-setting panel considerations for decision makers. *Educational and Psychological Measurement*, 29, 38–44.
- Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R.L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 433–470). Westport, CT: Praeger.
- Kane, M. (1994). Validating performance standards associated with passing scores. *Review of Educational Research*, 64, 425–461.
- Kane, M. T. (2002). Conducting examinee-centered standard-setting studies based on standards of practice. *The Bar Examiner*, 71, 6–13.
- Tannenbaum, R. J., & Wylie, E. C. (2008). *Linking English language test scores onto the Common European Framework of Reference: An application of standard setting methodology* (TOEFL iBT Series Rep. No. TOEFLibt-06, RR-08-34). Princeton, NJ: ETS.

Tannenbaum R. J., & Wylie, E. C. (2005). *Mapping English Language Proficiency Test Scores onto the Common European Framework* (TOEFL Research Report RR-80). Princeton, NJ: ETS.

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

Appendix
Panelist Affiliations

<i>Panelist</i>	<i>Affiliation</i>
Michael Adubato	Supreme Headquarters Allied Power Europe (SHAPE), Belgium
Lieutenant Colonel Pierre Barbe	Gendarmerie Nationale, France
Mary Jo Di Biase	Army Language School, Italy
Capitaine Patrice Fauchard	Air Force, France
Ulla Gudnason	Royal Danish Defence College, Denmark
Capitaine Anthony Hascoët	Field Forces, France
Corina Ispas	Ministry of Defense, Romania
Didier Le Gall	Air Force, France
Enseigne de Vaisseau Sylie Lisslillour	Navy, France
Pawel Madej	Land Forces Command, Poland
François Maguin	Air Force, France
Frank Marion	Navy, France
Katarzyna Misiejuk	Land Forces Command, Poland
Justyna Odymala	Land Forces Command, Poland
Karen Sacre	Defense Language Institute English Language Center, United States [retired]