Research Memorandum
ETS RM–22-01

Mapping *TOEIC*® Writing Test Scores to the American Council on the Teaching of Foreign Languages (ACTFL) Proficiency Guidelines for Writing

Shinhye Lee
Kathryn Hille
Renka Ohta

November 2022

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*® Writing Test Scores to the American Council on the Teaching of Foreign Languages (ACTFL) Proficiency Guidelines for Writing

Shinhye Lee, Kathryn Hille, and Renka Ohta
ETS, Princeton, New Jersey, United States

November 2022

Corresponding author: Shinhye Lee, Email: SLEE004@ets.org

Suggested citation: Lee, S., Hille, K., & Ohta, R. (2022). *Mapping* TOEIC® *Writing test scores to the American Council on the Teaching of Foreign Languages (ACTFL) Proficiency Guidelines for Writing* (Research Memorandum No. RM-22-01). ETS.

Action Editor: Jonathan Schmidgall

Reviewers: Ching-Ni Hsieh and Alexis Lopez

Abstract

Proficiency frameworks and standards have been widely used as common reference points, not only in direct service for educators as educational road maps, but also for local/international testing programs as they strive to provide clarity and transparency in score interpretations and uses. One common example of the latter, which is also the primary focus of the current research report, is when test developers engage in the so-called standard setting process. In this paper, we report a standard setting study that aimed to establish an interpretive link between *TOEIC*® Writing test scores and the American Council on the Teaching of Foreign Languages (ACTFL) Proficiency Guidelines for writing. Following well-established procedures and methods recommended in the field, we describe the process of evaluating construct congruence, composing and training an expert panel, and incorporating multiple sources of information to recommend final cut scores. In addition, we present numerous types of information that function as evidence supporting the validity of the procedures and outcomes of the study.

Keywords: TOEIC® Writing test, ACTFL Proficiency Guidelines, standard setting, cut scores, score interpretation

Table of Contents

The purpose of this paper is to report on a standard setting study that aimed to identify the cut scores on one module within the *TOEIC®* suite of assessments—the TOEIC Writing test—that correspond to the sublevels of the American Council on the Teaching of Foreign Languages (ACTFL) Proficiency Guidelines (ACTFL, 2012b). In so doing, we aim to establish "an interpretive bridge" (Tannenbaum & Wylie, 2008, p. 2) between the descriptors of the ACTFL Proficiency Guidelines for Writing and the TOEIC Writing test performances, thereby enhancing score interpretability to score users who operate within contexts where the ACTFL standards are relevant.[1]

Broadly speaking, standard setting denotes a collective set of systematic procedures by which levels of an external framework (e.g., basic, proficient, advanced) are projected—or mapped—onto a test score scale of interest (Kane, 2001). More precisely, standard setting involves applying expert judgments to create one or multiple sets of cut scores and linking them to performance level descriptors (PLDs) or the verbal elaborations of the external level descriptors (Hambleton, 2001). The cut scores are what create interpretive boundaries within a given score scale, which essentially add "qualitative meaning to the quantitative test scores" (Harsch & Malone, 2020, p. 40). In reference to the PLDs, cut scores establish the minimum requirement of the kinds of skills a test taker is expected to exhibit at a particular score level on the scale. In this way, the established cut scores offer score users a clearer, accessible indication of what an examinee is able to do (or not do) at a specific score level, which depending upon assessment contexts, may lead to significant consequences for individuals or policy-making bodies (Cizek, 2012).

Recent years have witnessed the expanding line of work in standard setting in the context of second/foreign language testing, which largely coincides with the appearance and application of international frameworks and scales specific to describing language proficiency (Kenyon & Römhild, 2013). One of the influential proficiency frameworks is the ACTFL Proficiency Guidelines (ACTFL, 2012b). The ACTFL Proficiency Guidelines conceptualizes the development of foreign language proficiency in terms of a hierarchy of global language-use tasks as well as corresponding performance features across five major levels (novice, intermediate, advanced, superior, and distinguished). The first three major levels (novice,

intermediate, advanced) are divided into high, mid, and low sublevels, and the descriptors are further differentiated both *within* and *across* the major levels in terms of what learners can perform consistently at one level relative to the next higher or lower adjacent levels.

Modeling "an independent measure of real-life, communicative abilities" (Clifford, 2012, p. 50), one key premise of the ACTFL Proficiency Guidelines has been the description of language proficiency independent of a particular language learning theory or teaching methodology (ACTFL, 2012b). As such, the descriptors tend to be in the form of action-oriented statements; that is, they intend to describe what language learners *can do* as opposed to what they *should do* with the language. Such a functional approach, although at times being criticized for its lack of "a firm theoretical grounding" (Hudson, 2013, p. 492), has generally led to the guidelines' widespread application and acceptance in a variety of institutional and workplace settings. Specifically, they have been utilized to guide educators as they strive to inform, and at times enhance, both local- and national-level curriculum, instruction, and assessment practices (Abbott & Phillips, 2011). The guidelines have also been the subject of numerous alignment endeavors, particularly in relation to another major language proficiency framework that is its European counterpart, the Common European Framework of Reference (CEFR) for Languages (Council of Europe, 2001; e.g., Tschirner, 2012).

In this research report, we describe the procedures implemented to identify the minimum scores on the TOEIC Writing test corresponding to the levels of writing proficiency set forth in the ACTFL Proficiency Guidelines. Conforming to the professional and technical guidance provided in the literature (e.g., Cizek & Bunch, 2007; Tannenbaum & Cho, 2014), in what follows, we break down the steps taken to facilitate the standard setting process into three large sections. In the first section, we provide evidence of construct congruence between the TOEIC Writing test content and the ACTFL Writing descriptors. We then outline the specific procedures taken to facilitate the standard setting study and the panel-recommended cut scores. We present these cut scores in particular reference to the poststudy adjustments we carried out, which in part were informed by the existing correspondence between TOEIC Writing scores and other external benchmarks that have a hypothesized alignment with ACTFL proficiency levels (e.g., Tannenbaum & Wylie, 2008). In the third and final section, we

document validity evidence supporting the quality of the study procedures and the final cut scores, particularly drawing upon three types of evidence, namely, procedural, internal, and external validity evidence (Council of Europe, 2009; Kane, 2001).

Construct Congruence

An integral preliminary step for a standard setting study is providing adequate evidence of alignment between the test content and a given proficiency framework (Cizek & Bunch, 2007). This process typically involves a content analysis of both the test and the framework to identify whether a reasonable amount of "construct congruence" (Tannenbaum & Cho, 2014, p. 237) between the two exists—that is, the degree to which the specific skill areas as well as proficiency levels described in the framework are of relevance to test performance. Given that most tests are not directly developed based upon an existing framework from the outset (Tannenbaum & Cho, 2014), demonstrating sufficient coverage is critical in justifying the use of test scores for making interpretations relative to the framework descriptors.

Following a brief overview of the TOEIC Writing test, we describe an investigation we undertook to examine the evidence of construct congruence. The purpose of this investigation is to highlight the alignment between the nature of language ability elicited by the TOEIC Writing test tasks and that described in the ACTFL Writing descriptors.

The TOEIC Writing Test

The TOEIC Writing test is designed to evaluate the ability of English language learners to carry out written communication tasks in the context of everyday and workplace environments. The test is delivered by computer and may be administered separately or with the TOEIC Speaking test depending on the user's need. The test comprises eight test tasks in total, developed based on an evidence-centered design (ECD) approach (Mislevy et al., 2003) to support three distinct claims about a learner's writing ability. In the ECD approach, claims about test-taker knowledge, skills, or abilities define the construct of measurement, and tasks are designed to elicit evidence in relation to claims (see Hines, 2010, for a detailed description of the ECD-based task development process). Table 1 summarizes this task—claim relationship.

Table 1. Claims Articulated for the TOEIC Writing Test and Corresponding Test Tasks

Claim	Task type	Number of questions
Claim 1, test takers can produce well-formed sentences	Write a sentence based on a picture	5 (Questions 1–5)
Claim 2, test takers can produce multi-sentence-level texts	Respond to a written request	2 (Questions 6–7)
Claim 3, test takers can produce multi-paragraph-level texts	Write an opinion essay	1 (Question 8)

Underlying the hierarchical ordering of the claims and corresponding tasks is the assumption that task difficulty increases as test takers progress through the test. Along similar lines, it is also assumed that test takers who can successfully manage the tasks supporting the higher level claims (e.g., writing a multi-sentence-length text) are likely to perform well on tasks covering the lower level claims (e.g., producing sentences). Accordingly, a weight-based scoring system is applied that awards the least weight to the sentence-writing task and the greatest weight to the opinion-writing task.

For the first task type, *Write a Sentence Based on a Picture*, test takers view a picture and use two supplied words (or phrases) to write one sentence. The task type is intended to produce evidence in relation to Claim 1 and is evaluated on the relevance of each produced sentence to its pictures as well as test takers' use of appropriate grammar to construct a sentence. In *Respond to a Written Request*, test takers are instructed to provide requested information to questions posed in two email messages. In this way, the task type elicits evidence of proficiency in relation to Claim 2 and aims to evaluate the quality and variety of sentences used as well as aspects related to coherence and language use. Finally, in *Write an Opinion Essay*, test takers plan and write an opinion about a workplace topic and support it with appropriate reasons and examples. This task type is intended to produce evidence of proficiency in relation to Claim 3, which requires additional aspects of writing competence, such as organization and development.

Construct Congruence Evidence

Establishing the conceptual alignment between a test and an existing framework is a necessary task to justify the mapping of test scores to that framework (Tannenbaum & Cho,

2014). Although scores on a given test may not map to every level of a framework due to design differences (Tannenbaum & Wylie, 2008), an examination of construct congruence can investigate how the skills and abilities to be considered in the score-mapping process are represented in both the test and the targeted frameworks. In addition, the context-independent nature of the frameworks necessitates careful consideration to effectively apply the descriptors to the specific context in which the tests operate (Hudson, 2013). In order to most effectively establish the correspondence between test scores and a given framework, Papageorgiou and his colleagues (in press) suggested taking a holistic approach in lieu of pursuing "point-by-point comparisons" (p. 9) of minute details and content. This approach involves making use of holistic judgments in determining which major skill areas as well as proficiency levels in the framework broadly correspond to those that are intended to be measured by the test.

In the present study, we applied a simplified version of Papageorgiou et al.'s (in press) approach to deriving the holistic judgments. Instead of adopting an elaborated scale of the degree of coverage between the two entities, we sought overall alignment as expressed through expert agreement. In our case, two ETS staff members who had appropriate background/experience with the TOEIC Writing test and ACTFL Proficiency Guidelines were deemed suitable to serve as expert evaluators. The experts consisted of a researcher from the project team, who had experience utilizing and applying the ACTFL Proficiency Guidelines in practice, as well as an assessment developer, who was directly engaged in the item-writing and scoring procedures for the TOEIC Writing test. The former served as the first evaluator and provided initial judgments (i.e., that a given ACTFL sublevel descriptor is elaborating language knowledge or abilities aligned with the language knowledge or abilities evaluated by the test) as well as the rationale supporting these judgments. The second evaluator then reviewed the initial judgments and indicated her agreement (e.g., agreed, not agreed) with those of the first evaluator.

Prior to providing their respective judgments, both evaluators conducted a review of the ACTFL sublevel descriptors, particularly those that were deemed as relevant to describing the range of proficiency targeted by the TOEIC Writing test (i.e., novice to advanced). The

evaluators' review also focused on the seven performance dimensions delineated in the ACTFL Proficiency Guidelines to distinguish the performance of learners as defined across the sublevel descriptors. The performance features included for the review were as follows: *functions*, *contexts and content*, *text type*, *language control*, *vocabulary*, *communication strategies*, and *cultural awareness*.

Functions prescribes a range of distinctive language-use tasks across the ACTFL sublevels, from basic (e.g., producing lists and notes) to more complex (e.g., formal correspondence). *Contexts and content* indicates the situations within which the learner can function (context) and the topics that the learner can handle (content). *Text type* is the unit of language that learners understand and produce in order to perform the functions of the level. This dimension spans the continuum of limited usage (isolated words/phrases) to more elaborative units (strings of sentences and paragraphs). The dimensions of *language control* and *vocabulary* each describe the level of control the learner has over certain linguistic features, organization, or parameters of vocabulary as they perform at a certain level. *Communication strategies* refers to a set of strategies used to negotiate meaning and express oneself in direct reference to an interlocutor/recipient present in the context. Finally, *cultural awareness* depicts the cultural products, practices, or perspectives the learner may employ to communicate more successfully in a given cultural setting. See ACTFL (2012a) for a detailed overview of the performance dimensions.

Table 2 provides the alignment between the ACTFL sublevel descriptors and the TOEIC Writing test content as indicated via the agreements between the two evaluators. The check marks show areas of positive alignment whereas *X* denotes the descriptors for which alignment was not identified.

Table 2. Final Alignment Results of ACTFL Levels and Performance Dimensions

ACTFL levels	Functions	Contexts and content	Text type	Language control	Vocab-ulary	Commun-ication strategies	Cultural aware-ness
Adv. High	✓	✓	✓	✓	✓	X	X
Adv. Mid	✓	✓	✓	✓	✓	X	X
Adv. Low	✓	✓	✓	✓	✓	X	X
Int. High	✓	✓	✓	✓	✓	X	X
Int. Mid	✓	✓	✓	✓	✓	X	X
Int. Low	✓	✓	✓	✓	✓	X	X
Nov. High	✓	✓	✓	✓	✓	X	X
Nov. Mid	X	X	X	X	X	X	X
Nov. Low	X	X	X	X	X	X	X

Note. ACTFL = American Council on the Teaching of Foreign Languages; Adv. = advanced; Int. = intermediate; Nov. = novice; ✓and gray shading = areas of alignment as agreed between the two evaluators; X = evaluators agreed no alignment.

Both evaluators were in complete agreement and came to a consensus that the majority of descriptors, particularly those for sublevels novice high to advanced mid, relate to the language knowledge and skills evaluated by the TOEIC Writing test and thus are suitable for standard setting considerations. The rationale as to the incongruence noted for the two lower novice levels (novice low and novice mid) primarily lay in the mismatch between the features of written performance that learners at those levels would typically exhibit (e.g., word-level production) and the claims that constitute the construct definition for the TOEIC Writing test; that is, the TOEIC Writing test tasks are designed to elicit sentence-level text to a limited extent (e.g., Claim 1). On the other hand, the evaluators perceived the novice high descriptors as corresponding to the entering level writing ability needed to perform on the sentence-writing task; as the descriptors state, learners at this level, albeit inconsistently, are able to provide emerging evidence of producing sentence-level text.

It should be noted that the evaluators and the project team had some discussions about whether the advanced high descriptors might have less overlap with TOEIC Writing. This was due to the expectation that writers in the advanced high level would be able to produce text for wide array of topics and contexts, something that is difficult to elicit evidence of in a practical testing context. However, the project team and evaluators eventually concluded that the

opinion-writing task may provide evidence of the writing skills required to carry out the core advanced high writing functions (e.g., narrating and describing in major time frames, presenting arguments supported by elaboration and examples) and therefore included the advanced high descriptors in the standard setting process.

Judgments also indicated that among the seven performance dimensions, evaluators agreed that two performance features (communication strategies, cultural awareness) were not particularly relevant to what is measured by the TOEIC Writing test tasks. In terms of communication strategies, the rationale was that the test tasks are primarily monological in nature and only indirectly reflect the contexts of real-time, spontaneous written correspondence (e.g., text messaging) wherein active use of strategies may be observed. The dimension of cultural awareness was also deemed as irrelevant as the situations presented in the tasks are intended to be common across a variety of language use domains (ETS, 2016). Therefore, cultural awareness is not considered as a critical task-completion requirement in the TOEIC Writing test context, and thus, not an explicit component of scoring rubrics.

Overall, the findings from the construct congruence phase indicate that the TOEIC Writing tasks demonstrate evidence of various ranges of writing ability as well as high construct congruence for the relevant descriptor scales (e.g., novice high to advanced high; the shaded area in Table 2). The result of this process directly informed the design of the standard setting study, as described in the following section.

Standard Setting Study

Standard setting is first and foremost a socially moderated process (Kenyon & Römhild, 2013) primarily driven by the deliberations and judgments of content and domain-specific experts. Expert judgments form the basis of the final recommended cut scores, which may be applied to enact real-life and, at times, high-stakes decisions about individuals (Cizek & Bunch, 2007). Given the significant consequences that the judgments may bring, well-researched procedures and quality decision-making are key to mitigate possible arbitrariness and eventually yield defensible standard setting outcomes. Key elements of the process include rigorous selection and training of expert panelists, appropriate selection and implementation of a standard setting method, expert facilitators, and a judgment process that is informed by

feedback and discussion. We describe how these elements were integrated into the study procedure in the following two subsections before and during the standard setting meeting.

Before the Standard Setting Meeting

Selection of TOEIC Writing Test Form and Test Data

This standard setting study involved the review of sample test-taker responses to one form of the TOEIC Writing test. A retired operational form of the TOEIC Writing test was selected for this purpose. This form had been administered to more than 1,000 test takers worldwide from 2015 to 2021 and included geographical representation from the regions that typically generate the highest testing volumes. Existing data, including both scoring and test-taker response data for this form, were used in our study.

In order to select representative test-taker responses for the panelists to review, first the scoring data for this form were analyzed. Test-taker data were ordered by total weighted raw score, and the corresponding scale score for each test taker was noted in an approach based on that of Tannenbaum and Baron (2015). Total weighted raw scores for the TOEIC Writing test range from 0 to 26, and scale scores range from 0 to 200, although not all possible scores were represented in the data set. For scale score points with many test-taker responses, the most common total weighted raw score to map to that scale score was taken as representative. Then, within test takers with that total weighted raw score, the pattern of scores on individual items was identified. Because the first five items on the TOEIC Writing test are all the same task type (*Write a Sentence Based on a Picture*) with the same scoring rubric and the same score range (0 to 3), scores across these five items were considered interchangeable for a given test taker (i.e., a scoring pattern of 2-2-3-2-3 on these five items was deemed equivalent to a scoring pattern of 3-3-2-2-2 from another test taker). Similarly, scores on Items 6 and 7 (*Respond to a Written Request*) were considered interchangeable in terms of the scoring pattern (which ranges from 0 to 4 for each item), while scores on Item 8 (*Write an Opinion Essay*) were considered unique.

Using this approach to examine scoring patterns, the most common scoring pattern across all eight items was identified for each total weighted raw score that had been selected in the previous step. Then, test-taker responses with those scoring patterns were reviewed.

Representative samples were selected among those responses that were judged to best preserve both geographical representation among test takers and typicality of content and writing characteristics for a given total weighted raw score point. The project team and assessment developers collaborated, using this process, to select two representative sets of written samples for each scale score point from 80 through 200. For scale score points 40 through 70, there were few test takers, so all available responses needed to be used for these score points. The final result was a collection of 30 sets of test-taker responses, each comprising written responses to the eight writing items, arranged in order from lowest to highest total weighted raw score.

Panelist Selection Criteria and Recruitment

The project team attempted to gather a panel of 15 to 20 individuals with relevant content expertise and familiarity with the target test-taker population of the TOEIC Writing test. In particular, the selection criteria prioritized those who had extensive background in teaching English as a second/foreign language (ESL/EFL; especially the skill of writing) as well as related experience in curriculum and assessment development. Another consideration, although not strictly mandated, was for the panelists to have familiarity with the ACTFL Proficiency Guidelines and preferably the standard setting process. In addition, a number of background characteristics (e.g., gender, age, region), as well as educational and professional experience, were considered relevant with the goal of forming a panel that reflects various perspectives.

With these criteria in mind, the project team first identified potential contact leads (e.g., university professors, university-level ESL/writing program directors and coordinators) through their professional and personal networks who were believed to have access to qualified candidates. The project team asked contact leads to distribute an invitation letter to potential candidates, which included a link to an online screening questionnaire. Interested candidates were then asked to submit their responses to the survey along with their resume or curriculum vitae.

All survey responses were merged into an Excel spreadsheet and candidates were rank-ordered based on years of teaching and curriculum/materials development. In addition, the project team took into account the extent to which the candidates had indicated familiarity

with the TOEIC Writing test, the ACTFL Proficiency Guidelines, and the characteristics of the target test-taker/learner population. In addition to the survey responses, each panelist's resume or curriculum vitae was reviewed thoroughly. In total, 50 individuals completed the questionnaire and 18 of these individuals were invited to participate as panelists, as elaborated further in the next sections (also see Appendix A for a list of panelists and their affiliations). Panelists were then asked to read and sign the consent form and an agreement to not disclose any secure test materials included in the study.

Results From the Screening Survey

The panel represented diverse and balanced profiles of experts, including eight females and 10 males from 13 states in the United States. Eleven of them were between 31 and 40 years old, six were between 41 and 50 years old, and one was between 51 and 60. The panel was limited to experts currently residing within the United States primarily due to some logistical considerations (e.g., time zone differences), but also in part due to the U.S. origin and focus of the ACTFL Proficiency Guidelines.

Regarding their educational and professional background, the panelists were working at the following educational institutions at the time of the study:

- Four-year college/university (14 panelists)
- Community college (three panelists)
- High school (one panelist)

Panelists had different professional positions, including lecturer/instructor (n = 8), professor (6), ESL program assistant director/director (4), ESL program coordinator (1), and an ESL and English teacher (1). Some panelists held multiple positions. All panelists had obtained at least a master's degree, and seven had a doctorate degree. Panelists had an average of 16 years of experience teaching ESL/EFL learners, and 10 panelists reported that they had more than 15 years of teaching experience, including five who taught more than 20 years. In addition to teaching experience, 12 panelists had more than 10 years of experience developing learning materials or assessments for ESL/EFL learners.

In the screening survey, panelists expressed familiarity with the ACTFL Proficiency Guidelines, the TOEIC Writing test, and the target test-taker population. We converted the

familiarity levels to numerical values and generated a total familiarity score, which was then used to compare candidates and to select those who have a better understanding of these areas. The majority of the selected panelists were either very familiar (*n* = 9) or familiar (5) with the ACTFL Proficiency Guidelines, while a smaller portion of panelists noted that they were somewhat familiar (3). Only one panelist reported having a little bit of familiarity. The panelists' familiarity with the TOEIC Writing test was relatively more varied; while more than half of the panelists noted having some extent of familiarity (10), the remaining seven had indicated limited familiarity, with one being not familiar with the test at all. As experienced English language educators, the majority of panelists (15) indicated that they were very familiar with the characteristics and needs of English language learners who may take the TOEIC Writing test, while three noted having moderate familiarity. In terms of prior experience with standard setting, four of the 18 panelists had served on a panel in the past, which underscored the significance of providing quality training and orientation so as to ensure all panelists would have adequate familiarity with the specific standard setting method implemented in the study.

Panelist Preparation Process

To ensure panelists formed a sound understanding of the standard setting process and its key components, the project team provided panelists with a panelist preparation guide 1 week prior to the standard setting meeting. The study preparation required approximately 4 to 5 hours to complete. The guide included information about the ACTFL Proficiency Guidelines for Writing and the TOEIC Writing test, as well as the instructions to the two preparation activities described below. Panelists were required to review the guide carefully and complete the preparation activities prior to the meeting. Both preparation activities were administered using an online survey platform.

Preparation Activity 1 comprised two sections (see Appendix B). The first section asked panelists to read a set of descriptors from the ACTFL Proficiency Guidelines for Writing and sort them first into three major ACTFL levels (advanced, intermediate, and novice) and then into each of the ACTFL sublevels (from novice mid to advanced high) based on their judgment. Note that the descriptors for novice mid were only included to serve as a frame of reference, allowing panelists to compare them with the characteristics of writing at an adjacent level (i.e.,

novice high). The second section of Preparation Activity 1 asked panelists to consider a full spectrum of eight ACTFL sublevels from novice mid to advanced high and come up with three to five distinguishing features that separate one ACTFL sublevel from another. Upon completing the activities on the survey platform, panelists received a copy of their responses via an email as well as feedback/answer keys to the first section.

Preparation Activity 2 asked panelists to complete a mock TOEIC Writing test rendered via the same online survey platform. This simulated test was deemed important given that none of the panelists indicated in the background questionnaire that they were very familiar with the test. In addition, the test tasks they experienced in the mock test were identical to the ones that were used in the standard setting meeting. Thus, this mock test not only provided panelists with an opportunity to form a better understanding of the structure and content of the test, but also to facilitate the judgment process by reducing the time they needed to familiarize themselves with the test tasks.

During the Standard Setting Meeting

After completing the preparation process, panelists participated in a full-day (8 hour) group meeting conducted remotely via Microsoft (MS) Teams (see Appendix C for the meeting agenda). To promote a more comfortable online meeting experience, the project team provided panelists prior to the meeting with detailed information on what to expect during the meeting, step-by-step instructions on solving potential technical problems, and general tips for a successful online meeting. The team was also on standby prior to the meeting time to allow panelists to test various functions of MS Teams and seek technical assistance if necessary.

The project team served as facilitators of the meeting, during which they guided the panelists through a series of activities to orient them to the standard setting process. This set of activities, elaborated in the following subsections, was first preceded by panelists' self-introduction, then followed by a brief overview of the construct definition of the TOEIC Writing test.

The Just Qualified Candidate Definition

The most critical part of the orientation activities was establishing the concept of the just qualified candidate (JQC), also referred to as the borderline candidate. More precisely, the JQC is defined as someone who can be considered "just good enough" to be classified at a given level of proficiency (e.g., ACTFL level novice high; Tannenbaum & Katz, 2013). Setting a defensible set of cut scores hinges in part upon clearly establishing and articulating the JQC descriptions (Hambleton et al., 2012). Yet, the JQC is a challenging concept to grasp given that it can be typically misconstrued as an examinee exhibiting performance features typical of a given proficiency level. Therefore, it was important for the facilitators to use concrete illustrations as well as clarify any confusion among panelists to aid them in clearly internalizing the concept.

Based on the established conceptualization, panelists were asked to develop the JQC descriptions at each ACTFL level of interest—novice high to advanced high. Considering the number of JQC descriptors needed and the limited time allotted for this activity, the panelists were divided into two groups to produce the JQC descriptors for different ACTFL levels simultaneously. To facilitate this process, all panelists participated in a facilitated whole group discussion, the purpose of which was to jointly create the JQC definitions for a midlevel category (i.e., intermediate high) that would serve as an anchor for subsequent JQC descriptors. This step was necessary to model the discussion process that each subgroup would need to follow in the subsequent discussions of other ACTFL levels. To produce each JQC description, panelists were thus asked to focus on the language knowledge and skills needed by the minimally competent candidate at that level; if needed, they were also encouraged to refer to relevant ACTFL performance features as well as Preparation Activity 1, which they had completed prior to the meeting. To contribute to the discussion of the JQC descriptors, panelists were encouraged to either use the chat box or raise their virtual hand on MS Teams while facilitators took notes on a shared screen.

When the panel came to a consensus on the JQC descriptors for intermediate high, they were divided into two groups, each of which met in separate virtual breakout rooms moderated by one of the facilitators. The first group of panelists was assigned to develop JQC descriptors for the three lower levels (i.e., novice high to intermediate mid), and the second group

developed JQC descriptors for the three upper levels (i.e., advanced low to advanced high). The three sets of JQC descriptors produced within each group were then presented to the full panel, who worked together as a group to revise, refine, and finalize the descriptors. The full set of revised JQC descriptors from novice high to advanced high was sent to panelists via email (see Appendix D) so it could be utilized during the judgment procedure.

Standard Setting Training and Judgment Procedure

The standard setting method used in the present study was the Performance Profile method (Zieky et al., 2008), which involves eliciting panelists' judgments on a set of test-taker responses that are ordered from the lowest raw scores to the highest. This method is commonly used with tests that have constructed-response items (such as writing and speaking items) because it allows the panel to review the full performance (or a representative sample thereof) for a given test taker at a given total score and make holistic judgments about the typical abilities of test takers across the score scale. With the Performance Profile method, representative test-taker samples are usually provided for various score points across the score scale to help panelists make informed comparisons between JQC descriptors and actual test-taker performances. In the present study, two representative test-taker samples were selected where possible for each raw score point. However, a restricted number of samples were available to account for the lower ends of the raw score points from the data set used in the study; hence, panelists were presented with a total of 30 test-taker responses for 18 different raw score points, as noted above. Following best practices, panelists were to go through three iterations of judgment tasks during which they were to review the written responses and arrive at their judgments to recommend cut scores.

As part of the training for the standard setting method, facilitators guided the judgment process by asking panelists to review one test taker's responses and compare them to the JQC descriptors developed for ACTFL level intermediate high. In so doing, facilitators encouraged the panelists to navigate the following set of guiding questions: "Is this test taker as able as a JQC at intermediate high? Is the performance less able or more able than a JQC at this level?" If the test taker was deemed less able, then panelists would need to review another set of responses with a higher raw score. If the test taker was as able as the JQC, then they would fill

out the rating form with that raw score. If the test taker was more able, then they would need to review test-taker responses with a lower raw score. These sets of questions, while posed as recommendations, were provided to aid panelists in orienting themselves to the judgment process and filling out the rating form. At the end of the training session, panelists were asked to respond to an online survey (ready-to-proceed form) to indicate their understanding of the process and readiness for completing the judgment tasks.

The three-round judgment process was iterative and standardized to follow a similar structure—panelists individually provided their ratings, followed by feedback and discussion sessions. For the Round 1 judgment, panelists read a series of test-taker writing responses, compared the features of these responses with the established JQC descriptors, and entered that test taker's total raw score to the rating form. They then submitted the finalized rating form to the project team via email and took a break. Between the judgment rounds, facilitators compiled results and generated summary statistics (e.g., the mean, median, mode, maximum and minimum scores, range, standard deviation) to characterize recommended TOEIC Writing test cut scores for the ACTFL proficiency levels considered (novice high to advanced high). Feedback for Rounds 2 and 3 judgments additionally included consequential data (Hambleton et al., 2012) in the form of percentiles, or the percentage of test takers who would be classified at each ACTFL proficiency level. Percentiles were based on the actual test administration data used for the study, which was a useful way to consider the potential impact on actual test-taking populations by looking at the cumulative percentage of examinees at a given recommended raw cut score and their locations along the score scale. This impact data thus gave panelists the opportunity to make sure they were comfortable with the approximate percentages of real-world test takers who would be classified into each ACTFL proficiency level with the cut scores from Round 2, before submitting their final cut score judgments in Round 3.

During the discussion phase, facilitators presented the summary statistics to the panelists, pointing out noticeable trends in the data. For example, facilitators drew the panelists' attention to the ACTFL sublevels for which greater variation in judgments occurred. Panelists were then encouraged to share their decision-making process and rationale for recommending a certain cut score, as well as specific challenges they had encountered during

the process. Therefore, the feedback and discussion sessions were key to helping panelists make informed decisions and potentially identify any misconceptions (with the judgment process, or interpretation of JQC descriptions) that could be of practical consequence. Note, however, that panelists were not obliged to change any of their initial judgments based on the feedback provided unless they saw a particular need to do so. This iterative process continued for the second and third (and final) rounds of judgments.

The panel's final recommended cut scores were computed by averaging 18 panelists' cut scores rounded down to the nearest raw score. These recommended cut scores, along with the accompanying summary statistics, were shared with all the panelists as they completed the final meeting evaluation survey.

Standard Setting Results

The results from the three rounds of panel judgments are summarized in Table 3. The recommended cut scores per ACTFL sublevels are presented as mean raw scores, supplemented by corresponding summary statistics (e.g., median, mode, minimum, maximum, and standard deviation). Accompanying these results are the standard errors of judgment (SEJs) for each recommended cut score. SEJs are used to index the cut-score consistency in standard setting outcomes (Tannenbaum & Katz, 2013) by means of estimating the proximity between the current recommended cut scores to those derived from a panel with similar training and background. The final panel-recommended cut scores are the average of Round 3 ratings.

Raw scores on the TOEIC Writing test range from 0 to 26. The panel showed relatively high convergence after Round 1, although more variance in ratings was evident in the cut scores recommended for the mid-range sublevels (i.e., intermediate mid, intermediate high, advanced low). Convergence in recommendations across the sublevels—based on standard deviations and SEJs—improved as rounds progressed, with the lowest level of variance of judgments observed in Round 3. Across the three rounds, all panelists provided cut scores for each of the seven ACTFL sublevels.

Table 3. Standard Setting Results From Each Judgment Round

Category	Novice	Intermediate			Advanced		
	High	Low	Mid	High	Low	Mid	High
Round 1 (*N* = 18)							
Mean	7.9	9.9	12.4	15.2	17.2	19.1	21.2
Median	7.9	10.0	11.7	15.0	17.0	19.8	20.5
Mode	8.4	8.4	11.0	13.4	16.1	19.8	20.5
Max.	10.3	13.4	16.1	18.8	19.8	21.5	24.3
Min.	6.3	7.4	8.4	11.0	13.4	14.4	17.8
SD	1.2	1.6	2.3	2.2	1.9	1.7	1.9
SEJ	0.3	0.4	0.5	0.5	0.4	0.4	0.4
Round 2 (*N* = 18)							
Mean	7.9	9.7	11.9	14.3	16.6	19.2	21.1
Median	7.9	10.0	11.7	13.9	16.1	19.8	21
Mode	8.4	8.4	11.7	13.4	16.1	19.8	21.5
Max.	10	11.7	15.6	16.1	19.8	20.5	24.3
Min.	6.3	8.4	10.0	11.4	14.4	15.6	19.8
SD	1.1	1.1	1.4	1.4	1.5	1.4	1.0
SEJ	0.3	0.3	0.3	0.3	0.4	0.3	0.2
Round 3 (*N* = 18)							
Mean	7.9	9.6	11.6	14.1	16.6	19.2	21.1
Median	7.9	10.0	11.7	13.4	16.1	19.8	21.5
Mode	8.4	8.4	11.7	13.4	16.1	19.8	21.5
Max.	10.0	11.7	14.4	16.1	18.8	20.5	24.3
Min.	6.3	8.4	10.0	11.4	14.4	15.6	19.8
SD	0.9	1.1	1.0	1.3	1.3	1.4	1.0
SEJ	0.2	0.3	0.2	0.3	0.3	0.3	0.2

Post-Standard-Setting Evaluation

A post-standard-setting evaluation is a process of synthesizing multiple sources of information and viewpoints for making necessary adjustments to the panel-derived cut scores (Geisinger & McCormick, 2010). Typically, this process involves comparing the cut scores recommended by the expert panel to one or multiple sets of alternative solutions (Papageorgiou, Morgan, & Becker, 2015; Schmidgall, 2021). Each of the solutions are then evaluated in particular reference to their psychometric soundness as well as the extent to which they are consistent with the policy-making bodies' needs. Essentially, what underscores these considerations is the desire to minimize the likelihood of classification errors (Ercikan & Julian, 2002)—that is, the possibility of misclassifying an examinee to higher (false positive) or

lower (false negative) level of ability than their actual level of ability. Depending on the score use context, it may be preferable to slightly raise cut scores (thereby minimizing false positive classifications) or slightly lower cut scores (thereby minimizing false negative classifications). Notwithstanding the fact that score users ultimately make the final decisions about prioritizing the minimization of either false positive or false negative classifications (Geisinger & McCormick, 2010), it is also the responsibility of the testing programs to be transparent in how the recommended set of cut scores function (psychometrically) and the impact that may ensue for actual classifications and score interpretations (Papageorgiou, Tannenbaum, et al., 2015).

Our efforts for carrying out the post-standard-setting evaluation are encapsulated in a three-step procedure. First, we applied the raw-to-scale score conversions to the panel-recommended raw scores. After so doing, we considered the coherence between the panel-recommended cut scores and prior alignment results. This involved examining a proposed alignment between ACTFL and CEFR levels (ACTFL, 2012c) and a preexisting mapping between TOEIC Writing test scores and CEFR levels (Tannenbaum & Wylie, 2008; see Table 4). Specifically, we used these two prior studies when making decisions regarding borderline cut scores. We did so by first identifying the corresponding CEFR level for a given proposed TOEIC Writing cut score and then finding the ACTFL level corresponding to that CEFR level. For instance, for novice high, the panel's recommended (raw) cut score initially converted to a location between two scale scores. Among these two, the lower scale score was deemed more appropriate as the scale score cut, given that the mapping between TOEIC Writing scores and CEFR levels (Tannenbaum & Wylie, 2008) implied a considerably lower cut score for CEFR level A1 (30); this was corroborated by the ACTFL-CEFR alignment results (ACTFL, 2012c), which indicated that ACTFL level novice high corresponds to CEFR A1. For advanced high, a higher scale score point of 180 was considered as the scale score cut based upon the TOEIC Writing–CEFR mapping (Tannenbaum & Wylie, 2008), also suggesting a higher scale score for CEFR level C1 (200). This decision was further confirmed by the ACTFL-CEFR alignment results (ACTFL, 2012c), which indicated correspondence between CEFR level C1 and ACTFL level advanced high. No adjustments were made for the remaining levels as the converted cut scores were at least 20 score points apart across all adjacent levels.

Table 4. Correspondence Between ACTFL–CEFR and TOEIC Writing–CEFR

Level	ACTFL-CEFR[a]	TOEIC Writing-CEFR[b]
Advanced High	C1	200
Advanced Mid Advanced Low	B2	150
Intermediate High Intermediate Mid	B1	120
Intermediate Low	A2	70
Novice High	A1	30

Note. ACTFL = American Council on the Teaching of Foreign Languages; CEFR = Common European Framework of Reference for Languages.
[a] ACTFL, 2012c. [b] Tannenbaum & Wylie, 2008.

This set of cut scores served as a baseline solution (Solution 1), which was used to derive two additional sets of cut-score combinations, namely, Solutions 2 and 3. As shown in Table 5, the newly adjusted cut scores were similar to those of Solution 1, except for the midlevel categories (i.e., intermediate mid, intermediate high, advanced low) for which slight adjustments were made. These adjustments were based upon the existing concordance results (ACTFL, 2012c; Tannenbaum & Wylie, 2008), both of which suggested that higher cut scores may be appropriate for several sublevels.

Table 5. Three Cut-Score Solutions

Solution	Novice High	Int. Low	Int. Mid	Int. High	Adv. Low	Adv. Mid	Adv. High
1	50	70	90	110	130	160	180
2	50	70	90	120	140	160	180
3	50	70	100	120	140	160	180

Note. Int. = intermediate; Adv. = advanced.

As a next step, these three solutions were each evaluated statistically in terms of their overall accuracy and reliability of classification. *Accuracy of classification* estimates the likelihood of a test taker's score being classified in the same level as their true score, and *reliability of classification* estimates the likelihood of a test taker being classified into the same level both times if they were to take two different parallel versions of the test (Livingston & Lewis, 1995). Although values over 0.6 are preferred for each of these measures (Powers et al., 2016), their calculations are sensitive to the number of cut scores being implemented, so studies conducted with larger numbers of cut scores will inherently have lower accuracy and

reliability of classification than studies conducted with smaller numbers of cut scores (see Ercikan & Julian, 2002). However, in such cases, these measures can still be successfully utilized to compare alternate solutions and consider the relative psychometric properties of each, as one facet of the post-study evaluation and adjustment process.

For our study, Solution 1 was found to have higher overall accuracy and reliability of classification than the other two solutions. The overall accuracy and reliability of classification values for all three solutions were moderate, but this was anticipated, given the relatively large number of cut scores (7) being implemented. The conditional accuracy and reliability of classification values for the advanced low and advanced mid levels were higher than the overall values, with the conditional accuracy of classification at or above the 0.6 benchmark for both of these levels. Since roughly 70% of the test takers for this operational form earned scores that would place them at either the advanced low or the advanced mid level under Solution 1, the majority of test takers would benefit from the higher conditional accuracies and reliabilities of classification at these two levels. Furthermore, the mean score across forms on the TOEIC Writing test is approximately 147 (ETS, 2021), which is in between the Solution 1 cut scores for advanced low and advanced mid; hence, one can expect that these higher conditional accuracies and reliabilities would apply to a large proportion of TOEIC Writing test takers each year. This higher end of the score range is also where higher stakes decisions, in terms of employment or other opportunities, are likely to be made.

As shown in Table 6, we recommend Solution 1 for mapping the TOEIC Writing test scores to the seven ACTFL sublevels. As described above, these cut scores are primarily based on expert judgment but also supported by statistical evidence and coherence with preexisting mapping and alignment research.

Table 6. Final Score Mapping of TOEIC Writing Test Score and ACTFL Proficiency Levels

Score	Novice Low and Mid	High	Int. Low	Int. Mid	Int. High	Adv. Low	Adv. Mid	Adv. High
TOEIC Writing Score	40 or lower	50–60	70–80	90–100	110–120	130–150	160–170	180–200

Note. ACTFL = American Council on the Teaching of Foreign Languages; Int. = intermediate; Adv. = advanced.

Cut-Score Validation

Cut-score validation concerns the documentation of evidence used to collectively support the defensibility of the standard setting process as a whole and its primary outcomes (i.e., cut scores). Among multiple sources of information, our efforts for the current study focused on addressing three types of validity evidence: procedural, internal, and external evidence.

Procedural Validity Evidence

Procedural validity evidence focuses on the documentation of the standard setting methods and procedures implemented to operate a given standard setting workshop; specifically, this may concern how the implemented procedures were perceived by the individual panelists (Tannenbaum & Cho, 2014). For the current study, procedural validity evidence was collected by surveying the panelists via two sources of evaluation forms: the ready-to-proceed form given at the end of the standard setting training, and the meeting evaluation survey administered at the end of the standard setting meeting. The former survey asked the panel's readiness for the judgment task and the adequacy of the training they have received in regard to the standard setting method. The meeting evaluation contained questions pertaining to the sufficiency, clarity, effectiveness, and appropriateness of the specific elements constituting the standard setting process (e.g., preparation activities, training, explanation by the facilitators).

Table 7 summarizes the results of the ready-to-proceed form. Specifically, the average rating to each of the questions is provided regarding the different types of training and orientation activities provided prior to proceeding to the judgment tasks. The ready-to-proceed form used a 4-point Likert scale (1 = *strongly disagree*, 2 = *disagree*, 3 = *agree*, *4* = *strongly*

agree). Thus, the higher the average, the better the panelists' reported understanding of the procedure, materials, and facilitators' explanations.

Table 7. Panelists' Perception About Training

Statement	1	2	3	4	Average
I understand the purpose of the study.	0	0	3	15	3.8
The facilitator explained things clearly.	0	0	8	10	3.6
I understand the definition of the just qualified candidate (JQC).	0	0	4	14	3.8
The training in the standard setting method adequately prepared me for the judgment task.	0	0	9	9	3.5
I understand how to complete my judgment task.	0	0	12	6	3.7
I am ready to proceed and to complete the judgment task.	Yes		No		
	18		0		

Note. 1 = strongly disagree; 2 = disagree; 3 = agree; 4 = strongly agree.

Overall, panelists indicated that they understood the purpose of the study, the definition of the JQC, and how to complete the judgment tasks. None of the panelists held particularly negative viewpoints over how the meeting was facilitated; there were no indications of either strong disagreements or disagreements with any of the statements. The results also show that all panelists indicated that they were ready to proceed to the judgment rounds, thereby underscoring the clarity and appropriateness of the trainings provided. Concerns, however, were noted in specific regard to the development of the JQC descriptors. In response to an optional, constructed-response question, two panelists expressed that discussions for developing JQC descriptors were particularly constrained by time and stressed the need for a better structuring of the group discussions. Nevertheless, these two panelists concluded that they were ready to proceed. All other comments and questions submitted were addressed by the facilitators during the meeting.

The final meeting evaluation survey provided a basis for collecting and evaluating another set of procedural evidence. The survey included a total of five sections, each of which included a series of statements designed to elicit perceptions on the quality of the standard setting process as a whole. The statements specifically touched upon dimensions such as the

clarity of the instructions provided for the standard setting procedures, the effectiveness of various factors contributing to the cut-score judgments, the adequacy and efficiency of the meeting process, and the comfort level with the final recommended cut scores (Papageorgiou & Tannenbaum, 2016; Tannenbaum & Cho, 2014).

For the first section, panelists indicated the degree to which they agreed with 11 statements that touched upon the clarity of the instructions provided for the various pre- and during-meeting activities. Panelists used a 4-point scale (1= *strongly disagree*, 2 = *disagree*, 3 = *agree*, *4* = *strongly agree*) to provide their judgments. As Table 8 shows, generally, panelists found the pre- and during-meeting activities (e.g., discussion) helpful. They also agreed that the facilitators' explanations and instructions given at various points during the meeting were clear. In addition, the panel indicated that they had obtained a sufficient understanding of the purpose of the study as well as the distinguishing features of writing at various ACTFL sublevels. The only statement that generated less favorable evaluation concerned the definitions of the JQC descriptors. Written comments provided by panelists illuminate relevant challenges. For example, one panelist perceived a lack of consensus at the whole panel-level in terms of what constitute the concept of JQC, which in their view, led to unsuccessful development of the JQC descriptors. Another panelist pointed out that the developed JQC descriptors included language skills that were not directly relevant to the ACTFL level descriptors (see Appendix E for precise comments).

As shown in Table 9, the second section of the survey pertained to the degree to which different factors influenced the panelists for arriving at their cut-score judgments. Panelists used a 3-point scale to indicate their evaluations (1 = *not influential*, 2 = *influential*, 3 = *very influential*). The most influential factor in their evaluation was the list of JQC descriptors, which was followed by a group discussion. This is a desirable and expected result given that JQC descriptors denote the minimum performance features at a given ACTFL level. Other factors influencing panelists' judgements included the between-round discussions, the summary statistics presented after each round of judgment, and panelists' own professional experience. Panelists' summaries of the distinguishing features of ACTFL sublevels (Preparation Activity 1) and the scoring rubrics were evaluated to be the least influential.

Table 8. Panelists' Feedback on Meeting Procedure

Statement	Strongly disagree	Disagree	Agree	Strongly agree	Average
I understood the purpose of the study.	0	0	4	14	3.8
The first part of the preparation activity 1 (sorting ACTFL descriptors in the online survey) was helpful.	1	0	8	9	3.4
The second part of the preparation activity 1 (summarizing the distinguishing features of ACTFL levels Novice Mid to Advanced High) was helpful.	0	2	5	11	3.5
Preparation activity 2 (TOEIC Writing test task experience) was helpful.	0	1	6	11	3.6
I understood the distinguishing features of the 7 ACTFL levels, Novice High to Advanced High.	0	0	8	10	3.6
I was satisfied with the definitions of the just qualified candidates (JQCs) produced and used during the meeting.	1	5	8	4	2.8
The instructions and explanations provided by the facilitators were clear.	0	0	10	8	3.4
The explanation of the process for the judgment task helped me complete my assignment.	0	0	8	10	3.6
The explanation of how the recommended cut scores are computed was clear.	0	1	7	10	3.5
The statistical information presented between rounds was helpful.	0	0	6	12	3.7
Feedback and discussion between judgement task rounds was helpful.	0	1	11	6	3.3

Note. strongly disagree = 1; disagree = 2; agree = 3; strongly agree = 4.

Table 9. Factors That Might Have Influenced Panelist Score Judgment

Factors	Not influential	Influential	Very influential	Average
The just qualified candidate (JQC) descriptors	0	5	13	2.7
The summary I prepared of the distinguishing features of ACTFL levels in the Preparation activity 1	4	13	1	1.8
The scoring rubrics (i.e., the rubrics used by raters to score each task type)	10	7	1	1.5
The discussion of the descriptors to distinguish the 7 ACTFL levels, novice high to advanced high	0	7	11	2.6
The between-rounds discussion	1	13	4	2.2
The summary statistics presented after each round of judgments	0	10	8	2.4
My own professional experience	0	10	8	2.4

Note. Not influential = 1; Influential = 2; Very influential = 3.

Table 10 shows the results of the third section of the survey that pertained to the panelists' perception about the quality of the meeting process. Panelists used a 5-point scale to rate the meeting in terms of the following dimensions: efficiency, coordination, understandability, and satisfaction. The panelists had generally positive perceptions about the meeting, particularly with regard to the coordination. Panelists also reacted positively to how the meeting was efficiently structured and the instructions being understandable, with no one indicating extremely negative perceptions (a rating of 1 or 2). The degree of satisfaction was rated somewhat lower than the other characteristics, with one panelist giving a rating of 2.

Table 10. Rating on the Meeting Process

Characteristics of meeting	1	2	3	4	5	Average
Inefficient (1) – Efficient (5)	0	0	6	6	6	4.0
Uncoordinated (1) – Coordinated (5)	0	0	3	7	8	4.3
Confusing (1) – Understandable (5)	0	0	4	10	4	4.0
Dissatisfying (1) – Satisfying (5)	0	1	7	5	5	3.8

The fourth section of the survey asked panelists to indicate their comfort level with the group's recommended cut scores, using a 4-point Likert scale (1 = *very uncomfortable*, 2 =

somewhat uncomfortable, 3 = *somewhat comfortable*, 4 = *very comfortable*). As shown in Table 11, panelists were generally comfortable with all the seven cut scores recommended, with none expressing a very low level of comfort. In particular, the majority of the panelists were very comfortable with the cut scores recommended for ACTFL sublevels intermediate mid and advanced mid.

Table 11. Panelists' Comfort Level With the Panel's Recommended Cut Scores

ACTFL level	1	2	3	4	Average
Novice High	0	0	9	9	3.5
Intermediate Low	0	1	9	8	3.4
Intermediate Mid	0	0	7	11	3.6
Intermediate High	0	2	8	8	3.3
Advanced Low	0	1	8	9	3.4
Advanced Mid	0	0	6	12	3.7
Advanced High	0	1	9	8	3.4

Note. ACTFL = American Council on the Teaching of Foreign Languages; Very uncomfortable = 1; Somewhat uncomfortable = 2; Somewhat comfortable = 3; Very comfortable = 4.

The last and optional section of the survey provided panelists with the opportunity to share any final comments they had about the study. Comments provided by 17 panelists were qualitatively analyzed to identify common themes and trends that shed light on their experience (see Appendix E for the themes and example quotes for each theme). Panelists' positive comments mostly concerned the facilitators as well as the overall standard setting experience; specifically, they had a positive view of how facilitators organized and led the meeting and expressed their appreciation for gaining new insights of the standard setting process (see Appendix E, comments from participant number 08).

On the other hand, panelists expressed mixed observations on various aspects of the orientation and the judgment activities. In regard to the development of the JQC descriptors, some commented that the purpose of this was unclear, while the others felt that the discussion of JQCs did not revolve around its core definition—that is, the minimum level of competence at a given proficiency level. In addition, panelists preferred smaller groups over the whole-group discussions and needed more time for developing JQCs and between-rounds discussions. In the authors' experience with various standard setting studies, creating and working with JQCs can be inherently challenging for panelists, especially when it is their first time to participate in a

standard setting study. The team did encourage panelists to consider subdimensions (e.g., structure, vocabulary, mechanics) for developing the JQCs at each level as a way to help structure the discussion. Future standard setting studies might consider various other ways of maximizing opportunities for participation, such as subdividing the panelists into smaller groups to develop and review JQCs.

Suggestions were also provided for improving the logistics of the meeting. For example, a few panelists believed the full-day meeting format was burdensome. Although widely adopted for practicality reasons, this meeting format presents both physical and cognitive challenges to panelists, especially those new to the process (Skorupski, 2012). Depending upon a given panel's availability and willingness, future studies can schedule meetings over different days to ease panelists' burden and ensure quality training is provided and received. Where a day-long meeting is the only viable option, studies may consider using well-trained, experienced panelists who are likely to set realistic expectations of the standard setting process (Skorupski & Hambleton, 2005). A few others also expressed concerns about using MS Teams, preferring a different remote-meeting platform that they were usually more familiar with (e.g., Zoom). MS Teams was chosen for its user-friendly functionality and the project team's strong familiarity. However, panelists using a web-browser version of MS Teams applications (versus a desktop version) seem to have issues with using the chat function, seeing a shared screen, or accessing the breakout room. While the project team had anticipated and strived to mitigate such challenges in advance, additional efforts should be taken to ensure a smoother experience (e.g., stationing a technical-assistance staff in the meeting) in addition to exploring and testing the functionality of other meeting platforms.

Internal Validity Evidence

The internal validity of a standard setting study can be supported by evidence of the consistency of panelist judgments as they proceed through the process of setting cut scores (Hambleton et al., 2012). For this study, panelists participated in three rounds of judgments, with summary statistics as well as the standard deviation (SD) and SEJ for each round reported in Table 3. These data support the consistency of panelist judgments in several ways, thereby providing evidence of internal validity for this study.

First, it can be noted that while panelists had the opportunity to change their cut-score decisions in each judgment round, the mean cut-score judgments for each level remained quite consistent across the three rounds. For four of the seven levels (novice high, intermediate low, advanced mid, and advanced high), the mean judgments varied no more than 0.3 score points across the rounds, and for the remaining three levels (intermediate mid, intermediate high, and advanced low—all nearer to the middle of the score scale), the mean judgments varied no more than 1.1 score points. This is a small amount of variation given the 26-point range of the raw score scale.

While the mean judgments for each level remained stable across rounds, the consistency among the panelists in their judgments remained stable or increased with each round. This change can be seen in the stability or reduction of the range of judgments (maximum minus minimum for each level), of the SD and of the SEJ. For example, the range for the intermediate mid cut score decreased from 7.7 points (maximum panelist judgment of 16.1 minus minimum judgment of 8.4) in Round 1 to 5.6 points (maximum of 15.6 minus minimum of 10) in Round 2 and to 4.4 points (maximum of 14.4 minus minimum of 10) in Round 3. This reflected an increase in agreement among panelist judgments and a decrease in more divergent judgments. Similarly, the SD for intermediate mid decreased from 2.3 points in Round 1 to 1.4 points in Round 2 and to 1.0 points in Round 3, further supporting the finding that the overall agreement among panelists increased across rounds. Finally, the SEJ, which is an indication of how similar the judgment of another group of similar panelists would likely be, decreased from 0.5 in Round 1 to 0.3 in Round 2 and to 0.2 in Round 3. In fact, by Round 3, all seven levels had a SEJ in the 0.2 to 0.3 range. This indicates that another panel of similar experts would likely have mean cut-score judgments closely mirroring those of this panel, given the small variation that 0.2 to 0.3 points represents across the 26-point raw score scale.

External Validity Evidence

External validity, also referred to as *convergent validity* (Kane, 2001), pertains to the extent to which the results of the standard setting converge with external objective criteria. Validation of this sort is commonly based on comparing the recommended cut scores with different sources of information, such as other standard setting procedures (Hambleton et al.,

2012) or the performance of the same group of test takers on a related measure (Tannenbaum & Cho, 2014). A comparison of measures, however, presents its own limitations due to comparability issues (Zieky et al., 2008). In addition, divergence noted as a result of the comparison does not necessarily imply the untrustworthiness of the panel-constructed cut scores (Tannenbaum & Wylie, 2008). As such, the external validity evidence presented in the current report is intended to shed insights into the reasonableness and potential applicability of the results derived from the study, rather than their absoluteness. Two sources of information guided the process of external validation: (a) prior concordances reported for the TOEIC Writing and CEFR (Tannenbaum & Wylie, 2008) and the ACTFL Proficiency Guidelines for Writing and CEFR (ACTFL, 2012c) and (b) findings from a criterion validity study of the TOEIC Writing test (Schmidgall & Powers, 2020).

In terms of the former, prior alignment results were combined with the final recommended cut scores resulting from the current study, as shown in Figure 1. Due to the absence of an anchor point that links all three sets of studies (i.e., the same proficiency framework), direct and complete mapping of the previous results to that of the current study is not plausible. This triangulation, however, may hint at potential overlaps for specific proficiency levels across the frameworks. For example, both the scale score ranges recommended for ACTFL sublevel novice high in the current study (50–60) and the CEFR alignment identified for novice high (ACTFL, 2012c) seem to commonly correspond to the CEFR level A1, as evidenced in Tannenbaum and Wylie (2008). Taking this approach, potential correspondences may be hypothesized between the scale score points derived for ACTFL sublevel intermediate low and CEFR level B2. More empirical evidence and score-mapping research, however, is needed to verify any hypothesized relationships as well as the equivalency between the TOEIC Writing scores and external frameworks.

A study by Schmidgall and Powers (2020) offered an additional perspective regarding the external validity of the results of the current study. The authors proposed conceptualizing performance on the TOEIC Writing test in terms of the construct of *functional adequacy*, a criterion operationalized as the impressionistic evaluations provided by naïve readers of the text, otherwise termed as “linguistic laypersons” (p. 46). A total of 100 professionals currently

employed in international workplace settings, who lacked specialized linguistic training, provided holistic judgements of test-taker responses to the email-request and opinion-writing tasks in particular respect to the following subdimensions of functional adequacy: comprehensibility, adequacy of content, effectiveness, coherence, and support.

Figure 1. Concordance Results

ACTFL (2012c)		Tannenbaum & Wylie (2008)	Current study
ACTFL Proficiency Level	**CEFR**	**TOEIC Writing Score (0-200)**	**ACTFL Proficiency Level**
Advanced High	C1	200	Advanced High
Advanced Mid Advanced Low	B2	190	
		180	
		170	Advanced Mid
		160	
		150	Advanced Low
Intermediate High Intermediate Mid	B1	140	
		130	
		120	Intermediate High
Intermediate Low	A2	110	
		100	Intermediate Mid
		90	
		80	Intermediate Low
		70	
Novice High	A1	60	Novice High
		50	
		40	Novice MID / LOW
		30	
N/A	N/A	0-20	

Source: ACTFL (2012c), Tannenbaum and Wylie (2008), and the current study.

The judgment task was conceived as role plays, providing a specific workplace context and purpose for which the professionals are to encounter the written responses. A subset of test-taker response data was pulled from the same test administration and test form used in the current study. Results of the regression analysis indicated that higher performance on the TOEIC Writing test predicts higher perceptions of functional adequacy. More precisely, a functional adequacy score of 4 (based on a range of 0 to 6) is projected onto TOEIC Writing scale score of 120, denoting that linguistic laypersons generally perceive that written responses at this scale score point are comprehensible and effective for meeting a given writing need. This finding may be taken to corroborate the results of the current study, particularly supporting the cut scores and JQC definitions derived for ACTFL sublevels intermediate high and above. In the current study, a scale score point of 120 was mapped onto the ACTFL level intermediate high, a point at which the writing becomes comprehensible to a wider audience, including laypersons who are not accustomed to the writing of language learners (see Appendix D, JQC descriptor for intermediate high). Writing at the intermediate high level is also characterized by meeting most basic and professional writing needs, which may reflect the kind of communicative competence needed to navigate the target-language-use domain defined by the TOEIC Writing test, the everyday and workplace environment (Hines, 2010).

Conclusion

In this paper, we reported a standard setting study that aimed to establish an interpretive link between TOEIC Writing test scores and the ACTFL Proficiency Guidelines for Writing. Following well-established procedures and methods recommended in the field, we described the process of evaluating construct congruence, forming and training an expert panel, and determining recommended cut scores. In addition, we presented key information to support the validity of the procedures and outcomes of the study.

Standard setting is a useful practice for establishing cut scores that can be used by score users to understand the meaning of test scores and make classification decisions. The current study thus has practical implications, particularly in terms of serving the needs of relevant stakeholders and, by extension, prospective test takers of the TOEIC Writing test, whereby high-stakes employment and career-advancing decisions are likely to occur (Tannenbaum, 2013).

However, a given standard setting process is not intended to offer a "true" cut score, nor does it assume a fixed link between the test and the external framework (Cizek & Bunch, 2007). In concluding the paper, we thus emphasize the importance of practicing flexibility in interpreting the final cut scores and considering them as recommendations (North, 2014). That is, score users should consider the recommended cut scores relative to their own particular needs and outcomes, with an eye toward making decisions on which classification errors to minimize.

References

Abbott, M., & Phillips, J. (2011). *A decade of foreign language standards: Influence, impact, and future directions: Survey results.* ACTFL. https://www.actfl.org/sites/default/files/publications/standards/NationalStandards2011.pdf

ACTFL. (2012a). *ACTFL performance descriptors for language learners 2012 edition.* https://www.actfl.org/sites/default/files/publications/ACTFLPerformance_Descriptors.pdf

ACTFL. (2012b). *ACTFL proficiency guidelines 2012*. https://www.actfl.org/sites/default/files/pdfs/public/ACTFLProficiencyGuidelines2012_FINAL.pdf

ACTFL. (2012c). *Assigning CEFR ratings to ACTFL assessments.* https://www.actfl.org/sites/default/files/reports/Assigning_CEFR_Ratings_To_ACTFL_Assessments.pdf

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). Routledge. https://doi.org/10.4324/9780203848203

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Sage. https://doi.org/10.4135/9781412985918

Clifford, R. Y. (2012). It is easier to malign tests than it is to align tests. In E. Tschirner (Ed.), *Aligning frameworks of reference in language testing: The ACTFL proficiency guidelines and the Common European Framework of Reference* (pp. 49–56). Stauffenburg.

Council of Europe. (2001). *The Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge University Press.

Council of Europe. (2009). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR). A manual.*

Ercikan, K., & Julian, M. (2002). Classification accuracy of assigning student performance to proficiency levels: Guidelines for assessment design. *Applied Measurement in Education*, *15*(3), 269–294. https://doi.org/10.1207/S15324818AME1503_3

ETS. (2016). *TOEIC Speaking and Writing tests examinee handbook*. https://www.ets.org/s/toeic/pdf/speaking-writing-examinee-handbook.pdf

ETS. (2021). *2021 report on test takers worldwide—TOEIC Speaking & Writing tests.* https://www.ets.org/s/toeic/pdf/sw-report-on-test-takers-worldwide.pdf

Geisinger, K. F., & McCormick, C. A. (2010). Adopting cut scores: Post-standard-setting panel considerations for decision makers. *Educational Measurement: Issues and Practice*, *29*(1), 38–44. https://doi.org/10.1111/j.1745-3992.2009.00168.x

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 89–116). Erlbaum.

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). Routledge.

Harsch, C., & Malone, M. E. (2020). Language proficiency frameworks and scales. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 33–44). Routledge.

Hines, S. (2010). Evidence-centered design: The TOEIC Speaking and Writing tests. In D. Powers (Ed.), *TOEIC compendium* (pp. 7.1–7.31). ETS.

Hudson, T. (2013). Standards-based testing. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 479–494). Routledge.

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). Erlbaum.

Kenyon, D. M., & Römhild, A. (2013). Standard setting in language testing. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 944–961). John Wiley & Sons.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, *32*(2), 179–197. https://doi.org/10.1111/j.1745-3984.1995.tb00462.x

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*(1), 3–62. https://doi.org/10.1207/S15366359MEA0101_02

North, B. (2014). *The CEFR in practice*. Cambridge University Press.

Papageorgiou, S., Davis, L., Ohta, R., & Gomez, G. G. (in press). *Mapping* TOEFL® Essentials™ *test scores to the Canadian Language Benchmarks* (ETS Research Report Series). ETS.

Papageorgiou, S., Morgan, R., & Becker, V. (2015). Enhancing the interpretability of the overall results of an international test of English-language proficiency. *International Journal of Testing*, *15*(4), 310–336. https://doi.org/10.1080/15305058.2015.1078335

Papageorgiou, S., & Tannenbaum, R. J. (2016). Situating standard setting within argument-based validity. *Language Assessment Quarterly, 13*(2), 109–123. https://doi.org/10.1080/15434303.2016.1149857

Papageorgiou, S., Tannenbaum, R. J., Bridgeman, B., & Cho, Y. (2015). *The association between TOEFL iBT® test scores and the Common European Framework of Reference (CEFR) levels* (Research Memorandum No. RM-15-06). ETS.

Powers, D., Schedl, M., & Papageorgiou, S. (2016). Facilitating the interpretation of English language proficiency scores: Combining scale anchoring and test score mapping methodologies. *Language Testing, 34*(2), 175–195. https://doi.org/10.1177/0265532215623582

Schmidgall, J. (2021). *Mapping the redesigned* TOEIC Bridge® *test scores to proficiency levels of the Common European Framework of Reference for Languages* (Research Memorandum No. RM-21-01). ETS.

Schmidgall, J., & Powers, D. E. (2020). TOEIC Writing test scores as indicators of the functional adequacy of writing in the international workplace: Evaluation by linguistic laypersons. *Assessing Writing*, *46*, Article 100492. https://doi.org/10.1016/j.asw.2020.100492

Skorupski, W. P. (2012). Understanding the cognitive processes of standard setting panelists. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 135–147). Routledge.

Skorupski, W. P., & Hambleton, R. K. (2005). What are panelists thinking when they participate in standard-setting studies? *Applied Measurement in Education*, *18*(3), 233–256. https://doi.org/10.1207/s15324818ame1803_3

Tannenbaum, R. J. (2013). Setting standards on the TOEIC Listening and Reading test and the TOEIC Speaking and Writing tests: A recommended procedure. In D. Powers (Ed.), *TOEIC compendium* (2nd ed., pp. 8.0–8.12). ETS.

Tannenbaum, R. J., & Baron, P. A. (2015). *Mapping 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). ETS.

Tannenbaum, R. J., & Cho, Y. (2014). Critical factors to consider in evaluating standard-setting approaches to map English language test scores to frameworks of language proficiency. *Language Assessment Quarterly, 11*(3), 233–249. https://doi.org/10.1080/15434303.2013.869815

Tannenbaum, R. J., & Katz, I. R. (2013). Standard setting. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology: Vol. 3. Testing and assessment in school psychology and education* (pp. 455–477). American Psychological Association. https://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. 6). ETS. https://doi.org/10.1002/j.2333-8504.2008.tb02120.x

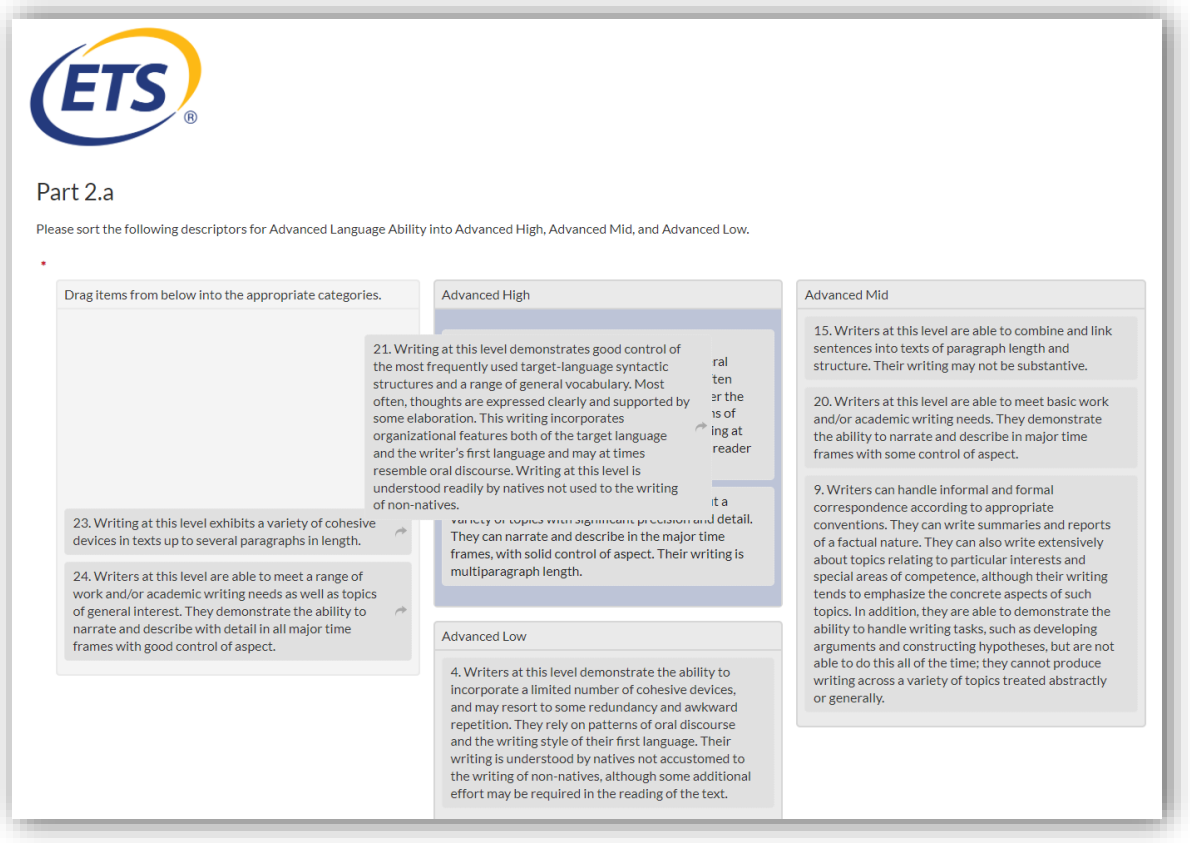
Tschirner, E. (2012). *Aligning frameworks of reference in language testing: The ACTFL proficiency guidelines and the Common European Framework of Reference.* Stauffenburg.

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

Appendix A. The Standard Setting Panelists, Affiliation, and State

Panelist name	Institution	State
Brian Thomas Hampson	Northwestern University	Illinois
David Daniel Sparks	Purdue University	Indiana
Emily Spurgeon	The University of Texas at Austin	Texas
Fiona Hu	Rutgers University	New Jersey
Kyle Patrick Butler	Ohio University	Ohio
Levin Arnsperger	Emory University	Georgia
Mark C Shea	Mount Holyoke College	Massachusetts
Matthew Clark Allen	Purdue University	Indiana
Matthew Jeffrey Kessler	University of South Florida	Florida
Matthias Peter Maunsell	University of Virginia	Virginia
Olga Ellis	The University of Arizona	Arizona
Rebecca Ruth Yeager	The University of Iowa	Iowa
Rhonda Renee Petree	University of Wisconsin - River Falls	Wisconsin
Robert McCarthy Sheppard	Temple University	Pennsylvania
Ryan Dehner	Kirkwood Community College	Iowa
Sally Jo Hatfield	Maumee Valley Country Day School	Ohio
Seo Jung (Linda) Park	Kirkwood Community College	Iowa
Valeriia Bogorevich	Arizona Western College	Arizona

Appendix B. Screenshots of Sample Task for Panelist Familiarization



Part 2.a

Please sort the following descriptors for Advanced Language Ability into Advanced High, Advanced Mid, and Advanced Low.

*

Drag items from below into the appropriate categories.

23. Writing at this level exhibits a variety of cohesive devices in texts up to several paragraphs in length.

24. Writers at this level are able to meet a range of work and/or academic writing needs as well as topics of general interest. They demonstrate the ability to narrate and describe with detail in all major time frames with good control of aspect.

Advanced High

21. Writing at this level demonstrates good control of the most frequently used target-language syntactic structures and a range of general vocabulary. Most often, thoughts are expressed clearly and supported by some elaboration. This writing incorporates organizational features both of the target language and the writer's first language and may at times resemble oral discourse. Writing at this level is understood readily by natives not used to the writing of non-natives.

They can narrate and describe in the major time frames, with solid control of aspect. Their writing is multiparagraph length.

Advanced Low

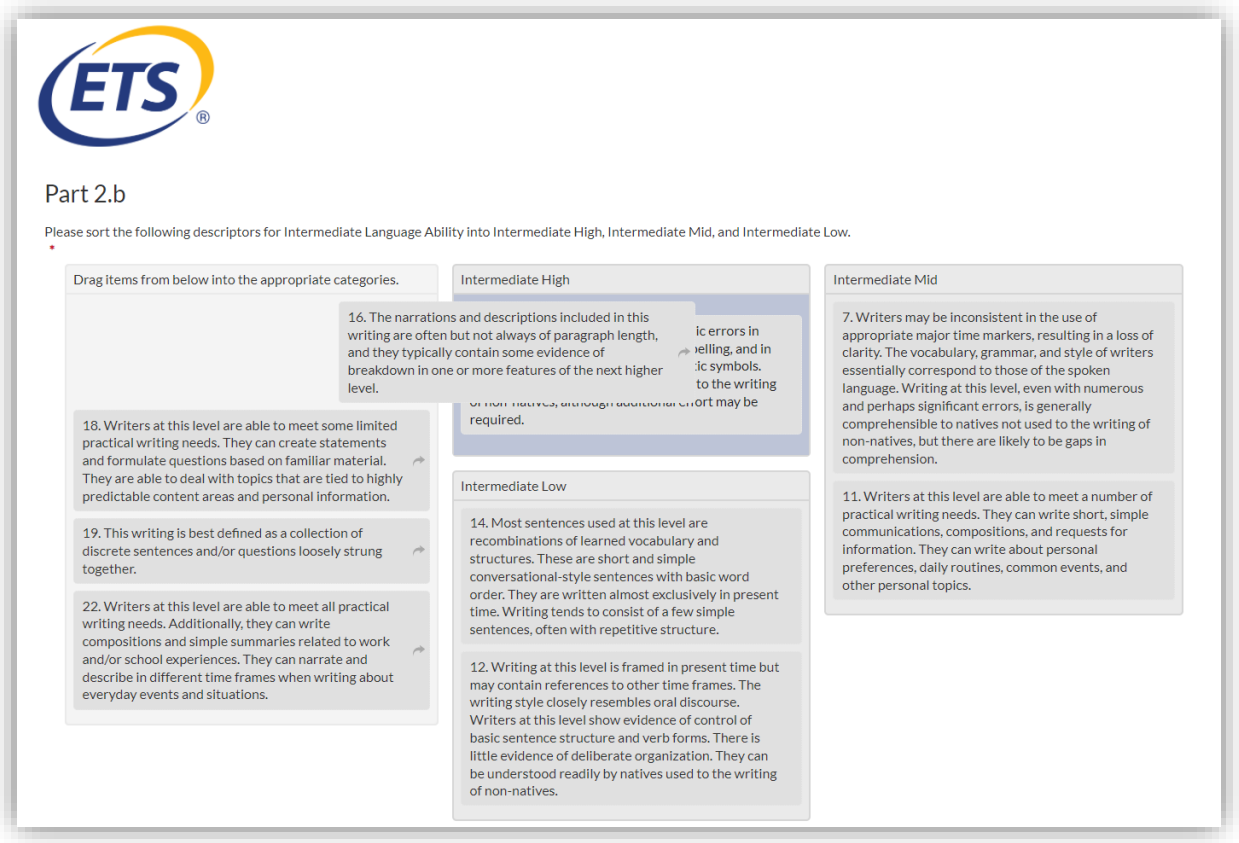
4. Writers at this level demonstrate the ability to incorporate a limited number of cohesive devices, and may resort to some redundancy and awkward repetition. They rely on patterns of oral discourse and the writing style of their first language. Their writing is understood by natives not accustomed to the writing of non-natives, although some additional effort may be required in the reading of the text.

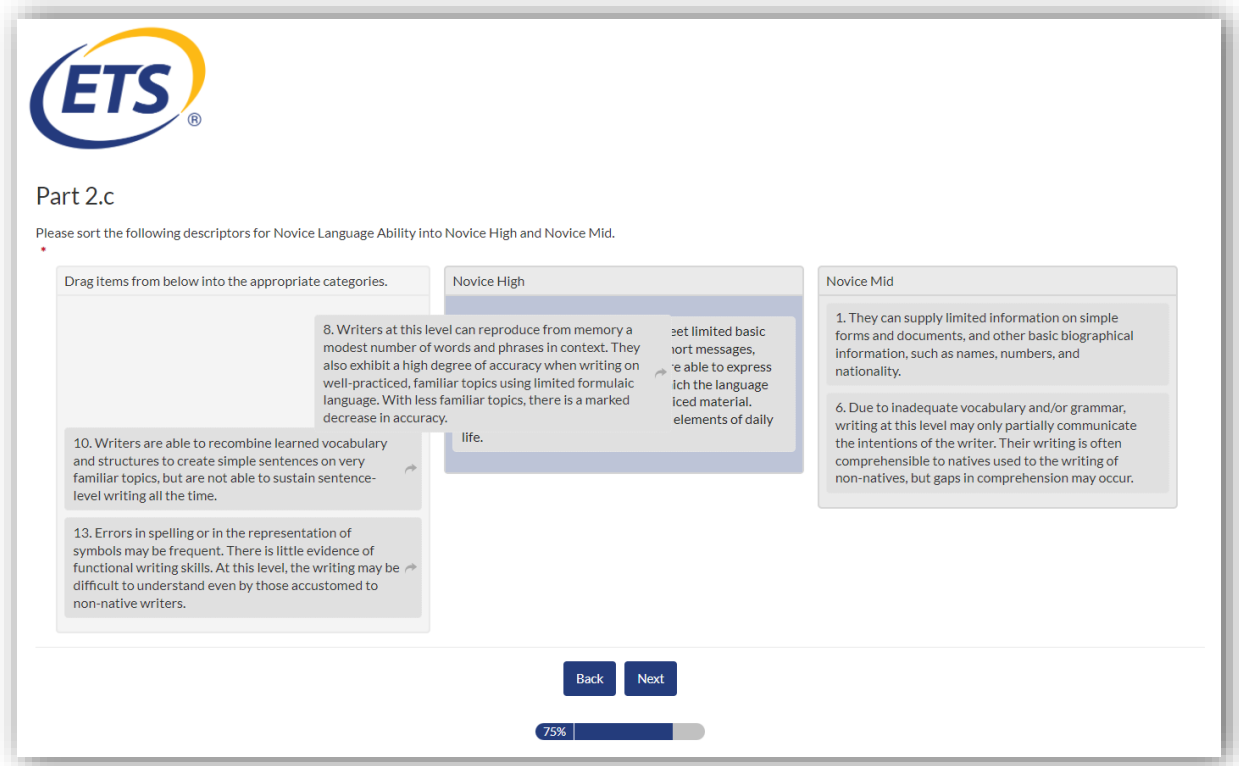
Advanced Mid

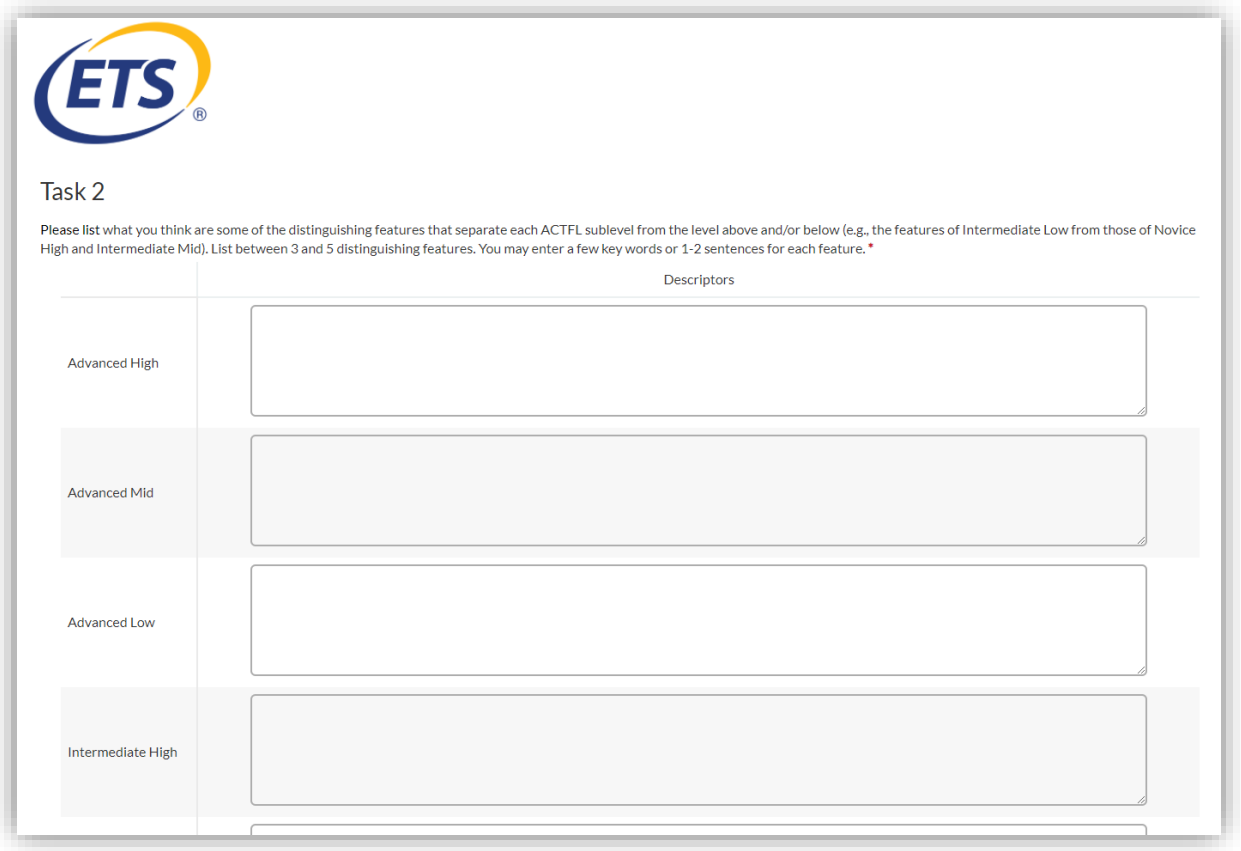
15. Writers at this level are able to combine and link sentences into texts of paragraph length and structure. Their writing may not be substantive.

20. Writers at this level are able to meet basic work and/or academic writing needs. They demonstrate the ability to narrate and describe in major time frames with some control of aspect.

9. Writers can handle informal and formal correspondence according to appropriate conventions. They can write summaries and reports of a factual nature. They can also write extensively about topics relating to particular interests and special areas of competence, although their writing tends to emphasize the concrete aspects of such topics. In addition, they are able to demonstrate the ability to handle writing tasks, such as developing arguments and constructing hypotheses, but are not able to do this all of the time; they cannot produce writing across a variety of topics treated abstractly or generally.


Part 2.b
Please sort the following descriptors for Intermediate Language Ability into Intermediate High, Intermediate Mid, and Intermediate Low.
*
Drag items from below into the appropriate categories.
18. Writers at this level are able to meet some limited practical writing needs. They can create statements and formulate questions based on familiar material. They are able to deal with topics that are tied to highly predictable content areas and personal information.
19. This writing is best defined as a collection of discrete sentences and/or questions loosely strung together.
22. Writers at this level are able to meet all practical writing needs. Additionally, they can write compositions and simple summaries related to work and/or school experiences. They can narrate and describe in different time frames when writing about everyday events and situations.
Intermediate High
16. The narrations and descriptions included in this writing are often but not always of paragraph length, and they typically contain some evidence of breakdown in one or more features of the next higher level.
Intermediate Low
14. Most sentences used at this level are recombinations of learned vocabulary and structures. These are short and simple conversational-style sentences with basic word order. They are written almost exclusively in present time. Writing tends to consist of a few simple sentences, often with repetitive structure.
12. Writing at this level is framed in present time but may contain references to other time frames. The writing style closely resembles oral discourse. Writers at this level show evidence of control of basic sentence structure and verb forms. There is little evidence of deliberate organization. They can be understood readily by natives used to the writing of non-natives.
Intermediate Mid
7. Writers may be inconsistent in the use of appropriate major time markers, resulting in a loss of clarity. The vocabulary, grammar, and style of writers essentially correspond to those of the spoken language. Writing at this level, even with numerous and perhaps significant errors, is generally comprehensible to natives not used to the writing of non-natives, but there are likely to be gaps in comprehension.
11. Writers at this level are able to meet a number of practical writing needs. They can write short, simple communications, compositions, and requests for information. They can write about personal preferences, daily routines, common events, and other personal topics.


Part 2.c
Please sort the following descriptors for Novice Language Ability into Novice High and Novice Mid.
*
Drag items from below into the appropriate categories.
10. Writers are able to recombine learned vocabulary and structures to create simple sentences on very familiar topics, but are not able to sustain sentence-level writing all the time.
13. Errors in spelling or in the representation of symbols may be frequent. There is little evidence of functional writing skills. At this level, the writing may be difficult to understand even by those accustomed to non-native writers.
Novice High
8. Writers at this level can reproduce from memory a modest number of words and phrases in context. They also exhibit a high degree of accuracy when writing on well-practiced, familiar topics using limited formulaic language. With less familiar topics, there is a marked decrease in accuracy.
Novice Mid
1. They can supply limited information on simple forms and documents, and other basic biographical information, such as names, numbers, and nationality.
6. Due to inadequate vocabulary and/or grammar, writing at this level may only partially communicate the intentions of the writer. Their writing is often comprehensible to natives used to the writing of non-natives, but gaps in comprehension may occur.
Back
Next
75%



Task 2

Please list what you think are some of the distinguishing features that separate each ACTFL sublevel from the level above and/or below (e.g., the features of Intermediate Low from those of Novice High and Intermediate Mid). List between 3 and 5 distinguishing features. You may enter a few key words or 1-2 sentences for each feature. *

	Descriptors
Advanced High	
Advanced Mid	
Advanced Low	
Intermediate High	

Appendix C. Meeting Agenda

Saturday, April 23, 2022, 11:00 am–7:00 pm (EDT)

- Welcome and overview
- Overview of the TOEIC Writing test
- Developing "Just Qualified Candidate" definitions for ACTFL levels

  Novice High to Advanced High
- Practice on standard setting method and training evaluation
- Round 1 judgments
- Round 1 discussion and Round 2 judgments
- Round 2 discussion and finalization of cut scores (Round 3)
- Wrap-up and meeting evaluation

Appendix D. Panel's Just Qualified Candidate (JQC) Descriptors

JQC Advanced High

Time Frame

Use major time frames with high degree of accuracy and control of aspect

Context/Content/Register

- Write about a variety of concrete topics with precision and detail
- Demonstrate accurate conventions for informal and formal correspondence
- Consistently able to narrate and describe in detail
- Attempted/emerging discussion of abstract concepts and basic arguments and hypotheses

Organization, Cohesive Devices

- Developing arguments in well-developed and unified paragraphs with strong evidence of cohesion
- Organization is mostly similar to the conventions of English language, with some characteristics of first language

Comprehensibility (Vocabulary/Grammar/Sentence Structure/Error)

- Good control of a range of grammatical structures and a fairly wide general vocabulary and ease of expression
- Consistently comprehensible to readers who are not accustomed to non-native writing

JQC Advanced Mid

Time Frame

Use major time frames with some degree of accuracy and control of aspect

Context/Content/Register

- Write about a variety of concrete topics with some precision and detail
- Demonstrate moderately accurate conventions for informal and formal correspondence
- Can write summaries on topics of general interest

Organization, Cohesive Devices

- Developing arguments in sequenced paragraphs with some evidence of cohesion
- Organization is mostly similar to the conventions of English writing, with some characteristics of first language

Comprehensibility (Vocabulary/Grammar/Sentence Structure/Error)

- Good control of the most frequently used grammatical structures and a range of general vocabulary
- Primarily comprehensible to readers who are not accustomed to non-native writing
- Thoughts are typically expressed and supported by some elaboration

JQC Advanced Low

Time Frame

Use major time frames with emerging degree of accuracy and control of aspect

Context/Content/Register

- Write about familiar/common topics with detail
- Demonstrate moderately accurate conventions for informal and formal correspondence
- Can write summaries only on familiar topics and able to write straightforward summaries on topics of general interest

Organization, Cohesive Devices

- Able to combine and link sentences into basic/minimal paragraphs

- Organization is somewhat similar to the conventions of English writing, with noticeable characteristics of first language
- Limited control of cohesive devices

Comprehensibility (Vocabulary/Grammar/Sentence Structure/Error)

- With good deal of effort, comprehensible to readers who are not accustomed to non-native writing
- Some thoughts are supported by elaboration
- Attempts to produce advanced grammatical and sentence structures with general vocabulary

JQC Intermediate High

Time Frame

- Can describe and narrate using correct consistent control of present, past, future time frames; occasional successful use of other time frames
- Lack control of aspect

Context/Content/Register

- Simple summaries of work or school experience
- Routine formal and informal communication tasks (e.g., emails)
- Vocabulary and grammar generally reflect oral speech

Organization, Cohesive Devices

- Sometimes writes paragraph length and shows some organization of thoughts
- Demonstrate emerging/inconsistent control of connected paragraphs and cohesion between phrases

Length

- Often but not always of paragraph length
- Sometimes writes paragraph length and shows some organization of thoughts

Comprehensibility (Vocabulary/Grammar/Sentence Structure/Error)

- Consistent control of simple sentence structure; emerging/inconsistent control of complex sentence structures
- Writing is mostly comprehensible to readers accustomed to non-native writing and generally comprehensible to readers who are not accustomed to non-native writing

JQC Intermediate Mid

Time Frame

- Mostly uses the present time, but demonstrates limited usage of major time frames

Context/Content/Register

- Demonstrates general control of simple, practical, or personal written communication about daily life or routines in texts that are short simple compositions or requests for information
- Limited or nonexistent control of formal/informal register appropriate to the writing task

Organization, Cohesive Devices

- Emerging but mostly unsuccessful attempts at coherence

Length

- Emerging but inconsistent evidence of multisentence structure or connections between sentences

Comprehensibility (Vocabulary/Grammar/Sentence Structure/Error)

- Text may require some limited effort for native speakers familiar with non-native writing

JQC Intermediate Low

Time Frame

- Almost exclusively uses the present time and does not demonstrate usage of other major time frames

Context/Content/Register

- Demonstrates limited control of simple, practical, or personal written communication about daily life or routines in texts that are similar to oral language or conversational style of statements and questions

Organization, Cohesive Devices

- No meaningful paragraph structure, organization, or cohesive devices

Length

- Some simple multisentence communication is demonstrated

Comprehensibility (Vocabulary/Grammar/Sentence Structure/Error)

- Text may require some consistent effort for native speakers familiar with non-native writing

JQC Novice High

Context/Content/Register

- Demonstrates basic control of very short and simple messages such as lists or notes

Language Control/Time Frame

- Heavily reliant on practiced/learned material, with some ability to recombine learned grammar and vocabulary in some new but basic ways

Organization, Cohesive Devices

- No discourse level organization above the sentence level is demonstrated. Connections are nonlinguistic based on lists or notes

Length

- Inconsistent attempts of some simple sentences

Comprehensibility (Vocabulary/Grammar/Sentence Structure/Error)

- Text requires some consistent effort to be comprehensible even for native speakers accustomed to non-native writing
- Vocabulary and grammar are too basic to convey complex ideas. Language use is formulaic and based on learned structures nearly exclusively

Appendix E. Themes and Quotes From the Meeting Evaluation Survey

Theme	Example quote
Meeting process and procedure	
Meeting facilitation	• They did a great job of keeping us on task while also facilitating the freedom to share our ideas. (ID#02) • I certainly benefited from both the written and oral instructions. (ID#13)
Development of JQCs	• The purpose of the JQCs became even more clear as we went along. I felt like I knew the levels well because of the two practice tasks. (ID#13) • …I think there was a potentially fatal flaw in the step from ACTFL descriptors to JQCs…what happened was that we ended up straying off into creating our own descriptors that are only half-connected to the ACTFL ones, and I don't know that we actually achieved what we were intended to in defining that lower bound of each level. (ID#14) • …it felt like many have been drawn to the notion of "average" candidate not JQC during the discussion… (ID#18)
Judgment task	• …it was actually easier for me to go through and make the judgments starting from Novice High and working up. Moving in two directions was a bit weird. (ID#12) • It would have been nice to look at least one sample and discuss together. We spent a lot of time early on coming up with JQC criteria, but I ended up using other documents more when making my judgements. (ID#03)
Group discussion	• More small group discussions. (ID#15) • …I do feel that we could have spent a bit more time discussing the descriptors (some criteria such as vocabulary and grammar could've been developed further) and then also a bit more time discussing and deciding on the cut scores; I sensed that it was a bit rushed, and maybe we could have been put into small groups for this activity, looking at and reviewing individual tests for a few minutes. (ID#06)
Personal reflection	• I really enjoyed this study and felt that it was incredibly informative. (ID#04) • This was an interesting process—it was quite challenging at first, to understand several constructs, concepts, and tasks. But the process felt easier and made more sense as we progressed through each stage. (ID#08) • Great experience with everyone concerned. I learned a lot also. (ID#10)
Logistical suggestions	• In the breakout room, it would be better to have a neutral "scribe" rather than a volunteer from among the participants… (ID#2) • Zoom instead of Teams? (ID#15) • It would be helpful if you break it into shorter sessions into two or more days. (ID#17)

Notes

[1] Scores on the TOEIC Writing test have also been mapped onto the Common European Framework of Reference for Languages (CEFR; see Tannenbaum & Wylie, 2008). The score-mapping results that this study will provide can thus be used to establish comparability with other preexisting performance standards and provide a basis for score users to make comprehensive score interpretations.