Quality Beyond Measu

Check for updates

TOEFL® Research Report
TOEFL–RR-89
ETS Research Report No. RR–19-44

Mapping the *TOEFL iBT*® Test Scores to China’s Standards of English Language Ability: Implications for Score Interpretation and Use

Spiros Papageorgiou

Sha Wu

Ching-Ni Hsieh

Richard J. Tannenbaum

Mengmeng Cheng

December 2019

The *TOEFL*® test is the world's most widely respected English language assessment, used for admissions purposes in more than 130 countries including Australia, Canada, New Zealand, the United Kingdom, and the United States. Since its initial launch in 1964, the TOEFL test has undergone several major revisions motivated by advances in theories of language ability and changes in English teaching practices. The most recent revision, the *TOEFL iBT*® test, contains a number of innovative design features, including integrated tasks that engage multiple skills to simulate language use in academic settings and test materials that reflect the reading, listening, speaking, and writing demands of real-world academic environments. In addition to the TOEFL iBT, the TOEFL Family of Assessments has expanded to provide high-quality English proficiency assessments for a variety of academic uses and contexts. The TOEFL Young Students Series (YSS) features the *TOEFL*® *Primary*™ and *TOEFL Junior*® tests, designed to help teachers and learners of English in school settings. The *TOEFL ITP*® Assessment Series offers colleges, universities, and others an affordable test for placement and progress monitoring within English programs.

Since the 1970s, the TOEFL tests have had a rigorous, productive, and far-ranging research program. ETS has made the establishment of a strong research base a consistent feature of the development and evolution of the TOEFL tests, because only through a rigorous program of research can a testing company demonstrate its forward-looking vision and substantiate claims about what test takers know or can do based on their test scores. In addition to the 20-30 TOEFL-related research projects conducted by ETS Research & Development staff each year, the TOEFL Committee of Examiners (COE), composed of distinguished language-learning and testing experts from the academic community, funds an annual program of research supporting the TOEFL family of assessments, including projects carried out by external researchers from all over the world.

To date, hundreds of studies on the TOEFL tests have been published in refereed academic journals and books. In addition, more than 300 peer-reviewed reports about TOEFL research have been published by ETS. These publications have appeared in several different series historically: TOEFL Monographs, TOEFL Technical Reports, TOEFL iBT Research Reports, and TOEFL Junior Research Reports. It is the purpose of the current TOEFL Research Report Series to serve as the primary venue for all ETS publications on research conducted in relation to all members of the TOEFL Family of Assessments.

To obtain more information about the TOEFL programs and services, use one of the following:

E-mail: toefl@ets.org **Web site:** www.ets.org/toefl

As part of its educational and social mission and in fulfilling the organization's non-profit Charter and Bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

RESEARCH REPORT

Mapping the *TOEFL iBT*® Test Scores to China's Standards of English Language Ability: Implications for Score Interpretation and Use

Spiros Papageorgiou,[1] Sha Wu,[2] Ching-Ni Hsieh,[1] Richard J. Tannenbaum,[1] & Mengmeng Cheng[2]

1 Educational Testing Service, Princeton, NJ

2 National Educational Examinations Authority, Beijing, China

The past decade has seen an emerging interest in mapping (aligning or linking) test scores to language proficiency levels of external performance scales or frameworks, such as the Common European Framework of Reference (CEFR), as well as locally developed frameworks, such as China's Standards of English Language Ability (CSE). Such alignment is ultimately a claim about the interpretation of test scores in relation to external levels of language proficiency. To support such a claim, established procedures should be carefully implemented and multiple sources of evidence should be collected. In this research report, we demonstrate the application of a series of steps in building an argument for aligning the scores of the *TOEFL iBT*® test, an international, large-scale language proficiency test of English as a foreign language (EFL), to the levels of the CSE. The alignment process comprised the following steps: (a) establishing construct congruence between the TOEFL iBT test and the CSE; (b) establishing recommended minimum test scores (cut scores), set by local experts, to classify language learners into the local proficiency levels; (c) collection of scores by test takers ($N = 1{,}326$) and evaluations of the test takers' proficiency levels by their teachers, based on the local framework; and (d) consideration of the results of other alignment studies in the local context as well as the link between the CEFR and the CSE levels. We conclude with a discussion of the contextual issues that should be considered when interpreting test scores in relation to external proficiency levels. These contextual issues are important considerations because they have the potential to impact score-based decisions on individuals and institutions. We also discuss the implications for similar alignment research.

Keywords *TOEFL iBT*® test; China's Standards of English Language Ability; standard setting; cut scores; alignment; score interpretation

doi:10.1002/ets2.12281

Test-based decisions can have important consequences for students, teachers, and institutions. Therefore, it is critical that test results be communicated in ways as transparent and meaningful as possible (Hambleton & Zenisky, 2013; Tannenbaum, 2019; Zenisky & Hambleton, 2012). However, the most important product of testing, that is numerical scores, typically do not convey direct information about what test takers actually know and are able to do. To address this issue, language testers have attempted to enhance the meaning of test scores by describing examinee performance in narrative terms, such as performance levels or performance descriptors (Alderson, 1991; Ryan, 2006).

Mapping (aligning or linking) test scores to external proficiency levels and descriptors, such as those in the Common European Framework of Reference (CEFR) of the Council of Europe (2001), is a common approach to facilitate the interpretation of test scores (Tannenbaum & Cho, 2014). Another approach is the development of internal, test specific performance level descriptors by "anchoring" exemplar test items to characterize particular score points within a level; hence the term *scale anchoring* (Beaton & Allen, 1992; Haberman, Sinharay, & Lee, 2011) for describing this procedure.

As Powers, Schedl, and Papageorgiou (2017) pointed out, when external levels and frameworks are relevant to the constructs being measured by a particular test and widely known in the educational contexts where the test is administered, then alignment can often make the interpretation of test scores more meaningful to the community familiar with these levels and descriptors (see also Kane, 2012). Ultimately, alignment is a claim about the interpretation of test scores in relation to external levels of language proficiency. To support such a claim, established procedures should be carefully implemented and multiple sources of evidence should be collected. The widespread use of the CEFR in educational

Corresponding author: S. Papageorgiou, E-mail: spapageorgiou@ets.org

systems around the world led its developer to publish a manual (Council of Europe, 2009) to guide test developers in linking test scores to the CEFR levels.

Test scores are often used within an educational or social context and have consequences for both individuals and institutions. Therefore, contextual issues should be carefully considered when interpreting test scores in relation to external proficiency levels. Such contextual issues are particularly important in the Chinese educational system, which in recent years witnessed a major development in the conceptualization of Chinese learners' language proficiency with the introduction of China's Standards of English Language Ability (CSE) (National Education Examinations Authority [NEEA], 2018). Released by the Ministry of Education and National Language Commission of China in 2018, the CSE was designed within and for China's specific context of use. The CSE includes comprehensive English proficiency scales covering the full range of learners of English as a foreign language (EFL) in China, built upon a large-scale empirical study.

NEEA and Educational Testing Service (ETS) launched a joint research project in November 2017 to explore the mapping of the *TOEFL iBT*® test scores to the language proficiency levels of the CSE. NEEA and ETS staff, as well as the CSE project team members from various academic institutions in China (see names in Appendix A), were involved in the research project. In this research report, we present the steps we followed in building an argument for the alignment of the test scores to the CSE levels. First, the content of the test was examined in relation to the description of language ability in the CSE to establish adequate construct congruence (content alignment). Such a step is critical because lack of construct congruence is a threat to an alignment claim. The second step involved 16 experts in the Chinese educational context who worked with researchers from the ETS headquarters in Princeton, NJ, to recommended minimum test scores (cut scores) to classify a test taker at a given CSE level. The cut score recommendation process, based on standard setting methodology, is central to an alignment claim because it establishes the "decision rule" as to how to classify language learners into the external levels based on the test scores. The third step involved examinee-centered data to evaluate the reasonableness of the recommended cut scores. The scores of a representative test taker sample ($N = 1{,}326$) were collected along with evaluations of the test takers' proficiency levels by their teachers, based on the CSE. To make the final score mapping recommendation, project steering group and working group members considered all collected data, as well as additional sources of external data, which included

- concordance of the TOEFL iBT test scores with the scores of another language test, which had already been mapped to the CSE levels; and
- comparison of the link between the CSE levels and CEFR levels, as well as the alignment of TOEFL iBT scores to the CEFR levels, as established in separate studies.

We conclude this research report with a discussion of the contextual issues that should be considered when interpreting test scores in relation to local proficiency levels, because of the potential impact of score-based decisions on individuals and institutions. We also discuss the implications for similar alignment research.

Context of the Research Study

The Purpose of the CSE

With an increasing awareness of the need to scale English learners' proficiency, language educators, teaching practitioners, and policymakers in China reached a consensus that a unified proficiency scale is urgently needed to describe learners' performance and streamline their competence across different educational stages and different regions in the Chinese EFL context. In 2014, the State Council of P. R. China issued a document titled "The Implementation Opinions of the State Council on Deepening the Reform of the Examination and Enrollment System." One pressing task, as highlighted in the document, was to develop a foreign language assessment framework to improve the quality of language tests; enhance the communication between teaching, learning, and assessment; and, thus, raise the overall effectiveness and efficiency of foreign language education in China. Against this, NEEA, endorsed by the Ministry of Education, P.R. China, initiated a nationwide project to develop an English language proficiency scale, known as the CSE, which set out to (a) define and describe the English proficiency of the English learners in China; (b) provide references and guidelines for English learning, teaching, and assessment; and (c) enrich the existing body of language proficiency scales for alignments on a global basis (Liu, 2015).

Theoretical Underpinnings of the CSE

The CSE adopted a use-oriented approach to the description of language ability based on a composite of several language ability models. These models include the communicative language ability (CLA) model (Bachman, 1990; Bachman & Palmer, 1996, 2010), Bloom's Revised Taxonomy of educational objectives (Anderson & Krathwohl, 2001), and the functional linguistic model (Jackson & Stockwell, 2011).

Bachman (1990) proposed the CLA model, where language ability is perceivably constructed as "consisting of both knowledge, or competence, and the capacity for implementing, or executing that competence in appropriate, contextualized communicative language use." The CLA model not only inherits organizational competence from its traditional sense, but also embeds strategic competence and regards it as not just serving a compensatory function, which, to a certain extent, alludes to Canale's (1983) refined model. More importantly, it recognizes the roles of cognitive strategies and pragmatic competence, together with their impact on the realization of communicative competence. On the whole, this model is theoretically sound and empirically validated, and it is merited as the state-of-the-art representation (Alderson & Banerjee, 2002). Referring to the CLA model (Bachman, 1990) and its revisions (Bachman & Palmer, 1996, 2010), the CSE defines language ability as the ability to comprehend and express information that learners exhibit when they apply their language knowledge and world knowledge and the strategies to perform language use tasks in a variety of contexts (Liu & Han, 2018).

The CSE adopted a use-oriented approach to the description of language ability. The term *use-oriented approach* is applied in response to the emergent demand of cultivating the learners' ability to use the language in the real world rather than learning the language as a static body of knowledge. To this end, the CSE treats language ability as a type of dynamic cognitive activity instead of an abstract and static system of rules. Adapting Bloom's Revised Taxonomy of educational objectives, the CSE embeds the cognitive levels in the descriptions of language performance. According to functional linguistics, texts are the carrier of language activities. Texts can be static like a reading passage or dynamic as a social process (Knapp & Watkins, 2005). So text types or functions (i.e., description, narration, exposition, argumentation, interaction, and instruction) were introduced in the CSE to categorize the language activities, aiming to make the CSE relevant to a wider context of uses and, more importantly, to encourage language learners and users to expand the scope of language uses.

Components of the CSE

Based on the theoretical framework, CSE developers formulated a descriptive scheme, as illustrated in Figure 1. Language ability, the core notion, is further divided into language comprehension (listening and reading), language expression (oral and written), and mediation (translation and interpreting). In congruence with the different functions that communication mainly serves, different subabilities deal with a plethora of texts, including narrative, descriptive, expository, argumentative, instructional, and interactional texts.

The two-headed arrow between translation/interpreting and language activities means English learners or users need to enable two channels: the source and the target languages, both of which are manifested by the texts of various communicative functions. In order to streamline the framework across different subability scales, the CSE developers designed the four-layer framework. The use of the word *four-layer* means that the description of language ability is structured in a hierarchical system. Language ability stands on the top layer, beneath which there are language comprehension and language expression. Mediation is also placed at this layer, though it is arguable whether mediation, translation, and interpretation in the CSE can be regarded as a kind of language ability. The third layer includes six subabilities (i.e., listening comprehension, reading comprehension, oral expression [speaking], written expression [writing], translation, and interpreting). All six subabilities are described respectively based on six functions or text types, which construct the fourth layer. Global scales for overall language ability and each subability are provided, as well as the subscales for all the functions specific to each subability mentioned above. Pragmatic ability is described in terms of pragmatic comprehension and pragmatic expression. Pragmatic comprehension concerns interpreting the speaker's intentions or the writer's intentions, while pragmatic expression involves expressing intentions in either speaking or writing.

Communication success depends upon the language knowledge learners and users resort to, as well as the strategies they employ in an activity. Consequently, language knowledge and strategies are another major component in the CSE descriptive framework. To specify what constitutes language knowledge, the CSE developers mainly referred to the CLA

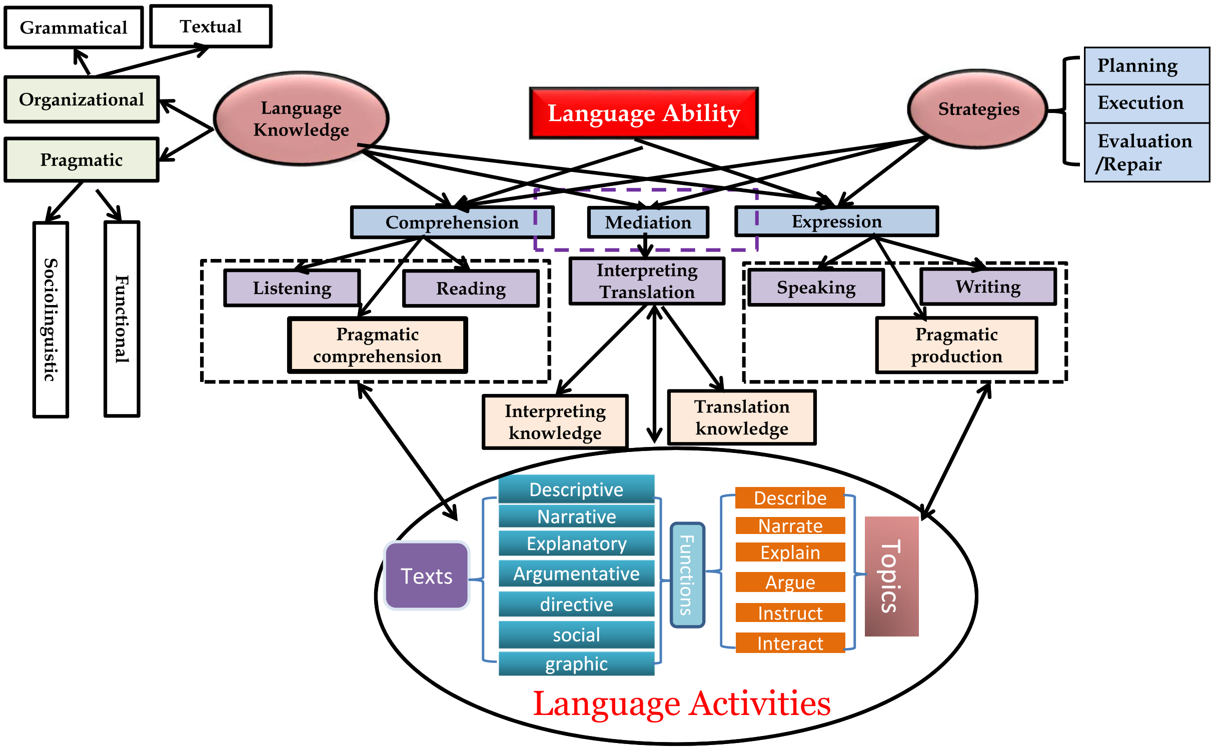


Figure 1 Components of China's Standards of English Language Ability.

model (Bachman & Palmer, 2010) and divided language knowledge into organizational knowledge and pragmatic knowledge. The former can be further broken down into grammatical knowledge and textual knowledge; the latter includes functional knowledge and sociolinguistic knowledge.

Apart from language knowledge, strategies can be divided into planning, execution, and evaluation/repair. Strategies are described globally as well as separately with six language subabilities. It is noteworthy that the strategies related to different language subabilities vary in naming. For example, the specific name for evaluation/repair in writing is revision, whereas assessment and compensation are used in the case of speaking.

Development and Validation of the CSE

The CSE project was launched in June 2014 and was completed at the end of 2017, around the time that representatives from ETS and the CSE team held their first face-to-face meeting in Shanghai. Figure 2 outlines the CSE development procedure. As is illustrated, the CSE development can be briefly divided into three phases. The first phase primarily dealt with collecting descriptors, which derived from not only a wealth of literature but also a database of descriptors generated by students and teachers of different educational levels. In the second phase, the CSE developers, based on expert and teacher judgments, conducted trial validation on a working-group basis. During this process, the developers removed duplicate descriptors, blended similar descriptors, and categorized descriptors into the framework of the CSE, as expanded above. The last phase was composed of two field studies for the finalization of scaling. In the first field study, all the polished descriptors were randomly spread into different sets of questionnaires, which were then administered to language education experts and classroom teachers as well as learners/users. They reported the extent to which their students (if the participants were teachers) or they themselves (if the participants were learners/users) could perform in relation to each descriptor provided. Based on the results, statistical analyses were conducted to determine the cut-off points of each proficiency level. The second field study, which was smaller in scale, aimed to elicit responses from teachers of various educational stages to the same set of descriptors, so that vertical scaling could be done for the calibration of the cut-off points.

Based on the composite analysis of the research results mentioned above, CSE descriptors were scaled into nine levels (CSE 1 through CSE 9) and were arranged in an ascending order from lower proficiency levels to higher ones. For an

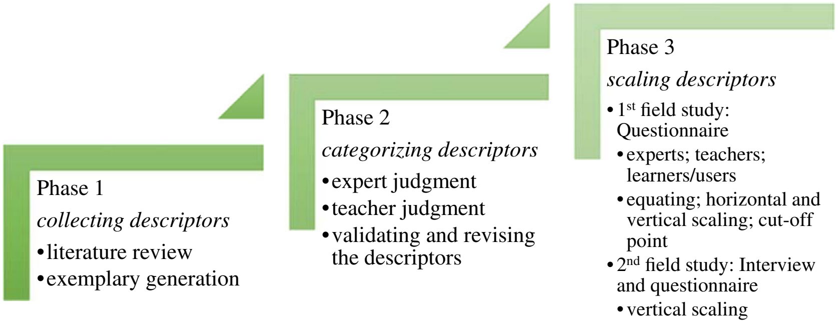

Phase 1
collecting descriptors
• literature review
• exemplary generation
Phase 2
categorizing descriptors
• expert judgment
• teacher judgment
• validating and revising the descriptors
Phase 3
scaling descriptors
• 1st field study: Questionnaire
• experts; teachers; learners/users
• equating; horizontal and vertical scaling; cut-off point
• 2nd field study: Interview and questionnaire
• vertical scaling

Figure 2 The development procedure of China's Standards of English Language Ability.

Table 1 Three-Element Model of Descriptors

Descriptor	Performance	Criteria	Conditions
Can, after preparation, present views coherently on hot social issues. (CSE, Oral Expression-Oral Exposition, Level 5)	Present views on hot social issues.	Coherently	After preparation
Can briefly describe recent experiences or mood. (CSE, Written Expression-Written Description, Level 3)	Describe experiences or mood.	Briefly	Recent

Note. CSE = China's Standards of English Language Ability.

easier reference, these levels are further grouped into three stages: elementary (CSE 1–3), intermediate (CSE 4–6), and advanced (CSE 7–9). However, it is worth noting that CSE sets out to describe progression of general English language proficiency regardless of learners' age and educational background.

Three guidelines have been consistently followed through the whole process of descriptor screening and revision. First, each descriptor takes the form of "can-do statement." In other words, what is described should point to learners' or users' accomplishments rather than their weaknesses. Caution should also be taken in using hedging and degree adverbs, such as *comparatively* and *in general* for scaling purposes. That means each descriptor can stand alone. Long and complex-structure descriptors should also be revised, for CSE users may be at a loss as to what is focused on in an individual descriptor if it is too long or complicated. Ambiguity, vagueness, atypical language activities, and linguistic jargon ought to be avoided wherever possible. Second, the intended construct of an individual descriptor should be unique. If more than one ability is included in a descriptor, this could give rise to misunderstandings among CSE users. Third, each descriptor follows a three-element model (Pearson Standards and Quality Office, 2014; see Table 1):

- performance: the language operation itself (e.g., present views on hot social issues)
- criteria: the intrinsic quality of the performance, typically in terms of the range of language used (e.g., coherently)
- conditions: any extrinsic constraints or conditions defining the performance (e.g., after preparation)

As such, the elements of performance and criteria, which stipulate the *doing* with the English language and how well the *doing* is, are compulsory for all descriptors. In comparison, condition is optional, given its role of adding or removing constraints.

Description of the TOEFL iBT Test

The first *TOEFL*® test was launched in 1964. The TOEFL iBT test was introduced in 2005 to better reflect the language demands of real-life academic tasks than did previous versions of the test. The purpose of the test is to evaluate the English proficiency of people whose first language is not English, and its scores are primarily used as a measure of the ability of international students to use English in an academic environment. To quote the original TOEFL working paper, the purpose of the test is "to measure the communicative language ability of people whose first language is not English … in situations and tasks reflective of university life" (Jamieson, Jones, Kirsch, Mosenthal, & Taylor, 2000, p. 10). The test content, described in this section, was based on extensive research of the tasks students typically perform in a university

Table 2 Overview of the Content of the TOEFL iBT Test

Test section	Test content (through July 2019)	Test content (from August 2019)
Reading	3–4 passages 12–14 questions per passage Duration: 60–80 min	3–4 passages 10 questions per passage Duration: 54–72 min Main content revision: fewer questions per passage
Listening	4–6 lectures 6 questions per lecture 2–3 conversations 5 questions per conversation Duration: 60–90 min	3–4 lectures 6 questions per lecture 2–3 conversations 5 questions per conversation Duration: 41–57 min Main content revision: fewer lectures
Speaking	6 tasks 2 independent 4 integrated Duration: 20 min	4 tasks 1 independent 3 integrated Duration: 17 min Main content revision: fewer tasks
Writing	2 tasks 1 integrated 1 independent Duration: 50 min	2 tasks 1 integrated 1 independent Duration: 50 min No content changes made in this test section

context (see Chapelle, Enright, & Jamieson, 2008). Since the launch of the test, an ongoing research program has provided funding to ETS and external researchers to support validation research. Arguably, the TOEFL research program is the most extensive program of its kind, resulting in over 200 research publications related to the TOEFL iBT test. These publications provide evidence supporting the validity of the test scores and the claims that are made based on these scores.[1]

The TOEFL iBT test is administered worldwide via computer from a secure, Internet-based testing network. Some tasks on the test require the use of two or more language skills. Test takers wear noise-reducing headphones and speak into a microphone to record their responses to the speaking tasks and type their responses to the writing tasks. The spoken and written responses are digitally recorded and sent to the ETS online scoring network.

As Table 2 illustrates, each test form includes four sections: reading, listening, speaking, and writing. Each section is reported on a scale of 0–30, resulting in a total score of 120. As of August 2019, the test time was shortened by 30 minutes to 3 hours. The data collection for this project was completed before August 2019; therefore, the longer version of the test was used. However, because the shorter version uses the same section and total score scales, with no changes to the overall test format or item types, there is no effect on the overall design and results of this research project.

The reading section measures test takers' ability to understand university-level academic texts. TOEFL iBT test takers read three or four passages of approximately 700 words each and answer 10 questions about each passage (13–14 questions per passage prior to August 2019). The passages represent a variety of academic areas and contain all the information needed to answer the questions; they require no special background knowledge. The questions are intended to assess the test takers' ability to comprehend facts, infer information from the passage, understand vocabulary in context, and understand the author's purpose. Other types of questions assess the test takers' ability to recognize relationships among facts and ideas in different parts of a passage.

The listening section measures test takers' ability to understand spoken English in an academic setting. Test takers listen to three to four lectures (four to six lectures prior to August 2019) representing different academic areas, each about 5 min long, and listen to two or three conversations representing typical campus interactions with faculty, staff, and fellow students, each about 3 min long. Each listening passage is associated with a set of questions intended to assess test takers' ability to understand main ideas or important details, recognize a speaker's attitude or function, understand the organization of the information presented, understand relationships between the ideas presented, and make inferences or connections among pieces of information.

The speaking section measures test takers' ability to use spoken English effectively in educational environments, both inside and outside the classroom. There are four tasks in the speaking section: one independent task and three integrated tasks (there were six tasks, two independent and four integrated, prior to August 2019). The independent task is designed to allow test takers to draw on their own personal experience and opinions. The integrated tasks require the test taker to use information presented in a short spoken text or both a short spoken text and a related short written text. These tasks assess integrated skills; that is, they require test takers to respond orally about things they listen to and read. The three types of integrated tasks are as follows:

- Read/Listen/Speak (campus situation). Test takers read a short passage communicating a typical campus situation or policy and then listen to a conversation in which a speaker expresses an opinion about the situation or policy. Test takers are then asked to give an oral summary of the speaker's opinion. A full response will require the test taker to combine and convey key information from both the reading and the listening tasks.
- Read/Listen/Speak (academic course topic). Test takers read a passage that broadly defines a term, process, or idea from an academic subject. They then listen to a lecture that provides specific examples to illustrate the term, process, or idea expressed in the reading passage. Finally, they are asked to explain how the illustration presented in the lecture supports the broader concept defined in the reading. A full response will require test takers to combine and convey key information from both the reading and the listening input materials.
- Listen/Speak (academic course topic). Test takers listen to an excerpt from a lecture that explains a term or concept (often by explaining two aspects or perspectives) and gives concrete examples to illustrate it. Test takers must then demonstrate understanding of the concept by providing a brief oral summary of the explanation and the related examples.

The writing section measures test takers' ability to write in an academic environment and includes two tasks: one independent and one integrated. The independent task requires test takers to draw on their own knowledge and experience to write a short essay that states, explains, and supports their opinion on a specific issue. The integrated task requires test takers to first read a passage on an academic topic. They then listen to part of a lecture that evaluates and criticizes the information and arguments presented in the reading. Finally, test takers must write a summary, in connected English prose, of the important points in the lecture, explaining how these points relate to those in the reading passage.

For both the speaking and the writing sections, test developers carefully design integrated tasks to ensure that a successful response will consider information from both the listening and reading materials. Additional resources about the TOEFL iBT test, including printable materials and a video library, can be found at https://www.ets.org/toefl/institutions/resources. Sample items of the TOEFL iBT item types are available at https://www.ets.org/toefl/ibt/prepare/quick_prep/.

Establishing Construct Congruence

External levels and descriptors cannot be specific to any given test and are likely to suffer from what has been called "descriptional inadequacy" (Fulcher, Davidson, & Kemp, 2011, p. 8); consequently, external level descriptors may not provide information that is directly relevant to test performance. For example, the description of what learners are expected to do at different language proficiency levels of the CEFR is intentionally underspecified to allow for wide application (Milanovic & Weir, 2010). Given this limitation of external levels and descriptors, evidence of "construct congruence" (Tannenbaum & Cho, 2014) is needed first to establish that a test measures language skills in a manner consistent with the way the external levels describe language proficiency. For example, the manual (Council of Europe, 2009) requires the developer of an exam to describe various aspects of the test content for each language skill in relation to the CEFR levels and descriptors (i.e., communicative language activities, tasks, communication themes, and text types) and present a claim of how test content covers aspects of language ability described in the CEFR.

Given that external levels and descriptors cannot be specific to any given test, nor can they function as the blueprint for test design, the TOEFL iBT test is not necessarily a point-by-point reflection of the English language skills and expectations presented in the CSE. This lack of point-by-point reflection is not a limitation of the test, but it does mean that evidence is needed regarding where and for which levels of the CSE the content of the test is considered adequately aligned before engaging in a standard-setting process to conduct score mapping (Council of Europe, 2009; Tannenbaum & Cho, 2014). Therefore, before convening the standard-setting meeting, NEEA reviewed sample test forms and other materials related to the design and intended difficulty of the test that are available on the TOEFL iBT website (https://www.ets.org/toefl/). Based on this review, NEEA and the CSE project team members identified 32 CSE scales (eight per test section) that were the most relevant to the TOEFL iBT test content. These scales, listed in Table 3, were also used for the definition of

Table 3 CSE Scales Selected for the Definition of Borderline Students

Listening	Reading	Speaking	Writing
Overall listening comprehension	Overall reading comprehension	Overall oral expression	Overall written expression
Self-assessment scale for listening comprehension	Self-assessment scale for reading comprehension	Self-assessment scale for oral expression	Self-assessment scale for written expression
Understanding oral description	Understanding written description	Oral description	Written description
Understanding oral narration	Understanding written narration	Oral narration	Written narration
Understanding oral exposition	Understanding written exposition	Oral exposition	Written exposition
Understanding oral instruction	Understanding written instruction	Oral instruction	Written instruction
Understanding oral argumentation	Understanding written argumentation	Oral argumentation	Written argumentation
Understanding oral interaction	Understanding written interaction	Oral interaction	Written interaction

Note. CSE = China's Standards of English Language Ability.

Table 4 Number of CSE Descriptors Aligned With the Content of Each Task in the TOEFL iBT Test Form

TOEFL iBT task	CSE 4 descriptors	CSE 5 descriptors	CSE 6 descriptors	CSE 7 descriptors	CSE 8 descriptors	Total
Reading task 1	1	3	6	5	3	18
Reading task 2	1	3	6	2	3	15
Reading task 3	1	3	5	2	3	14
Listening task 1	3	4	2	3	6	18
Listening task 2	3	5	4	2	3	17
Listening task 3	1	4	4	4	4	17
Listening task 4	1	2	2	4	4	13
Listening task 5	2	4	4	2	3	15
Listening task 6	1	3	4	5	2	15
Speaking task 1	4	9	7	1	2	23
Speaking task 2	1	9	6	2	2	20
Speaking task 3	0	0	1	1	3	5
Speaking task 4	1	2	2	5	9	19
Speaking task 5	3	8	6	2	2	21
Speaking task 6	0	2	0	5	10	17
Writing task 1	2	1	1	4	4	12
Writing task 2	5	4	3	2	2	16

Note. CSE = China's Standards of English Language Ability.

borderline students during the standard setting workshop and will be described later in this report. In addition, NEEA and the CSE team suggested that the focus of the score mapping study should be on Levels 4–8.

To check NEEA's recommendations, and to further investigate construct congruence between the CSE levels and the TOEFL iBT test, two ETS researchers reviewed the test form to be used in the standard setting meeting and identified the CSE descriptors from Levels 4 to 8 that are most closely aligned with the test content. The two researchers examined the test form, as well as the speaking and writing scoring rubrics, and reached consensus on the selected descriptors. Table 4 presents the number of descriptors for each task in the TOEFL iBT test form used in the standard setting meeting. It should be noted that a descriptor could have been selected for more than one task. For this reason, Table 5 presents the number of unique descriptors selected for each of the four test sections. A total of 117 descriptors were selected for the entire TOEFL iBT test form, suggesting adequate construct congruence.

There are some caveats as to how Tables 4 and 5 should be interpreted. As noted elsewhere in this report, the description of what learners are expected to do at different levels, such as the CSE ones, is intentionally underspecified to allow for wide applications of these levels. Thus, the descriptor frequencies by test task or test section should not be interpreted as an indication of exact match between the CSE levels, which are generic, and the test content which is based on a detailed test blueprint. Instead, the frequencies should be interpreted only as a justification of the decision to focus on the specific CSE levels during the subsequent standard setting workshop with the expert panel. Moreover, because only one test form was used, the descriptor frequencies presented in this section might not apply to another TOEFL iBT test form (although similarities would be expected, given that all TOEFL iBT test forms are based on the same blueprint). Finally, the descriptor

Table 5 Number of Unique CSE Descriptors Aligned With the Content of the TOEFL iBT Test Form

TOEFL iBT test section	CSE 4 descriptors	CSE 5 descriptors	CSE 6 descriptors	CSE 7 descriptors	CSE 8 descriptors	Total
Reading	1	3	7	6	3	20
Listening	6	6	5	6	6	29
Speaking	7	11	11	6	10	45
Writing	6	4	4	3	6	23
Total	20	24	27	21	25	117

Note. CSE = China's Standards of English Language Ability.

frequencies should not be used as a tool to compare the content alignment of the TOEFL iBT test and another language test to the CSE levels.

Methodology for the Standard Setting Study

General Procedures for Setting Cut Scores

After construct congruence has been established, a minimum score (cut score) on the test needs to be identified for each proficiency level. A cut score indicates the point on the test score scale that separates examinees who have demonstrated a specific level of performance from those who have not. Cut scores are established with a well-researched process called *standard setting* (Cizek & Bunch, 2007). During standard setting, a panel of experts is typically required, under the guidance of one or more meeting facilitators, to make judgments about the difficulty of test questions (items or tasks). The outcome of the standard setting meeting is a set of cut score recommendations to the examination provider. Statistical information about the test (e.g., item difficulty estimates and distribution of test scores) is also used to help panelists with their judgment task. A fairly common practice in standard setting meetings is to conduct more than one round of judgments. Between rounds, the panel discusses individual judgments, receives statistical information about items and scores, and repeats the judgments. Even though the panel will recommend a cut score, the decision regarding whether to accept or adjust (lower or raise) this score rests with the examination provider or score user. Procedures for validating the recommended cut scores are also presented in the manual (Council of Europe, 2009, pp. 89–118); we implemented several of these procedures, as we discuss later in this report.

Selection of Empirical Data From the TOEFL iBT Test

In preparation for specific steps during the standard setting meeting described in subsequent sections, members of the project team collaborated with assessment developers and psychometricians at ETS to collect item-level response data and score distribution information on one TOEFL iBT test form. The data were collected from 2,185 test takers in China who took the same TOEFL iBT test form in 2017. The geographical distribution of the test takers and their mean test performance were representative of the performance of all test takers who took the test in China in 2017. For example, the mean total TOEFL iBT score of the sample (78.9) was similar to the mean total score of all test takers who took the test in China in 2017 (79.4). The geographical distribution of the test takers was also representative of the 2017 overall test taking population in China, as the data were collected from 89 test centers in 39 cities. Of the 2,185 test takers, about half (1,084 test takers) took the test in high-volume cities, specifically Beijing, Shanghai, Guangzhou, Hangzhou, and Nanjing.

Selection of Panelists

Sixteen panelists from China, 12 female and 4 male, served on the standard setting panel. The panelists represented a variety of institutions (see Appendix B). All panelists completed a background questionnaire prior to the standard setting meeting. At the time of the study, the panelists indicated that they were teaching the following groups:

- senior high school students (two panelists)
- non-English major undergraduate students (two panelists)
- English major undergraduate students (three panelists)

- non-English major graduate students (one panelist)
- English major graduate students (two panelists)
- high school teachers (one panelist)
- at least two or more groups of university students listed above (five panelists)

Except for one, all panelists had more than 5 years of experience teaching English, and the majority had more than 10 years (13 panelists). All panelists had experience in one or more of the following tasks:

- writing test items for examinations at municipal level or above (selected by 10 panelists)
- writing or compiling textbooks (selected by 10 panelists)
- teacher training (selected by eight panelists)
- developing national teaching guideline or curriculum (selected by one panelist)

In terms of familiarity with the CSE, 11 panelists indicated they were very familiar with it, three indicated that they were somewhat familiar, and two indicated they were not familiar. It should be noted that 13 panelists had been involved previously in work related to the CSE, with seven panelists indicating that they were members of the development team. Although most panelists were overall very knowledgeable about the CSE, the project team decided to provide all of them with preparation materials to study prior to the standard setting meeting (discussed in the Panelist Preparation Prior to the Standard Setting Workshop subsection).

In terms of familiarity with the test, only five panelists indicated that they were very familiar with the TOEFL iBT test content. Ten panelists indicated that they were somewhat familiar, whereas one panelist noted lack of familiarity with the test. As we discuss later in this report, part of the panelists' preparatory activities prior to engaging in the standard setting process was to take the TOEFL iBT test under operational conditions. Taking the test is critical to the overall standard setting process, as it enables all panelists to become familiar with the test content.

A final point that should be noted here is that 9 of the 16 panelists had participated in a similar standard setting meeting to map the band levels of the IELTS test to the CSE, in which similar standard setting methodology was employed. However, not all panelists were familiar with the standard setting methods used in this study; therefore, training activities were organized for all panelists (described in "Standard Setting Method for Selected-Response Items" and the "Standard Setting Method for Constructed-Response Items" sections).

Panelist Preparation Prior to the Standard Setting Workshop

Prior to the standard setting study, a preparation guide was created and sent to the panelists. The guide included information about the CSE and the TOEFL iBT test, as well as preparatory activities related to the 32 scales identified as part of the construct congruence study, discussed earlier in this report (all scales were included in the guide). All panelists were asked to complete a familiarization activity for the selected scales of each of the four language skills (see Appendix C for the reading familiarization activity). The purpose of the activity was to ensure that the panelists had a good understanding of the features that distinguished each of the five CSE levels (i.e., Levels 4–8) selected for the study. The panelists also brought their guide with the completed familiarization tasks to the standard setting meeting.

On the first day of the standard setting meeting (see Appendix D for the full schedule) all panelists signed a nondisclosure/confidentiality agreement and took all four sections of the test at an authorized TOEFL iBT test center that was managed by NEEA. The test was delivered on the computer, in exactly the same way it is presented to actual test takers. In general, the experience of taking the test in the operational setting is necessary for panelists to understand the scope of what the test measures and the difficulty of the questions and tasks on the test. Taking the test was particularly important for the panelists of this study, because, as discussed in the previous section, only a few panelists indicated in their background questionnaire that they were very familiar with the test content.

Borderline Student Definition

As discussed earlier, the panelists were introduced to the CSE, the TOEFL iBT test, and the standard setting methodology and took the test on Day 1 of the standard setting meeting. Panelists then recommended cut scores on each of the four sections for the remaining days of the meeting. The first task on each of these days was to define minimum language skills needed to reach each of the targeted CSE levels (CSE 4 to CSE 8). This was in a way a continuation of the premeeting assignment. A student (test taker) who has these minimally acceptable skills is referred to as a just qualified candidate (JQC). These JQC descriptions served as the frame of reference for the standard setting judgments; that is, panelists were asked to consider the test questions in relation to these definitions. The following steps were used to form the JQC definitions for all targeted CSE levels:

- Panelists were asked to refer to their premeeting assignments in the preparation guide and to review a list of distinctive features prepared by the CSE team (see Appendix E).
- Two subpanels were formed each day, with panelists rotating in each subpanel, to encourage collaboration across all panelists.
- Panelists were asked to consult their premeeting assignment and write at least five statements describing important skills of the JQC at CSE Level 6.
- A whole-panel discussion was then organized, whereby the whole group reached consensus on the definition for the Level 6 JQC.
- Next, one subpanel was asked to prepare the Level 4 JQC, whereas the other subpanel was asked to prepare the Level 8 JQC. Each subpanel was asked to present the suggested JQCs to the other subpanel so that agreement was reached on the definitions.
- A whole-panel discussion followed to complete the definitions for Level 5 JQC and the Level 7 JQC.

The above steps were considered an efficient and effective way to develop the JQC definitions while allowing for adequate discussions of the JQC features among the panelists. The borderline student definitions for all four language skills can be found in Appendix F.

Standard Setting Method for Selected-Response Items

For the two test sections containing selected-response items (listening comprehension and reading comprehension), a modified Angoff procedure was employed (see Plake & Cizek, 2012, for a description of modified Angoff procedures). Following the development of the JQC definitions (see "Borderline Student Definition" section) for reading, panelists were trained in the modified Angoff standard-setting process and given an opportunity to practice their judgments. The facilitator first asked panelists to make judgments on the first three reading test items and discuss the rationale behind their judgments. The facilitator guided this instructional discussion and provided clarification on the procedure as needed. Each panelist was asked to complete an evaluation form indicating the extent to which the training was clear and whether the panelist was ready to proceed. All panelists indicated their readiness to proceed and were then instructed to independently review the items and record their judgments on a rating form.

The modified Angoff approach was implemented in three rounds of judgments informed by feedback and discussion between rounds. In Round 1, panelists were asked to judge how many out of 100 JQCs at CSE 4, CSE 6, and CSE 8 would know the correct answer. For example, for CSE 6, the question was stated as follows: "Imagine 100 JQCs at CSE 6. For each test question, how many of them would know the correct answer?"

The panelists used a judgment scale from 0 to 100 with 10-point increments and entered their judgments electronically on a rating form in Excel format (see Appendix G for a sample). The panelists were instructed to focus only on the alignment between the English language skills demanded by the test question and the English language skills possessed by the JQCs and not to factor random guessing into their judgments. After completing their first round of judgments, panelists received feedback on individual- and panel-level judgments. The sum of each panelist's cross-item judgments (divided by 100) represented this panelist's recommended cut score. Each panelist's recommended cut score was shown to them at the bottom of their individual Excel rating form.

The panel-recommended cut score and highest and lowest cut scores by an individual panelist were compiled and presented to the panel to foster discussion. The panel-recommended cut score was computed by taking the mean of the panelists' recommended cut scores. Panelists were then asked to share their judgment rationales and consider any changes to their judgments for Round 2. Panelists were also shown the *p* values, or the percentage of the 2,185 test takers (discussed in "Selection of Empirical Data From the TOEFL iBT Test" section) who answered each question correctly. Panelists were instructed to use the *p* values as a guide when considering relative difficulty of items, not as an indicator of the probability that a JQC would get an item correct. Because panelists were making judgments for three cut scores for each item, they were reminded that their numeric judgments would, by definition, be lower for CSE 6 compared to CSE 8, and for CSE 4 compared to the other two levels. After the group discussion concluded, panelists were asked to make Round 2 judgments.

In Round 2, judgments were made again at the test question level. Panelists were asked to take into account the panel-recommended cut score and the discussion from Round 1, as well as the empirical difficulty of the items, and were informed that they could make changes to their ratings for any question(s), for any CSE level. The Round 2 judgments were compiled, and the recommended Round 2 cut score was presented to the panel. Each panelist's recommended cut score was shown to them at the bottom of their individual Excel rating form.

In Round 3, panelists were asked to make holistic judgments, that is, to provide one cut score recommendation for the overall test section (e.g., reading comprehension) instead of item-level judgments (see Appendix G for a sample of the Round 3 rating form). Specifically, panelists were asked to review the JQC definitions for CSE 4, CSE 6, and CSE 8 and to decide on the recommended cut scores for these three levels, taking into account the Round 2 cut score recommendations and group discussions. Upon completing the task for these three levels, the panelists were then asked to review the JQC definitions for CSE 5 and CSE 7 and to decide on the recommended cut scores for these two levels. The transition to a holistic-level judgment placed emphasis on the overall language skill of interest (i.e., reading comprehension or listening comprehension) and the setting of cut scores for each test section. Upon completion of Round 3, panelists were shown the panel-recommended cut score for each level, as well as impact data, that is, the distribution of the 2,185 test takers' scores by CSE level based on the recommended cut scores, and were asked to discuss the reasonableness of the cut scores in terms of how TOEFL iBT test takers were classified into the CSE levels.

The three-round process was repeated with the listening comprehension section on the following day.

Standard Setting Method for Constructed-Response Items

For the test sections containing constructed-response items (speaking and writing), a variation of the performance profile method (Hambleton, Jaeger, Plake, & Mills, 2000) was followed in the TOEFL iBT standard setting workshop. This holistic standard setting method is desirable for test sections with constructed-response items because it allows panelists to review a set of student performance samples. As educators, panelists have expertise making judgments about samples of actual student work in a holistic fashion (Kingston & Tiemann, 2012). Standard setting was completed first for the speaking section on Day 4 of the meeting, and the last day of the meeting was dedicated to the writing section.

Panelists reviewed the responses of 41 test takers to the six speaking prompts and the responses of 48 test takers to the two writing prompts. The test takers were selected based on their score profiles, which represented the most frequently occurring task-score patterns across the speaking and writing section scores respectively from the test taking population described earlier. Similar to the procedure followed for selected-response items, three rounds of judgments occurred with feedback and discussion between rounds. The judgment task was presented as follows for CSE 6: "Imagine **ONE** CSE 6 JQC. Listen to responses of actual TOEFL iBT test takers. What speaking score would the CSE 6 JQC earn?" The recommendations for the speaking and writing cut scores were based on the final round of judgments.

For speaking, the audio files of the responses of 16 of the 41 test takers were played upon request by the panelists as they refined their judgments for each cut score. In addition to listening to speaking samples, each panelist was provided with a printed student profile sheet to facilitate the judgment process (see a sample of the list in Appendix H). Similarly, panelists received a binder of the written responses of the 48 students to both tasks and a printed student profile list for the writing section.

To make cut score recommendations, panelists were asked to review the borderline student descriptions for CSE 4, CSE 6, and CSE 8. The task in this method was to review the test takers' responses to speaking and writing tasks and decide the TOEFL iBT speaking and writing section scores a borderline student at each CSE level would receive. After Round 1, the panel's mean cut score and the minimum and maximum cut scores recommended by a single panelist were presented, and panelists shared their judgment rationales. Although a second round for the same levels was planned, as shown in the sample rating form (Appendix I), the panelists decided after the Round 1 discussion that a high level of agreement regarding the cut scores was already present and a Round 2 judgment was not needed. Therefore, the panelists decided to revise their Round 1 judgments (if they wanted) and provide cut scores for CSE 5 and CSE 7, as originally planned for Round 3. To avoid confusion in this report, we therefore refer to Round 1, whereby cut scores were recommended for CSE 4, CSE 6, and CSE 8, and Round 2, whereby cut scores from Round 1 were reviewed and cut scores for CSE 5 and CSE 7 were added. Similar to the selected-response test sections, impact data were also shown after Round 2 to inform panelists about the percent of students who would be classified into each of CSE levels based on the Round 2 cut scores.

Feedback and Discussion

At the final debriefing, panelists were shown the final recommended cut scores based on their judgments, as well as the resulting impact data for all test sections, and were once again asked to discuss their reasonableness. At the end of the last day, panelists were asked to complete a final evaluation form that asked questions about the process, the importance of various factors in the process, and which factors influenced their judgments. Panelists were also asked to indicate their level of confidence in the final set of recommended cut scores constructed during the process.

Results of the Standard Setting Study

The first set of results in this section summarizes the panel's standard setting judgments by round for each of the test sections. The results include the mean, median, minimum, maximum, and standard deviation (SD) of each round of judgments. The mean cut scores in the final round of judgments for each test section are considered the panel's final recommendations. The results are presented in raw scores for reading and listening, which is the metric that the panelists used. The cut scores for speaking and writing are provided on the 0–30 reported scale for TOEFL iBT, as the panelists had access to that information during the standard setting process, and the judgment task involved listening to or reading test-taker responses to the prompts of the speaking or writing section. The standard error of judgment (SEJ) is also included along with the other statistics as an estimate of the uncertainty in the panelists' judgments. The SEJ is computed by dividing the standard deviation of the judgments by the square root of the number of panelists (Cizek & Bunch, 2007). The SEJ can be interpreted as an indication of how close each recommended cut score is likely to be to a cut score recommended by other panels of experts similar in composition to the current panel and similarly trained in the same standard setting methods. A comparable panel's cut score would be within one SEJ of the cut score 68% of the time and within two SEJs 95% of the time. In order to reduce the impact on misclassification rates (false positives and false negatives), Cohen, Kane, and Crooks (1999) suggested that an SEJ should be no more than half the value of the standard error of measurement (SEM).

The results from the end-of-meeting survey completed by the panelists are also presented in this section. The information from the survey was collected to provide procedural validity evidence, for example, whether the procedures followed were practical, implemented properly, whether feedback given to the panelists was effective, and whether documentation had been sufficiently compiled (for a summary of the different types of validity evidence for standard setting, see Papageorgiou & Tannenbaum, 2016; for a detailed discussion see Hambleton, Pitoniak, & Copella, 2012).

Panel-Recommended Cut Scores for the Four Test Sections

The results for the cut scores of the reading section of the TOEFL iBT test are presented in Table 6. The mean cut score across the three rounds was similar for CSE 6 and CSE 8, whereas it increased by 1 point in Round 2 for CSE 4. In Round 3, the CSE 4 mean cut score was between the Round 1 and Round 2 mean value. The variability in panelists' judgments decreased in general across rounds, as can be seen by the standard deviation, suggesting some convergence in the final round of judgments (one exception was CSE 6, for which there was a minor increase from 2.48 *SD* in Round 2 to 2.50 *SD* in Round 3). The SEJ for each cut score was within half of the SEM for the reading section of the particular test form (2.15 for all test takers around the world and 1.85 for the 2,185 Chinese test takers).

The results for the listening section are presented in Table 7. The mean cut score across the three rounds was similar for CSE 4, CSE 6, and CSE 8 and decreased slightly from round to round. The standard deviation tended to decrease from Round 1 to Round 2, and remain similar in Round 3. The SEJ for each cut score was within half of the SEM for the listening section of the particular test form (2.21 for all test takers around the world and 2.31 for the 2,185 Chinese test takers).

The results for the speaking section are presented in Table 8. The mean cut score across Round 1 and Round 2 was similar for CSE 4, CSE 6, and CSE 8, and the standard deviation tended to decrease, suggesting convergence in the judgments. The SEJ for each cut score was within half of the SEM for the speaking section of the particular test form (1.30 for all test takers around the world and 1.59 for the 2,185 Chinese test takers).

Table 6 Standard Setting Results for the Reading Section of the TOEFL iBT Test

	Round 1			Round 2			Round 3				
Description	CSE 4	CSE 6	CSE 8	CSE 4	CSE 6	CSE 8	CSE 4	CSE 5	CSE 6	CSE 7	CSE 8
Mean	12.94	25.57	36.43	13.98	25.89	36.67	13.44	19.38	25.44	30.88	36.19
Median	13.15	25.45	37.00	14.35	25.70	37.40	14.00	19.50	25.00	31.50	37.00
Minimum	6.80	20.60	30.20	7.60	21.50	31.40	7.00	15.00	21.00	25.00	31.00
Maximum	19.10	29.80	41.40	18.50	29.70	41.30	18.00	23.00	29.00	35.00	41.00
SD	3.70	2.57	3.74	2.62	2.48	3.42	2.45	1.78	2.50	2.90	2.45
SEJ	0.92	0.64	0.93	0.66	0.62	0.86	0.61	0.45	0.63	0.72	0.85

Note. CSE = China's Standards of English Language Ability; SEJ = standard error of judgment.

Table 7 Standard Setting Results for the Listening Section of the TOEFL iBT Test

	Round 1			Round 2			Round 3				
Description	CSE 4	CSE 6	CSE 8	CSE 4	CSE 6	CSE 8	CSE 4	CSE 5	CSE 6	CSE 7	CSE 8
Mean	8.37	18.16	26.84	8.26	18.04	26.39	7.75	12.38	17.63	21.56	25.75
Median	8.40	18.20	27.45	8.60	18.00	27.25	8.00	12.50	17.50	22.50	26.50
Minimum	4.00	14.60	19.20	4.40	14.60	21.90	4.00	10.00	14.00	16.00	21.00
Maximum	12.40	22.50	32.80	10.80	20.70	29.60	10.00	14.00	20.00	24.00	29.00
SD	2.23	2.63	3.92	1.53	1.81	2.41	1.44	1.15	1.82	2.10	2.44
SEJ	0.56	0.66	0.98	0.38	0.45	0.60	0.36	0.29	0.46	0.52	0.61

Note. CSE = China's Standards of English Language Ability; SEJ = standard error of judgment.

Table 8 Standard Setting Results for the Speaking Section of the TOEFL iBT Test

	Round 1			Round 2				
Description	CSE 4	CSE 6	CSE 8	CSE 4	CSE 5	CSE 6	CSE 7	CSE 8
Mean	13.94	20.31	26.50	13.56	16.94	20.25	23.63	26.56
Median	15.00	20.00	27.00	14.00	17.00	20.00	24.00	27.00
Minimum	9.00	18.00	24.00	11.00	14.00	18.00	22.00	24.00
Maximum	17.00	23.00	28.00	15.00	19.00	23.00	26.00	28.00
SD	2.21	1.40	0.89	1.41	1.48	1.24	0.96	0.89
SEJ	0.55	0.35	0.22	0.35	0.37	0.31	0.24	0.22

Note. CSE = China's Standards of English Language Ability; SEJ = standard error of judgment.

The results for the writing section are presented in Table 9. The mean cut score across Round 1 and Round 2 was similar for CSE 4, CSE 6, and CSE 8, and the standard deviation tended to decrease, suggesting convergence in the judgments. The SEJ for each cut score was within half of the SEM for the speaking section of the particular test form (2.00 for all test takers around the world and 2.10 for the 2,185 Chinese test takers).

TOEFL iBT test scores for each test section are reported on a 0–30 score scale. To facilitate the standard setting judgment task for the selected response items, the panelists made recommendations based on raw scores. The panel-recommended cut scores can be easily converted to scale scores using the score conversion table ETS provided to the project team for the test form used in the standard setting meeting. However, the conversion process first requires a decision to be made about rounding the panel's recommended cut scores, because the conversion to scale scores requires whole raw scores (no decimals). There are two options for the rounding of the raw scores for listening and reading:

- The raw score is rounded up to the next achievable raw score. The rationale behind this decision is that the decimals indicate ability beyond a given score point. For example, a raw score of 13.44 means that the cut score should be 14 to indicate that the minimum score is above 13.
- The raw score is rounded down. The rationale behind this decision is that although the decimals indicate ability beyond a given score point, still the next higher score has not been achieved. Using the example above, a raw score

Table 9 Standard Setting Results for the Writing Section of the TOEFL iBT Test

	Round 1			Round 2				
Description	CSE 4	CSE 6	CSE 8	CSE 4	CSE 5	CSE 6	CSE 7	CSE 8
Mean	13.13	20.25	26.94	13.06	17.00	20.31	23.63	26.94
Median	13.50	20.00	27.00	13.00	17.00	20.00	24.00	27.00
Minimum	10.00	18.00	26.00	10.00	15.00	19.00	22.00	26.00
Maximum	16.00	22.00	28.00	15.00	19.00	22.00	25.00	28.00
SD	1.86	1.00	0.85	1.39	1.03	0.79	0.81	0.77
SEJ	0.46	0.25	0.21	0.35	0.26	0.20	0.20	0.19

Note. CSE = China's Standards of English Language Ability; SEJ = standard error of judgment.

Table 10 Raw and Scale Cut Scores for TOEFL iBT Test Sections Recommended by the Panel

	Panel-recommended cut scores (raw)					Cut scores converted on the reporting scale using two rounding approaches				
Test section	CSE 4	CSE 5	CSE 6	CSE 7	CSE 8	CSE 4	CSE 5	CSE 6	CSE 7	CSE 8
Reading	13.44	19.38	25.44	30.88	36.19	7/8	13/14	17/18	21	25
Listening	7.75	12.38	17.63	21.56	25.75	2/4	9/10	15/16	19/20	22/23
Speaking	13.56	16.94	20.25	23.63	26.56	13/14	16/17	20	23/24	26/27
Writing	13.06	17.00	20.31	23.63	26.94	13/14	17	20/21	23/24	26/27

Note. CSE = China's Standards of English Language Ability. When two scale score are shown, the lower value is the results of the "round down" rule, whereas the higher value is the result of the "round up" rule. In some cases, rounding up or rounding down did not make a difference.

of 13.44 means that the cut score should be 13, because the next higher score, 14, was not recommended by the panel.

The following factors can be considered to make a decision about rounding:

- Collection of data outside this study, which is discussed in the next section
- Implications for false positive and false negative classifications (see discussion in Papageorgiou, Tannenbaum, Bridgeman, & Cho, 2015)

Table 10 provides the conversion from raw to scale scores, using both rounding approaches. Rounding for the speaking and writing cut scores is conducted directly on the reported 0–30 scale, because as discussed earlier, the standard setting process allowed panelists direct reference to the scale scores.

Results of Meeting Evaluation Survey

Table 11 summarizes the panel's feedback regarding the general process followed in the standard setting meeting. The majority of panelists strongly agreed or agreed that they understood the purpose of the study, that the instructions and explanations provided by the meeting facilitators were clear, that the training provided for both methods was adequate, that the explanation of how the recommended cut scores were computed was clear, that there was adequate amount for discussion and feedback, and that feedback between rounds in terms of item-level and score distribution data was helpful. No panelists strongly disagreed with any statement.

Panelists were also asked to indicate their level of comfort with the final cut score recommendations (Table 12). All panelists were very comfortable with the cut scores for the constructed-response sections (speaking and writing). The majority of the panelists reported they were either very comfortable or somewhat comfortable with the recommended cut scores for the selected-response sections (reading and listening), whereas one panelist indicated lower level of confidence in the listening section cut score. This particular panelist wrote in the survey: "I was surprised to find out that the cut scores for receptive skills are actually lower than productive skills. The former skills are supposed to be less challenging than the latter to the best of my knowledge."

Table 11 Panelists' Feedback on the Standard Setting Process

Question	Strongly disagree	Disagree	Agree	Strongly agree
I understood the purpose of this study.	0	0		16
The instructions and explanations provided by the facilitators were clear.	0	0	1	15
The training in the Angoff standard setting method (Reading and Listening) was adequate to give me the information I needed to complete my assignment.	0	0	1	15
The training in the Profile standard setting method (Speaking and Writing) was adequate to give me the information I needed to complete my assignment.	0	0	1	15
The explanation of how the recommended cut score is computed was clear.	0	0	3	13
The opportunity for feedback and discussion between rounds was helpful.	0	0	1	15
The inclusion of the item and task data was helpful.	0	0	1	15
The inclusion of the classification percentages was helpful.	0	0	1	15

Table 12 Panelists' Confidence in the Standard Setting Results

Test section	Very uncomfortable	Somewhat uncomfortable	Somewhat comfortable	Very comfortable
Listening	0	1	7	8
Reading	0	0	3	13
Speaking	0	0	0	16
Writing	0	0	0	16

The panelist's comment suggests some misunderstanding about the way TOEFL iBT scores should be interpreted. According to ETS (2018), the four test sections have the same scale score range (0–30) to indicate that all sections should be viewed as equally important in measuring English language proficiency. However, identical scores in different sections do not represent identical percentile ranks within those sections and cannot be directly compared; for example, a score of 25 in the reading section will have a different percentile rank than a score of 25 in the listening section. Moreover, the score distributions by CSE level for each test section, which were presented to the panelists, confirmed this panelist's misunderstanding. Specifically, fewer test takers in the specific test form would be classified at higher CSE levels in speaking and writing than in reading or listening. Despite this panelist's comment, the group's overall confidence in the cut scores is satisfactory, and the higher confidence in the cut scores of the constructed response sections might be, to some extent, affected by the standard setting method. Evaluating sample speaking and writing responses is likely to be a more natural task for the panelists than judging the difficulty of reading and listening items.

Examination of the Reasonableness of the Cut Scores Through Teacher Judgment Data

The cut scores recommended by the standard-setting panel represent one source of information. A second, independent source of information was provided by a representative sample of classroom teachers. These teachers classified their students, who already had TOEFL iBT scores, into different CSE levels. This second source of information provides an opportunity to consider the reasonableness of the panel-based cut score recommendations. Using teacher classification in this way is consistent with recommendations in the manual (Council of Europe, 2009) in the context of mapping test scores to the CEFR levels.

A total of 304 teachers and 1,326 students were recruited for the external data collection. Of these teachers, 285 evaluated the language proficiency of the students in relation to the CSE levels for the four language skills (reading, listening, speaking, writing) as well as overall language ability. Most teachers evaluated the language proficiency of up to 10 of their students; this decision was made after the first teacher training session (see below), as a greater number of students was deemed unmanageable or would cause teacher fatigue. The teachers were first trained in applying the CSE level descriptions by NEEA staff through face-to-face or online sessions. Following training, teachers were asked to perform one holistic and one analytic judgment task for each language skill (reading, listening, speaking, and writing; see Appendix J for a sample). In addition, teachers were asked to evaluate the overall language ability of each student. The five holistic ratings were then compared to the five TOEFL iBT test scores (reading, listening, speaking, and writing, and total score).[2]

Table 13 Number of Students by School Type

School type	Number of students
Public high school	184
International high school	282
University	367
Private language training school	493
Total	1,326

Table 14 Number of Students by Location

Locations	Number of students
Beijing	190
Yangzhou	180
Haerbin	147
Shenzhen	137
Shanghai	127
Hangzhou	121
Guangzhou	106
Chengdu	101
Chongqing	99
Suzhou	98
Changchun	14
Zhenjiang	3
Nanning	3
Total	1,326

The analytic judgment task was implemented to help teachers with their understanding of the CSE levels for each of the language skills. For this task, teachers were asked to state how well the students could perform in relation to a small number of CSE descriptors, for which the level was not disclosed to the teachers (the descriptors were selected from different CSE levels). The analytic task also functioned as a check for the teachers' understanding of the holistic task. For example, erratic ratings in the analytic judgment task, such as use of the same rating for all descriptors (despite the differences in the CSE level of the descriptors), resulted in removal of the holistic judgments for a given student.

Almost all students took the test between Winter 2018 and Spring 2019, typically within 3 months from the time teacher classifications and ratings were collected. The students came from four types of schools, as can be seen in Table 13. In addition, the cities where the students were studying at the time of data collection can be seen in Table 14.

Descriptive statistics of the TOEFL iBT scores by CSE level placement are presented in Table 15 for all four test section scores, as well as the total score. It should be noted that the total number of students by language skill in the table is lower than the total number of students recruited following removal of some data discussed above. The results in general show an increase in the mean and median of TOEFL iBT scores by CSE level, as expected. The wide range of scores for each CSE level, as well as the standard deviation, indicates that teachers vary in their evaluation of students' language proficiency in relation to the CSE levels. The moderate to high correlations between test scores and the teachers' CSE level placement in Table 16 suggest that the teachers are in general successful in ranking their students in terms of language ability; nevertheless, the teachers seem less successful in making finer distinctions with regard to the placement of individual students into adjacent CSE levels. The teachers' difficulty in distinguishing between adjacent CSE levels might not be surprising, given that the teachers in this study did not receive the same extensive training in the use of the CSE levels, which included targeted discussion of the distinguishing features of each level, as the standard setting panel did.

To estimate cut scores based on the teachers' judgments, equipercentile equating was performed (for a nonmathematical introduction to equipercentile equating, see Livingston, 2014). Based on this equating method, the CSE levels assigned by the teachers and the TOEFL iBT test scores are considered equivalent if they have the same percentile rank for this group of students. The lowest equated scores for each CSE level are shown in Table 17 and can be used as the cut score for that level. Although there are several ways to calculate cut scores using teacher judgments as the criterion

Table 15 Descriptive Statistics of Student Scores

			TOEFL iBT scale score				
Language skill (number of students)	CSE level placement	Number of students	Mean	Median	Min	Max	SD
Reading (N = 1,183)	Below 4	157	10.17	10.00	0	26	5.72
	4	167	16.83	17.00	2	30	5.83
	5	276	19.07	20.00	2	29	5.69
	6	315	22.54	23.00	2	30	4.93
	7	210	25.08	26.00	0	30	4.57
	8 or higher	58	25.71	28.00	5	30	5.31
Listening (N = 1,183)	Below 4	197	9.70	9.00	1	25	4.97
	4	159	15.54	16.00	3	30	5.06
	5	293	18.65	19.00	2	30	4.85
	6	312	21.35	22.00	5	30	4.96
	7	162	24.49	26.00	2	30	4.98
	8 or higher	60	25.12	27.50	2	30	6.17
Speaking (N = 1,188)	Below 4	132	12.90	14.00	0	25	5.36
	4	275	18.05	18.00	6	27	3.42
	5	366	19.52	20.00	10	30	2.77
	6	263	20.88	22.00	4	30	3.44
	7	126	21.47	22.00	6	28	3.53
	8 or higher	26	24.38	23.00	19	30	3.65
Writing (N = 1,212)	Below 4	193	14.72	15.00	0	28	5.62
	4	212	17.67	18.00	5	28	4.37
	5	329	20.09	20.00	10	29	3.61
	6	276	20.71	21.00	9	28	3.62
	7	148	22.83	23.00	8	28	3.52
	8 or higher	54	23.89	24.00	15	30	3.35
Overall/total (N = 1,183)	Below 4	141	47.48	48.00	11	105	17.84
	4	200	67.22	68.50	24	109	17.79
	5	348	77.34	79.00	33	114	14.57
	6	296	86.39	88.50	24	115	14.72
	7	162	93.14	95.00	48	120	13.34
	8 or higher	141	102.67	106.50	78	114	10.53

Note. CSE = China's Standards of English Language Ability.

Table 16 Correlations Between CSE Level Placement and TOEFL iBT Test Scores

Language skill/test section	Correlation
Reading	.66
Listening	.70
Speaking	.58
Writing	.56
Overall ability/total test	.69

Note. CSE = China's Standards of English Language Ability.

measure, equipercentile equating was preferred because it provides easily interpretable results. Given the nature of the analysis, truncated (rounded down) scores are included in Table 17.

The cut scores derived from teachers' CSE level placement (Table 17) and the cut scores recommended by the panelists (Table 10) are not identical. Such discrepancy is not surprising, given the following considerations:

- **Amount and type of training.** The panelists and the teachers received different types and amount of training in using the CSE scales, with the panelists spending considerably more time applying the scales. Several of the panelists were involved in the development of the CSE scales, as noted elsewhere, so overall, the panel was more familiar with the scales than the teachers.

Table 17 Results of Equipercentile Equating Between CSE Level Placement and TOEFL iBT Test Scores

	TOEFL iBT scale score				
CSE level	Reading	Listening	Speaking	Writing	Total
8	29	29	27	27	111
7	27	26	24	25	101
6	23	22	22	22	90
5	18	17	19	19	76
4	13	13	16	16	61

Note. CSE = China's Standards of English Language Ability. The total score in this table is equated to the teachers' judgments of the students' overall language ability; TOEFL iBT test takers receive a total score which is the sum of the four section scores.

Table 18 Classification of Students Into Levels Based on Different Sets of Cut Scores

Student classification comparison	Same level	Within 1 level	Within 2 levels	Within 3 or more levels
Teacher placement vs. panel's recommended cut scores (round up rule)	368 (31.11%)	559 (47.25%)	207 (17.50%)	49 (4.14%)
Teacher placement vs. panel's recommended cut scores (round down rule)	306 (25.87%)	563 (47.59%)	245 (20.71%)	69 (5.83%)

- **Nature of tasks.** The panelists and the teachers were asked to perform different tasks. The panelists focused more closely on the test content and the performance expected by borderline students at each level. The teachers might have focused on what constitutes typical performance at a level, as shown in relevant standard setting research (e.g., Giraud, Impara, & Plake, 2005; Papageorgiou, 2010). Research also shows that even when the same standard setting method and panelist training are implemented for the same test form with different but comparable panels, cut scores are not exactly the same (see Tannenbaum & Kannan, 2015).
- **Truncated student sample**. A common issue with studies involving teacher judgments of student performance is that lower ability students are not always represented (see discussion of this issue in Papageorgiou & Cho, 2014). In fact, when examining the relationship between academic language proficiency tests and academic success, the issue of a truncated student sample is prominent, because only the performance of students who have received high enough test scores for university admission can be investigated (Cho & Bridgeman, 2012; Harsch, Ushioda, & Ladroue, 2017). Although the teachers placed several students at CSE 4 or even lower, experience suggests that not many students would take the TOEFL iBT test until they are ready to receive scores that are more likely to help them be admitted to higher education institutions.

Given the above considerations, a more meaningful way to compare the two sets of cut scores (cut scores proposed by panelists and cut scores based on teachers' classification of students) is not necessarily through a direct comparison of these cut scores but through an examination of the classification of the same group of students into CSE levels based on either set of cut scores. Table 18 presents this classification comparison for the total TOEFL iBT scores of the 1,183 students in Table 15. When comparing the teacher classification decisions (Table 17) with the panel's rounded up cut scores (Table 10), 78.36% of the test takers would be placed either at the same or within one CSE level. When the rounded down cut scores are used, this percentage is somewhat lower but comparable (73.46%). It should be pointed out that taking into account classification within one CSE level is appropriate, given that some students might have received adjacent TOEFL iBT scores, which would nevertheless locate them into different CSE levels (e.g., a total score of 101 is CSE 7, but a total score of 100 is CSE 6 in Table 17). Given the considerations presented in this section (e.g., different type and amount of training, different tasks), it could be argued that the teachers' CSE level placement of their students offers some support to the reasonableness of the panel-recommended cut scores.

The next section of the report presents the recommended mapping of TOEFL iBT test scores to the CSE levels for which the score concordance study between the TOEFL iBT test and IELTS (ETS, 2010) will also be considered.

Final Cut Score Recommendation

Consideration of a Score Concordance Study

As discussed in the previous section, the meaningfulness and credibility of the panel-based cut score recommendations were reinforced by the results of the teachers' classification of their students. In this current section, we now introduce a third source of information to consider for potentially adjusting the panel-recommended cut scores (presented in Table 10). This third source is a concordance study conducted between TOEFL iBT test scores and IELTS band levels (ETS, 2010). The concordance study may be considered objective, in that no expert judgments were collected; instead, only the test scores (TOEFL iBT and IELTS) from the same group of test takers were compared. Given that the IELTS bands have been mapped to the CSE levels, the results of the concordance study offer additional information for possibly adjusting the panel-based cut scores. However, it must be noted that while the two tests assess similar constructs and may be used for similar purposes—and so the concordance study was justified—they are not identical and cannot be considered interchangeable.

When interpreting results from score concordance studies, an important consideration is the way in which scores are reported. In the case of the TOEFL iBT test and IELTS, there are some very noticeable differences. The TOEFL iBT test uses a 0–120 total score scale and a 30-point score scale for the four test sections. The total score is the sum of the four section scores. IELTS uses a nine-level reporting system, inherited from its predecessor, the ELTS test (Davies, 2008). The speaking and writing sections of IELTS are scored on a nine-level rubric and the scaled scores for reading and listening are converted to one of the nine levels; the overall level is also reported on the nine-level system with half level points, based on the average of the section scores (Lim, Geranpayeh, Khalifa, & Buckendahl, 2013). Because the two tests use different score reporting mechanisms and because the TOEFL iBT score scales are more refined than the IELTS 9-band reporting scale, multiple TOEFL iBT scores may be equivalent to the same IELTS band. That is, there is no one-to-one mapping of scores from one test to the other.

When examining the results of the score concordance study to compare TOEFL iBT and IELTS cut scores onto the CSE levels, limitations regarding the student population sample, acknowledged by the authors of the technical report, should also be considered. This limitation of the population sample is noted even though the overall number of test takers in the score concordance study was high compared to other similar studies in the language testing field. For example, 1,153 test takers participated in the score concordance study between the TOEFL iBT test and IELTS, whereas a study comparing performance on the IELTS test and the Cambridge Advanced test was based on scores from only 186 test takers (Lim et al., 2013). It should also be pointed out that the purpose of the score concordance study was to offer some guidance to score users who are familiar with one test but not the other, not to directly compare test scores for the purposes of score mapping onto the CSE levels.

One final consideration is that the standard setting methodology to map the scores of each test to the CSE levels, while similar, was not identical, and interpretation of the CSE levels across the two panels of each score mapping study might vary even though nine panelists participated in both efforts. Green (2018) makes a similar observation regarding the mapping of TOEFL iBT scores and IELTS bands onto the CEFR levels. While the cut score for the Level C1 level is very similar for both tests, the interpretation of Levels B1 and B2 differs across tests. For example, IELTS Band 6 was found to be equivalent to a TOEFL iBT score range of 60–78 in the score concordance study. However, while IELTS Band 6 is mapped to Level B2 (Lim et al., 2013), the TOEFL iBT test score range for Level B2 is 72–94 (Papageorgiou, Tannenbaum, et al., 2015); therefore, based on the aligned TOEFL iBT scores, the lower end of IELTS Band 6 could be interpreted as CEFR Level B1 rather than Level B2. It should be noted that the use of TOEFL iBT total score of 72 as the cut score for CEFR Level B2 is further corroborated by a criterion-related study, which examined the relationship between TOEFL iBT test scores and the scores of a locally administered test in an English-medium institution (O'Dwyer, Kantarcıoğlu, & Thomas, 2018).

Appendix K offers a visual comparison of the mapping of TOEFL iBT scores based on the recommendation of the standard setting panel, the teachers' CSE classification of their students, and the score concordance study between the TOEFL iBT test and IELTS bands discussed above. The results from the standard setting panel and the concordance study have a similarly wide spread of scores even though the cut scores vary, whereas the teacher score mapping is based on a narrower range of the scores for reasons discussed in the previous section, primarily the inclusion of students at lower proficiency levels.

Table 19 Classification Accuracy and Consistency of TOEFL iBT Scores Mapped to CSE Levels (Test Takers Whose Responses Were Presented to the Standard Setting Panel)

	Cut scores based on the round down rule		Cut scores based on the round up rule	
Test section/score	Accuracy	Consistency	Accuracy	Consistency
Reading	0.67	0.57	0.66	0.56
Listening	0.71	0.62	0.69	0.59
Speaking	0.69	0.59	0.72	0.62
Writing	0.54	0.43	0.54	0.44
Total	0.80	0.72	0.79	0.71

Note. CSE = China's Standards of English Language Ability. The mapping of the total score was based on the sum of the section scores, which is the same way the total score is calculated on the operational TOEFL iBT test.

Table 20 Classification Accuracy and Consistency of TOEFL iBT Scores Mapped to CSE Levels (All 2018 Test Takers in China)

	Cut scores based on the round down rule		Cut scores based on the round up rule	
Test section/score	Accuracy	Consistency	Accuracy	Consistency
Reading	0.67	0.57	0.66	0.57
Listening	0.72	0.63	0.70	0.61
Speaking	0.68	0.57	0.68	0.59
Writing	0.54	0.43	0.55	0.44
Total	0.79	0.71	0.79	0.71

Note. CSE = China's Standards of English Language Ability. The mapping of the total score was based on the sum of the section scores, which is the same way the total score is calculated on the operational TOEFL iBT test.

Examination of Classification Accuracy and Consistency

Because rounding rules might result in somewhat different cut scores proposed by the panel (see Table 10), an analysis of classification accuracy and consistency was first performed for both rounding rules prior to providing the final cut score recommendation in this report. The aim of this analysis was to examine if either rounding rule results in notably higher classification accuracy and consistency. Classification *accuracy* refers to the extent to which classifications based on observed test scores match those that would have been made based on theoretical (error-free) true scores. Classification *consistency* indicates the extent to which classifications based on one test form would match the decisions based on scores from a second, parallel form of the same test based on the theoretical (error-free) true scores. A technical discussion on classification accuracy and consistency is beyond the purpose of this report; the reader is referred to Livingston and Lewis (1995), whose method was used in this analysis, and several studies applying their method in the context of English language tests (Papageorgiou, Morgan, & Becker, 2015; Papageorgiou, Xi, Morgan, & So, 2015; Powers et al., 2017). Analysis of classification accuracy and consistency was performed twice, first for the test takers in China whose responses were presented to the standard setting panelists (Table 19), and second, for the all test takers who took the TOEFL iBT test in China in 2018 (Table 20). The analysis was conducted for cut scores based on both rounding rules.

As Young and Yoon (1998) pointed out, there does not seem to be a "rule of thumb" for acceptable classification accuracy and consistency figures, and to the best of our knowledge, their observation remains valid. The results are comparable to those reported in the studies cited above for other English language assessments, which have similar psychometric characteristics that affect classification accuracy and consistency, that is, reliability and the number of reporting levels. Classification accuracy and consistency was higher for the total TOEFL iBT test score than section scores, as the total score is psychometrically more reliable.[3] Overall, classification accuracy and consistency appeared similar when classifying students into CSE levels based on the TOEFL iBT scores, irrespective of the rounding rule.

Recommended Mapping of TOEFL iBT Test Scores Onto the CSE Levels Based on All Data

Table 21 offers the recommended mapping of TOEFL iBT test scores onto the CSE levels. To arrive at the recommended mapping, the following decisions were made:

Table 21 Recommended Mapping of TOEFL iBT Test Scores Onto the CSE Levels

	Cut scores				
Test scores	CSE 4	CSE 5	CSE 6	CSE 7	CSE 8
Reading (0–30)	7	13	17	21	25
Listening (0–30)	4	10	16	20	23
Speaking (0–30)	13	17	20	23	26
Writing (0–30)	13	17	21	23	27
Total (0–120)	37	57	74	87	101

Note. CSE = China's Standards of English Language Ability. Numbers indicate the minimum TOEFL iBT scores a test taker should receive to be classified in a CSE level. The mapping of the total score was based on the sum of the section scores, which is the same way the total score is calculated on the operational TOEFL iBT test.

- Reading score: Rounded down cut scores were preferred over rounded up cut scores because they allow for more score points for CSE 6 and align somewhat better with the score mapping based on the TOEFL iBT–IELTS score concordance study.
- Listening scores: Rounded up cut scores were preferred over rounded down cut scores, because they align somewhat better with the score mapping based on the TOEFL iBT–IELTS score concordance study. Rounded up cut scores also result in a higher cut score for CSE 4, which is less likely to be achieved through guessing if the lower cut scores, based on the round down rule, are adopted.
- Speaking scores: Except for CSE 5, rounded down cut scores were preferred over rounded up cut scores because they align somewhat better with the score mapping based on the TOEFL iBT–IELTS score concordance study (note that the cut scores for CSE 6 are not affected by the rounding rule).
- Writing scores: For CSE 6 and CSE 8, rounded up cut scores were preferred over rounded down cut scores, whereas the opposite was the case for CSE 4 and CSE 7, to offer optimal alignment with the score mapping based on the TOEFL iBT–IELTS score concordance study (note that the cut scores for CSE 5 are not affected by the rounding rule).

Examining the Reasonableness of the Recommended Score Mapping Through the Link Between the CSE and CEFR Levels

Comparing the levels of different language frameworks through separate score mapping studies is not straightforward (see research in the volume edited by Tschirner, 2012). Nevertheless, such a comparison offers an additional perspective regarding the reasonableness of the recommended score mapping, by triangulating the relationships between TOEFL iBT, the CSE, and the CEFR. The CEFR is regarded as an external criterion here, with the TOEFL iBT scores and the CSE levels being linked to the CEFR levels in separate studies. The mapping of the TOEFL iBT test scores onto the CEFR levels is presented in Table 22, based on Papageorgiou, Tannenbaum, et al. (2015). NEEA has also conducted an empirical study to investigate the relationship between CEFR and CSE levels during the CSE development process (Liu & Peng, 2018). The study included embedding CEFR descriptors in the data collection carried out to scale the CSE descriptors then comparing the CSE levels that the descriptors were placed at in relation to their original CEFR levels. The link between CEFR and CSE levels based on the results of the NEEA study is presented in Figure 3 and can be summarized as follows:

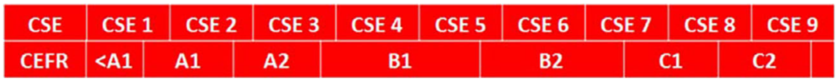
- CSE 8 is aligned with upper CEFR Level C1 and the lower CEFR Level C2.
- CSE 7 is aligned with upper CEFR Level B2 and lower CEFR Level C1.
- CSE 6 is aligned mainly with CEFR Level B2.
- CSE 5 is aligned with upper CEFR Level B1 and lower CEFR Level B2.
- CSE 4 is mostly aligned with CEFR Level B1.

The results from Tables 21 and 22, and Figure 3 are summarized in Table 23. The first line of Table 23 shows the TOEFL iBT total scores corresponding to the CSE levels based on Table 21. The second line in Table 23 shows how the CSE and CEFR levels are linked based on the separate mapping of TOEFL iBT total scores onto the CEFR levels (Table 22). The third line in Table 23 shows how the CSE and CEFR levels are linked based on the empirical study conducted by

Table 22 Score Mapping of the TOEFL iBT Test Onto the CEFR Levels

	Cut scores		
Test scores	CEFR B1	CEFR B2	CEFR C1
Reading (0–30)	4	18	24
Listening (0–30)	9	17	22
Speaking (0–30)	16	20	25
Writing (0–30)	13	17	24
Total (0–120)	42	72	95

Note. CEFR = Common European Framework of Reference.



CSE	CSE 1	CSE 2	CSE 3	CSE 4	CSE 5	CSE 6	CSE 7	CSE 8	CSE 9
CEFR	<A1	A1	A2	B1	B2	C1	C2		

Figure 3 Link between China's Standards of English Language Ability and the Common European Framework of Reference levels. From Liu and Peng (2018).

Table 23 Comparison of the Link Between the CSE and the CEFR Levels Based on Different Studies

Study	CSE 4	CSE 5	CSE 6	CSE 7	CSE 8
TOEFL iBT total scores mapped onto CSE levels	37–56	57–73	74–86	87–100	101–120
CEFR levels linked to CSE levels based on TOEFL iBT total scores	Below CEFR B1, CEFR B1	CEFR B1, CEFR B2	CEFR B2	CEFR B2, CEFR C1	CEFR C1
CEFR levels linked to CSE levels based on the NEEA study	CEFR B1	CEFR B1, CEFR B2	CEFR B2	CEFR B2, CEFR C1	CEFR C1, CEFR C2

Note. CSE = China's Standards of English Language Ability; CEFR = Common European Framework of Reference; NEEA = National Education Examinations Authority.

NEEA (Figure 3). The results in the second and third lines of Table 23 are nearly identical, with slight discrepancies at CSE 4 and CSE 8, given that TOEFL iBT does not currently target CEFR levels below B1 or above C1. This triangulation analysis shows that the results from separate studies converge, thus offering support to the proposed mapping of TOEFL iBT test scores onto the CSE levels. It should also be noted that the relationship between the CSE and the CEFR levels is not as precise as the mapping relationship between test scores and the two frameworks, given methodological differences in each study. Furthermore, given that the CSE and CEFR are separate frameworks with overlapping but different perspectives, contexts of use, and development procedures, exact alignment of their levels should not be expected.

Conclusion

In this report, we provided a detailed rationale behind the mapping of TOEFL iBT test scores onto the CSE levels, building on several sources of data to support the panel-based recommended cut scores. Although the different sources of data provide some support to the recommended cut scores, policymakers in the educational context where the CSE levels are used might want to further investigate the relationship between TOEFL iBT test scores and the CSE levels. Such exploration should focus on the relevance and usefulness of the recommended cut scores to facilitate score-based decision-making (see Papageorgiou, Tannenbaum, et al., 2015). In addition, it is important to investigate the impact of score alignment to the CSE levels to facilitate score-based decision-making, and to inform and guide educational policy (Wu, 2019).

In conclusion, we believe that this research project makes a useful contribution to the field of language assessment, as it demonstrates how evidence should be collected to support a claim about the alignment of test scores to the proficiency levels of a language framework. Our research underscores the importance of collecting data from multiple sources so that the alignment claim is convincing.

Based on our experience from this research project and our understanding of the literature on aligning test scores to the CEFR levels, we would also caution about potential issues in the context of aligning test scores to the CSE levels:

- Alignment to proficiency levels facilitates score interpretation, but score users and decision makers should not consider alignment to be sufficient evidence of the quality of a language test or sufficient support for score interpretation and use.
- Two language tests targeting the same proficiency levels should not be viewed as being equivalent in terms of content or difficulty, nor should their scores be considered interchangeable based solely on separate alignment studies.
- The relationship between a language test and the levels of a language proficiency framework is not necessarily simple, direct, or established as a one-time event. Revisions to the alignment of test scores to the proficiency levels might be required in light of additional evidence and improved understanding of how the language test and the language framework operationalize the underlying language ability construct.

Notes

1 A reference list of all publications to date related to TOEFL research can be found at https://www.ets.org/toefl_family/research

2 For the vast majority of the students, all five holistic ratings were provided by a single teacher. A total of 136 students were taught by different teachers, depending on the target language skill. As a result, some of these students received more than one rating of their overall language ability. For example, a student with two teachers, one for the receptive skills and another for the productive skills, could have received two holistic ratings of overall language ability, one by each teacher. For these few cases of students with more than one holistic judgment of overall language ability, the rating of the teacher who was listed first in the data set was used for convenience, as the correlation between the first teacher's holistic rating and the average of all teachers' holistic ratings was 0.99.

3 It should be noted that common statistical indices used to indicate score reliability are affected by the number of items in a test. However, as Bridgeman (2016) notes, treating tasks in constructed-response tests as items is conceptually problematic. For example, scores in the writing section of the TOEFL iBT test are not based on the total number of correct items, as is the case for the reading and listening sections, but on evaluation of test takers responses by raters using scoring rubrics combined with automated scoring.

References

Alderson, J. C. (1991). Bands and scores. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s* (pp. 71–86). London, England: Macmillan.

Alderson, J. C., & Banerjee, J. (2002). Language testing and assessment (Part 2). *Language Teaching*, *35*(2), 79–113. https://doi.org/10.1017/S0261444802001751

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. New York, NY: Longman.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, England: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford, England: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford, England: Oxford University Press.

Beaton, A., & Allen, N. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics*, *17*, 191–204. https://doi.org/10.2307/1165169

Bridgeman, B. (2016). Can a two-question test be reliable and valid for predicting academic success? *Educational Measurement: Issues and Practice*, *35*(4), 21–24. https://doi.org/10.1111/emip.12130

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 2–27). London, England: Longman.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York, NY: Routledge.

Cho, Y., & Bridgeman, B. (2012). Relationship of TOEFL iBT® scores to academic performance: Some evidence from American universities. *Language Testing*, *29*(3), 421 – 442. https://doi.org/10.1177/0265532211430368

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

Cohen, A. S., Kane, M. T., & Crooks, T. J. (1999). A generalized examinee-centered method for setting standards on achievement tests. *Applied Measurement in Education*, *12*, 343 – 366. https://doi.org/10.1207/S15324818AME1204_2

Council of Europe. (2001). *The Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge, England: 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*. Strasbourg, France: Council of Europe. Available from http://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680667a2d

Davies, A. (2008). *Assessing academic English: Testing English proficiency 1950 – 98 — The IELTS solution*. Cambridge, England: Cambridge University Press.

Educational Testing Service. (2010). *Comparing TOEFL® and IELTS™ total scores*. Retrieved from https://www.ets.org/toefl/institutions/scores/compare/

Educational Testing Service. (2018). *Reliability and comparability of TOEFL iBT® scores*. Retrieved from https://www.ets.org/s/toefl/pdf/toefl_ibt_research_s1v3.pdf

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing*, *28*(1), 5 – 29. https://doi.org/10.1177/0265532209359514

Giraud, G., Impara, J. C., & Plake, B. S. (2005). Teachers' conceptions of the target examinee in Angoff standard setting. *Applied Measurement in Education*, *18*(3), 223 – 232. https://doi.org/10.1207/s15324818ame1803_2

Green, A. (2018). Linking tests of English for academic purposes to the CEFR: The score user's perspective. *Language Assessment Quarterly*, *15*(1), 59 – 74. https://doi.org/10.1080/15434303.2017.1350685

Haberman, S. J., Sinharay, S., & Lee, Y.-H. (2011). *Statistical procedures to evaluate quality of scale anchoring* (Research Report No. RR-11-02). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2011.tb02238.x

Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. (2000). Setting performance standards on complex educational assessments. *Applied Psychological Measurement*, *24*(4), 355 – 366. https://doi.org/10.1177/01466210022031804

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

Hambleton, R. K., & Zenisky, A. L. (2013). Reporting test scores in more meaningful ways: Some new findings, research methods, and guidelines for score report design. In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology* (pp. 479 – 494). Washington, DC: American Psychological Association. https://doi.org/10.1037/14049-023

Harsch, C., Ushioda, E., & Ladroue, C. (2017). *Investigating the predictive validity of TOEFL iBT® scores and their use in informing policy in a U.K. university setting* (ETS Research Report No. RR-17-41). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12167

Jackson, H., & Stockwell, P. (2011). *An introduction to the nature and functions of language* (2nd ed.). London, England: Continuum International.

Jamieson, J., Jones, S., Kirsch, I., Mosenthal, P., & Taylor, C. (2000). *TOEFL 2000 Framework: A working paper*. Princeton, NJ: Educational Testing Service. Retrieved from https://www.ets.org/Media/Research/pdf/RM-00-03.pdf

Kane, M. (2012). Validating score interpretations and uses. *Language Testing*, *29*(1), 3 – 17. https://doi.org/10.1177/0265532211417210

Kingston, N. M., & Tiemann, G. C. (2012). Setting performance standards on complex assessments: The body of work method. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 201 – 224). New York, NY: Routledge.

Knapp, P., & Watkins, M. (2005). *Genre, text, grammar: Technologies for teaching and assessing writing*. Sydney, Australia: University of New South Wales Press Ltd. & University of New South Wales.

Lim, G. S., Geranpayeh, A., Khalifa, H., & Buckendahl, C. W. (2013). Standard setting to an international reference framework: Implications for theory and practice. *International Journal of Testing*, *13*(1), 32 – 49. https://doi.org/10.1080/15305058.2012.678526

Liu, J. (2015). Some thoughts on developing China common framework for English language proficiency. *China Examinations*, *1*, 7 – 11.

Liu, J., & Han, B. (2018). Theoretical considerations for developing use-oriented China's Standards of English. *Modern Foreign Languages*, *1*, 78 – 90.

Liu, J., & Peng, C. (2018, December). *Aligning CSE with CEFR*. Paper presented at the 4th International Conference on Language Testing and Assessment, Beijing, China.

Livingston, S. A. (2014). *Equating test scores (without IRT)*. Princeton, NJ: Educational Testing Service. Retrieved from https://www.ets.org/Media/Research/pdf/LIVINGSTON2ed.pdf

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.

Milanovic, M., & Weir, C. J. (2010). Series editors' note. In W. Martyniuk (Ed.), *Relating language examinations to the Common European Framework of Reference for Languages: Case studies and reflections on the use of the Council of Europe's draft manual* (pp. viii–xx). Cambridge, UK: Cambridge University Press.

National Education Examinations Authority (NEEA). (2018). *China's Standards of English Language Ability*. Beijing, China: Higher Education Press & Shanghai Foreign Language Education Press. Retrieved from http://cse.neea.edu.cn/html1/report/18112/9627-1.htm

O'Dwyer, J., Kantarcıoğlu, E., & Thomas, C. (2018). *An investigation of the predictive validity of the TOEFL iBT® test at an English-medium university in Turkey* (TOEFL Research Report No. RR-83). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12230

Papageorgiou, S. (2010). Investigating the decision-making process of standard setting participants. *Language Testing*, *27*(2), 261 – 82. https://doi.org/10.1177/0265532209349472

Papageorgiou, S., & Cho, Y. (2014). An investigation of the use of TOEFL Junior Standard scores for ESL placement decisions in secondary education. *Language Testing*, *31*(2), 223 – 239. https://doi.org/10.1177/0265532213499750

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). Princeton, NJ: Educational Testing Service.

Papageorgiou, S., Xi, X., Morgan, R., & So, Y. (2015). Developing and validating band levels and descriptors for reporting overall examinee performance. *Language Assessment Quarterly*, *12*(2), 153 – 177. https://doi.org/10.1080/15434303.2015.1008480

Pearson Standards and Quality Office. (2014). *Writing descriptors: Guidelines and best practice*. London, England: Pearson.

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

Powers, D., Schedl, M., & Papageorgiou, S. (2017). 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

Ryan, J. (2006). Practices, issues, and trends in student test score reporting. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 677 – 710). Mahwah, NJ: Erlbaum.

State Council of P. R. China. (2014). *The implementation opinions of the State Council on deepening the reform of the examination and enrollment system*. Retrieved from http://www.gov.cn/zhengce/content/2014-09/04/content_9065.htm

Tannenbaum, R. J. (2019). Validity aspects of score reporting. In. D. Zapata-Rivera (Ed.), *Score reporting research and applications* (pp. 9 – 18). New York, NY: Routledge. https://doi.org/10.4324/9781351136501-2

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

Tannenbaum, R. J., & Kannan, P. (2015). Consistency of Angoff-based standard-setting judgments: Are item judgments and passing scores replicable across different panels of experts? *Educational Assessment*, *20*(1), 66 – 78. https://doi.org/10.1080/10627197.2015.997619

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

Wu, S. (2019). The anticipated impact of aligning international English tests to China's Standards of English Language Ability. *Modern Foreign Languages*, *5*, 672 – 683.

Young, M. J., & Yoon, B. (1998). *Estimating the consistency and accuracy of classifications in a standards-referenced assessment*. Retrieved from https://cresst.org/wp-content/uploads/TECH475.pdf

Zenisky, A. L., & Hambleton, R. K. (2012). Developing test score reports that work: The process and best practices for effective communication. *Educational Measurement: Issues and Practice*, *31*(2), 21 – 26. https://doi.org/10.1111/j.1745-3992.2012.00231.x

Appendix A

Group Members

Steering Group Members

Name	Institution
Jennifer Brown	Educational Testing Service
Richard J. Tannenbaum	Educational Testing Service
Spiros Papageorgiou	Educational Testing Service
Wu Sha	National Education Examinations Authority
Liu Jianda	Guangdong University of Foreign Studies
He Lianzhen	Zhejiang University
Zhang Wenxia	Tsinghua University

Working Group Members

Name	Institution
Ching-Ni Hsieh	Educational Testing Service
Venessa Manna	Educational Testing Service
Susan Nissan	Educational Testing Service
John Norris	Educational Testing Service
Lixiong Gu	Educational Testing Service
Lin Wang	Educational Testing Service
Li Tao	ETS Global B.V., China
Cheng Mengmeng	National Education Examinations Authority
Zheng Mingming	National Education Examinations Authority
Yang Fan	National Education Examinations Authority
Min Shangchao	Zhejiang University
Zhang Jie	Shanghai University of Finance and Economics
Pan Mingwei	Shanghai International Studies University
Wang Weiqiang	Guangdong University of Foreign Studies
Cai Hongwen	Guangdong University of Foreign Studies
Gao Manman	Anhui University
Jie Wei	Shanghai Jiao Tong University
Wu Xuefeng	Nanjing Forestry University
Xu Yun	Minzu University of China
Yang Lvna	Beijing Normal University
Zhang Hao	Tsinghua University

Appendix B

The Standard Setting Panelists and Their Affiliation

Panelist name (English)	Panelist name (Chinese)	Institution
Luo Kaizhou	罗凯洲	Beijing Foreign Studies University
Li Mei	李梅	Tongji University
Gao Miao	高淼	Central University of Finance and Economics
Gu Xiangdong	辜向东	Chongqing University
Guo Qian	郭茜	Tsinghua University
Huang Yongliang	黄永亮	Hebei University
Jia Yidong	贾贻东	Shandong University of Finance and Economics
Li Ruilian	栗瑞莲	Beijing Haidian Teachers Training College
Liu Wei	刘威	Tsinghua High School
Rong Li	荣丽	Zhejiang University of Technology
Wang Mingyue	王明月	The High School Affiliated to Renmin University of China, Chaoyang Branch School
Wang Shuhua	王淑花	Beijing Wuzi University
Zhang Wenjuan	张文娟	China University of Political Science and Law
Xu Yun	胥云	Minzu University of China
Yan Yi	颜奕	Tsinghua University
Yu Weishen	余渭深	Chongqing University

Appendix C

Sample Task for Panelist Familiarization With the CSE Levels

In preparation for the workshop, please complete this activity about the distinguishing features of CSE levels 4, 5, 6, 7, and 8 for reading

Your task: For CSE levels 4 through 8, please list what you think are some of the distinguishing features that separate each level from the level above and below (e.g., the features of CSE level 5 which distinguish that level from CSE level 4 and CSE level 6). Please consult the overall reading scale above, as well as the reading scales in Appendix II. List between 3 and 5 distinguishing features. You may write a few key words or 1–2 sentences for each feature.

CSE level	*Distinguishing features for reading*
8	1.
	2.
	3.
	4.
	5.
7	1.
	2.
	3.
	4.
	5.
6	1.
	2.
	3.
	4.
	5.
5	1.
	2.
	3.
	4.
	5.
4	1.
	2.
	3.
	4.
	5.

Appendix D

Tentative Schedule for the Standard Setting Workshop, Included in the Panelist Preparation Guide

Monday, July 9, 2018

- Registration
- Welcome and overview of the week
- Introduction to China's Standards of English Language Ability
- Introduction to the TOEFL iBT test
- Introduction to the standard setting methodology
- Receive materials
- Lunch
- Taking the TOEFL iBT test in authorized test center
- Welcome dinner courtesy of ETS

Tuesday, July 10, 2018

- Developing just qualified definitions for relevant CSE levels for Reading

- Training on standard-setting method for receptive skills
- Break
- Round 1 judgments for Reading
- Lunch
- Round 1 discussion and Round 2 judgments for Reading
- Break
- Round 3 revision and finalization of cut scores for Reading
- Adjourn for the day

Wednesday, July 11, 2018

- Developing just qualified definitions for relevant CSE levels for Listening
- Review standard-setting method for receptive skills
- Break
- Round 1 judgments for Listening
- Lunch
- Round 1 discussion and Round 2 judgments for Listening
- Break
- Round 3 revision and finalization of cut scores for Listening
- Adjourn for the day

Thursday, July 12, 2018

- Developing just qualified definitions for relevant CSE levels for Speaking
- Training on standard-setting method for productive skills
- Break
- Round 1 judgments for Speaking
- Lunch
- Round 1 discussion and Round 2 judgments for Speaking
- Break
- Round 3 revision and finalization of cut scores for Speaking
- Adjourn for the day

Friday, July 13, 2018

- Developing just qualified definitions for relevant CSE levels for Writing
- Review standard-setting method for productive skills
- Break
- Round 1 judgments for Writing
- Lunch
- Round 1 discussion and Round 2 judgments for Writing
- Break
- Round 3 revision and finalization of cut scores for Writing
- End-of-workshop evaluation survey
- Adjourn for the day

Appendix E

List of Distinctive Features Compiled by the CSE Team

Reading

Level	Distinguishing features
9	The reading materials of Level 9 focus on **research literature and academic monographs, original literary works, highly specialized expository works, etc.** And the language of the reading materials is complicated or abstract with plenty of terms or technical expressions beyond one's field of study. Cognitive ability concentrates on **making appraisal about the inner meaning, implications, or connotations.**
8	Representative materials of this level are argumentations including **academic paper or research literature**, other text types like **literary works, political and economic newspaper articles** are also included. Language of the materials is commonly complex, but the topics are **relevant to the reader's field of study**. The cognitive ability mainly focuses on **making analysis, appraisals or judgments**.
7	Level 7 has the largest number of descriptors, indicating a wide range of reading materials which distribute evenly across different text types. Most of the materials are **linguistically complex**, with part of them relatively complex. Representative texts are **prose essays, play scripts, literary works, research literature, government documents, long letters**, etc. Cognitive ability focuses on **making inference, analysis, evaluation and appraisals, etc.**
6	Level 6 has the second largest number of descriptors and a wide range of text types. Most of the materials are relatively complex. Text types contain **descriptive articles, stories, literary works, novels, fairy tales, popular science articles, common news stories, expository texts, instructions, commercial correspondences, long dialogues, etc.** Cognitive skills consist of **inferring the author's attitude, making critical analysis, summarizing the main idea, evaluating the content or language of the materials, etc.**
5	The range of reading materials of this level is also wide, which converges on **short prose essays, stories, poems, popular science articles, practical writings, technical requirements, argumentations, commentaries, reviews, dialogues, etc.** The materials are all relatively complex, which is easier than those in Level 6 and the topics **are within social and daily areas. The cognitive aims concentrate on making comparative analysis, extracting specific information, and comprehending the key points, etc.**
4	**Narrations, expositions and argumentations** are the most representative reading materials of Level 4. Typical text types are **stories, anecdotes, short articles, excerpts of novels, popular science articles, expository essays, simple news stories, argumentative texts, speeches, etc.** Language difficulty of the reading materials ranges **from simple to relatively complex**. Reading in this level mainly requires the ability of **identifying details, understanding the general idea, analyzing the viewpoints, etc.**
3	Text types of Level 3 centre on simple short essays, **simple narrations like anecdotes, fables and abridged version of biographies. Simple news stories, popular science stories, simple manuals and daily letters are also included**. Readers read the materials mainly to understand the general meaning or key content.
2	The most common reading materials of Level 2 are **simple stories, dairies, short expository essays, notes, etc. with simple language. Topics of the texts are all about daily or personal life**. Cognitive aims mainly concentrate on **recognizing or picking out specific words or phrases**.
1	The text types of this level mainly include **simple and short narrations** like **picture books, nursery rhymes, simple stories, etc**. And learners read only to get **the main ideas or specific words.**

Listening

Levels	Speech rate	Phonological features	Words, phrases and content	Topic	Cognitive processing
9	Regardless of speech rate	Regardless of accent	Containing low-frequency words; containing low-frequency colloquial expressions or jargons; sophisticated vocabulary	All kinds of topics	Make analyses, inferences and evaluations; understand allusions; understand their social, cultural and historical meaning
8	Regardless of speech rate	Regardless of accent	Different English varieties; containing colloquial expressions; containing technical terms	A wide range of topics; academic talks close to one's own field	Evaluate the rationality and logic of opinions of different sides
7	Regardless of speech rate	Standard pronunciation	Containing slangs or idioms; containing puns and metaphors; containing technical terms	Academic talks in one's own field; political, economic, and cultural issues; public policies and social issues; familiar topics	Evaluate speakers' opinions and stance; summarize the main content; understand sociocultural connotations; evaluate speakers' main points
6	Normal	Standard pronunciation	Highly informative; subtle; complex	In one's own field; current and political issues	Summarize the main content; compare opinions of different sides; evaluate the appropriacy of speakers' expressions
5	Normal	Standard pronunciation; articulated clearly	Complex (novels)	General topics; familiar topics; social issues; related to job and study; popular science	Summarize the main idea; obtain key points and details; understand speakers' implied meaning.
4	Normal	Standard pronunciation; articulated clearly	Simple; short (conversations)	General topics; topics of personal interest; familiar topics	Grasp the main idea; obtain factual information; understand speakers' intentions
3	Slow but natural	Standard pronunciation; articulated clearly	In simple language; simple; short (conversations)	Familiar topics	Obtain key information; identify the themes
2	Slow	Standard pronunciation; articulated clearly	In simple language; simple; containing few low-frequency words	Daily life	Obtain specific information
1	Slow	Articulated clearly	In simple language; simple; common	–	Identify words and phrases

Speaking

CSE level	Distinguishing features
9	The speaking performance of Level 9 covers mainly evaluative talks, such as **negotiation, explanation and justification, etc**. And the topics of speaking are complicated or abstract professional/social/cultural issues. The linguistic features are **effective, extensive, accurate and skilful**.
8	Representative performances of this level are communicative and expressive talks, such as **discussion, consultation, interpretation, argumentation, etc.** The topics are wide range professional and academic issues in formal and informal settings. The linguistic features at this level are **accurate, fluent, concise, and coherent.**
7	Level 7 has a large number of descriptors, speaking performances at this level are **rich, extensive and various** across different communicative activities. The topics are familiar, personal and specific. The linguistic features are appropriate, clear, detailed and explicit.
6	Level 6 has a relatively large number of descriptors and **a wide range** of speaking performance. Most of the performance activities are relatively general, such as **hot social issues and familiar topics** in one's field. The linguistic features are appropriate, insightful, effective, complete and adequate.
5	Level 5 has the second **largest number of descriptors**. Speaking performances cover on factual oriented tasks such as description, explanation, comparison, etc. The topics at this level include e**veryday topics and familiar and popular** social issues, matters of daily life, etc. Linguistic features at this level are **brief, clear, organized and logical.**
4	Level 4 has the largest number of descriptors. Speaking performance mainly focuses on description, explanation, narration. Topics are often **personal or of one's interests, or related to one's work or school life.** Linguistic features at this level are brief, appropriate and coherent.
3	Level 3 descriptors focus on describing speaking performances of relatively factual expression, such as **description, instruction, discussion, retelling, etc**. Topics at this level are familiar, common events. Linguistic features are **short, brief, simple.**
2	The most common activities at level 2 are **narration, description**, using simple terms and words. Topics are familiar, general in everyday life. Linguistic features are **simple and clear**.
1	The speaking performances at this level are mainly **expressive and descriptive**. Linguistic utterances are simple terms and words, talking about **familiar events in routine communicative activities**.

Writing

CSE Level	Distinguishing Features
9	The topics of Level 9 are high-demanding creative or academic writing including **scenery depiction, responding to academic journal articles, product specifications, contracts and agreements, etc**. What is written should display a high level of **vividness, multiple angles, evidence sufficiency, persuasiveness, and formality**. Language of this level is supposed to be **appropriate, natural, elegant, etc**.
8	Level 8 covers topics that are less demanding but are still very complex, including **culture, history, society, scientific and academic research articles**. What is written should display the features as **objectiveness, clarity, logic, well-organized structure, comprehensiveness**, etc. Linguistic quality takes such factors into consideration as **appropriateness, coherence, clarity, etc.**
7	The topics of Level 7 are less abstract, covering **novel plots, personal experiences, news, social and natural phenomena, instructions, announcements, comments on artistic works, conference minutes, statistical description**, etc. What is written should be **clear, accurate, complete, standardized, logical and formal**. Language quality focuses on **appropriateness, accuracy, etc**.
6	Level 6 covers a wide range of topics including **product description, description of well-known figures, science-fiction stories, movie summaries, experiment reports, project plans, academic abstracts, social hot issues, research data analysis, posters, letters of congratulation or complaint**, etc. What is written should be clear, accurate, sufficient and salient with details, attentive to readers, complete, logical and persuasive. Language of this level is supposed to be **clear, accurate, well-organized, detailed,** etc.
5	Level 5 has the largest number of descriptors, covering topics that are much less complicated but closer to everyday life. The topics are **description of familiar people, scenes, products or settings, description of personal experiences, short plays, data description, activity plans, job application letters, letters or emails about school life**, etc. What is written should be of **concreteness, completeness, vividness, accuracy, clarity, logic**, etc. Language of this level is expected to be **appropriate, clear, coherent,** etc.

CSE Level	Distinguishing Features
4	Level 4 also has the largest number of descriptors covering topics that are much closer to everyday life, including **descriptions of hometowns or campuses, familiar people or activities, favorite movies, short stories, personal feelings, abstract but familiar concepts, suggestions, self-introduction,** etc. What is written should be **clear, brief, coherent**. Language of this level focuses on **clarity**.
3	The topics of Level 3 are much more familiar, covering **recent experiences, favorite places, short story completion, procedure of simple tasks, school rules, simple directions, simple letters or emails,** etc. What is written should be brief, complete and clear. Language of this level is supposed to be **clear**.
2	Level 2 has topics that are closely associated with the authentic life of all individuals, including **wishes, dreams, familiar objects, people or places, family life, weather, everyday transportation, plans, opinions or attitudes, listing reasons, simple notices, etc**. What is written should be **brief and clear**. Language quality of the level focuses on the use of **simple words or short sentences.**
1	The topics of Level 1 are quite simple covering **descriptions of simple pictures, description of common animals, likes and dislikes, family activities, simple greetings, modes of transportations,** etc. What is written should be **brief**. Language quality of this level is the **use of simple words or phrases**.

Appendix F

Borderline Student Definitions by the Panelists

Reading

CSE level	Distinguishing features
8	- Can understand the aims, methods, and conclusions of an academic paper in relevant fields of study. - Can differentiate opinions from facts in commentaries written in complex language. - Can comprehend and analyze the relationship between different factors in specialized fields. - Can summarize the main ideas in linguistically complex articles. - Can synthesize the content in linguistically complex academic materials of familiar -fields of study. - Can analyze the logic of arguments in linguistically complex argumentative writing.
7	**Linguistically complex materials in specialized fields (science and technology, social sciences)** - Discern the key information. - Comprehend the implicit meaning. - Summarize/extract main features/ideas. - Infer feelings/attitudes.
6	- Infer mood and attitude while reading medium difficulty linguistic materials, such as literary works. - Understand relatively complex, medium-difficulty language. - Understand subject-related materials. - Summarize the main ideas of relatively complex passages (expository, popular science). - Locate significant details in relatively complex passages (literary works, subject-related materials).
5	**Medium difficulty linguistic material, a variety of general topics** - Grasp the essential meaning. - Analyze linguistic features. - Understand the cultural implications. - Distinguish different positions.
4	**Simple linguistic materials on general topics** - Locate detailed information. - Summarize the main ideas. - Differentiate facts and opinions. - Make simple inferences. - Understand the relationship between ideas.

Listening

CSE level	Distinguishing features
8	- Academic discourse related to one's own field. - Mass media. - Regardless of accent and speech rate. - Comprehend main ideas and supporting details. - Understand implied meanings and social and cultural connotations. - Evaluate main points/opinions.
7	- Abstract topics (e.g., politics, economy, culture, etc.). - Regardless of speech rate. - Understand implied meaning embedded in common rhetorical devices. - Extract key information. - Evaluate main points/opinions.
6	- Identify speakers' attitudes, intentions and opinions. - Summarize main ideas and content. - Input is delivered in normal speed and standard pronunciation. - Information is within one's field. - Follow lectures, presentations, discussions and negotiations.
5	- Normal speed. - Standard pronunciation. - General topics. - Obtain main ideas and supporting details, factual information. - Understand courses, lectures and talks. - Understand conversations and interactions. - Understand radio, film and TV programs.
4	- Normal speed. - Standard pronunciation. - Familiar topic and general topics of personal interests. - Follow simple oral descriptions. - Understand main ideas, opinions and general instructions.

Speaking

CSE level	Distinguishing features
8	- A wide range of topics in formal and informal settings. - Professional topics at academic seminars. - Accurately, fluently, coherently, appropriately, vividly. - Explain, discuss, analyze, summarize, negotiate, elaborate.
7	- A variety of familiar topics; (personal opinions) on abstract topics; formal academic presentation. - Fully, relatively in-depth, extensively, confidently. - Discuss, comment, explain, synthesize.
6	- Describe, discuss, respond during interaction, comment on one's opinion, negotiate, paraphrase. - Hot social issues or familiar topics in one's own field. - Effectively, clearly, adequately, in detail.
5	- Everyday topics, familiar topics, popular social issues. - Comment, describe, explain, consult, present. - In an organized, logical and clear manner.
4	- Personal, familiar topics of interest. - Describe, explain, narrate, respond. - Briefly, coherently, smoothly, in a relatively complete manner.

Writing

CSE level	Distinguishing features
8	- Complex social issues; academic articles, relevant literature, formal letters. - Clear, well-organized, logical fluent, comprehensive. - Sufficient evidence, in-depth discussion, reliable conclusion, objective description and analysis. - Describe, discuss, summarize, evaluate, elaborate.
7	- Abstract topics, academic writing (process and findings), complex narrative, formal instructions and announcements. - Clear and convincing, complex sentences, cohesive devices, strong evidence. - Analyze, elaborate, summarize, describe, integrate (data), comment.
6	- Popular genres (e.g., news reports, book review). - Article abstracts. - Social issues/topics/phenomena. - Describe, explain, analyze, clarify, summarize. - Relatively accurate and clear. - Generally appropriate and complete structure. - Mostly provide sufficient evidence, appropriate detail. - Straightforward argument (pros and cons).
5	- Topics of interest, reports related to one's field, common practical writing. - Clearly, relatively detailed, clear and convincing viewpoints. - Describe, explain, discuss, comment.
4	- Familiar topics. - Briefly, coherently, relatively persuasive. - Some evidence. - Common rhetoric devices. - Narrate, explain, express, describe.

Appendix G

Sample Panelist Forms for Reading

Sample Panelist Rating Form for Reading (Round 1 and Round 2)

	Panelist		
	1		
Item	Level 4	Level 6	Level 8
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			
17			
18			
19			
20			
21			
22			
23			
24			
25			
26			
27			
28			
29			
30			
31			
32			
33			
34			
35			
36			
37			
38			
39			
40			
41			
42			
Total/100	0.00	0.00	0.00
Cut score	0	0	0

Sample panelist rating form for Reading (Round 3)

	Panelist				
	1				
Cut score	Level 4	Level 5	Level 6	Level 7	Level 8
Round 1					
Round 2					
Round 3					

Appendix H

Sample of Test Taker Scores Used for Setting Speaking Cut Scores

Test taker	Total raw score	Total scale score	Notes
41	24	30	
40	23	29	
…			
30	18	23	
29	18	23	
…			
3	8	10	
2	7	9	
1	6	8	

Appendix I

Sample Panelist Rating Form for Speaking (All Rounds)

	Panelist				
	1				
Score	Level 4	Level 5	Level 6	Level 7	Level 8
Round 1		▨		▨	
Round 2		▨		▨	
Round 3					

Appendix J

Sample Teacher CSE Judgment Form for Reading (Both Holistic and Analytic Judgment Task)

Teacher’s name: _______ Student’s name: _______

Holistic Judgment

Based on China’s Standards of English Language Ability (CSE) and your understanding of this student, please tick the appropriate CSE level for this student. **As a rule of thumb, to help you make your judgment, consider at what level a student is able to confidently do 50% or more of the tasks expected at that level. The overall reading comprehension scale is provided below for your reference.**

Below CSE 4	CSE 4	CSE 5	CSE 6	CSE 7	CSE 8	Beyond CSE 8

Analytic Judgment

Following the instruction scale below, please rate the student’s performance on each of the subsequent descriptors. Please tick the appropriate number from 0–4.

Instruction Scale:

0	**1**	**2**	**3**	**4**
No, cannot do it	**Yes, in favorable circumstances**	**Yes, in normal circumstances**	**Yes, even in difficult circumstances**	**Clearly better than this**
Could not be expected to perform like this. The descriptor describes a level which is definitely **beyond** his/her capabilities.	**Could** be expected to perform like this provided that **circumstances** are **favorable**.	**Could** be expected to perform like this **without support** in **normal circumstances**.	**Could** be expected to perform like this even in **difficult circumstances**.	**Could** perform **better** than this in any circumstance. The descriptor describes a performance which is **clearly below** his/her level.

Favorable circumstances: e.g., a familiar topic and circumstance, an appropriate support or hint; some time to think about what to say; the interlocutor who is tolerant and prepared to help out.
Difficult circumstances: e.g., an unfamiliar topic; a surprising or interfering circumstance; limited time to think about what to say; a less cooperative interlocutor.

Descriptors	Cannot do it	⟷			Better than this
Can understand and summarize the author’s viewpoints and stances in commentaries written in relatively complex language.	0	1	2	3	4
Can summarize the main ideas in linguistically complex articles in political or economic newspapers and journals.	0	1	2	3	4
Can get the main idea of book reviews in relevant fields of study.	0	1	2	3	4
Can recognize details (e.g., time, character, and place) in articles on social life, such as travel notes, written in relatively complex language.	0	1	2	3	4
Can generalize the main ideas of literature reviews in relevant disciplines.	0	1	2	3	4
Can comprehend the implicit meaning of specialized linguistically complex materials by relating the materials to similar topics.	0	1	2	3	4
Can summarize the viewpoints and arguments in commentaries on familiar topics.	0	1	2	3	4

Overall reading comprehension scale

CSE 9	• Can understand linguistically complex materials from a variety of fields, analyzing them synthetically from multiple perspectives. • Can synthetically appraise complex and abstruse specialized materials from relevant fields of study.
CSE 8	• Can discriminate and appreciate aesthetic language use and social significance of linguistically complex materials from a wide range of topics. • Can appraise, by means of text analysis, the language and content of linguistically complex academic materials from familiar fields of study.
CSE 7	• Can synthesize the content of specialized linguistically complex materials (e.g., original literary works, science and technology literature, social commentaries), and analyze the author's viewpoint and stance. • Can make critical comments on a variety of cultural phenomena from different cultures, as presented in linguistically complex works. • Can comprehend the implicit meaning of specialized linguistically complex materials by relating the materials to similar topics.
CSE 6	• Can grasp significant relevant information and briefly comment on the language and content of subject-related materials of medium linguistic difficulty (e.g., literary works, news reports, business documents). • Can infer the writer's mood and attitude while reading materials of medium linguistic difficulty (e.g., literary works, news reports). • Can locate target information by scanning the indices of academic literature.
CSE 5	• Can grasp essential meaning, analyze linguistic features, and understand cultural implications whilst reading materials of medium linguistic difficulty on a variety of topics likely to be encountered in the domains of education, technology, and culture. • Can distinguish different positions in materials of medium linguistic difficulty containing opposing argumentation (e.g., editorials, book reviews).
CSE 4	• Can locate detailed information and summarize the main idea whilst reading different kinds of linguistically simple materials (e.g., simple short stories, essays, letters). • Can differentiate facts and opinions and make simple inferences in linguistically simple narratives and argumentative texts on a variety of topics. • Can understand the relationship between ideas by analyzing the structures of sentences and discourse whilst reading materials of medium linguistic difficulty.
CSE 3	• Can locate key information in linguistically simple practical forms of writing (e.g., letters, notices, signs). • Can understand the implicit meaning and summarize the main points of short, linguistically simple materials on familiar topics. • Can understand the relationship between points of information with the help of connectors in linguistically simple argumentative texts on familiar topics.
CSE 2	• Can acquire specific information and understand the main idea of short, linguistically simple essays on familiar topics. • Can understand short, simple texts containing new words with the help of pictures or other methods.
CSE 1	• Can understand very short, simple texts and locate basic information (e.g., characters, time, place). • Can understand simple materials (e.g., children's songs and nursery rhymes) and identify common words.

Appendix K

Comparison of the Mapping of TOEFL iBT Scores Onto the CSE Levels

Total Score

Cut scores by the standard setting panel	Cut scores based on teachers' CSE level placement of students	Cut scores derived from the TOEFL iBT-IELTS score concordance study
120	120	120
119	119	119
118	118	118
117	117	117
116	116	116
115	115	115
114	114	114
113	113	113
112	112	112
111	111	111
110	110	110
109	109	109
108	108	108
107	107	107
106	106	106
105	105	105
104	104	104
103	103	103
102	102	102
101	101	101
100	100	100
99	99	99
98	98	98
97	97	97
96	96	96
95	95	95
94	94	94
93	93	93
92	92	92
91	91	91
90	90	90
89	89	89
88	88	88
87	87	87
86	86	86
85	85	85
84	84	84
83	83	83
82	82	82
81	81	81

80	80	80
79	79	79
78	78	78
77	77	77
76	76	76
75	75	75
74	74	74
73	73	73
72	72	72
71	71	71
70	70	70
69	69	69
68	68	68
67	67	67
66	66	66
65	65	65
64	64	64
63	63	63
62	62	62
61	61	61
60	60	60
59	59	59
58	58	58
57	57	57
56	56	56
55	55	55
54	54	54
53	53	53
52	52	52
51	51	51
50	50	50
49	49	49
48	48	48
47	47	47
46	46	46
45	45	45
44	44	44
43	43	43
42	42	42
41	41	41
40	40	40
39	39	39
38	38	38
37	37	37
36	36	36
35	35	35
34	34	34
33	33	33

Note. Green = CSE 8; yellow = CSE 7, blue = CSE 6; orange = CSE 5; gray = CSE 4; white = no reference to any level. Dotted lines in the first column of each table indicate possible cut scores depending on which rounding rule is followed (see Table 10 for details on the rounding rule)

Reading Cut Scores

Cut scores by the standard setting panel	Cut scores based on teachers' CSE level placement of students	Cut scores derived from the TOEFL iBT–IELTS score concordance study
🟩 30	🟩 30	🟩 30
🟩 29	🟩 29	🟩 29
🟩 28	🟨 28	🟩 28
🟩 27	🟨 27	🟩 27
🟩 26	🟦 26	🟨 26
🟩 25	🟦 25	🟨 25
🟨 24	🟦 24	🟨 24
🟨 23	🟦 23	🟦 23
🟨 22	🟧 22	🟦 22
🟨 21	🟧 21	🟦 21
🟦 20	🟧 20	🟦 20
🟦 19	🟧 19	🟦 19
🟦 18	🟧 18	🟦 18
🟦 17	▨ 17	🟦 17
🟧 16	▨ 16	🟦 16
🟧 15	▨ 15	🟦 15
🟧 14	▨ 14	🟦 14
🟧 13	▨ 13	🟦 13
▨ 12	12	🟧 12
▨ 11	11	🟧 11
▨ 10	10	🟧 10
▨ 9	9	🟧 9
▨ 8	8	🟧 8
▨ 7	7	▨ 7
6	6	▨ 6
5	5	▨ 5
4	4	▨ 4
3	3	▨ 3
2	2	2
1	1	1
0	0	0

Note. Green = CSE 8; yellow = CSE 7, blue = CSE 6; orange = CSE 5; gray = CSE 4; white = no reference to any level. Dotted lines in the first column of each table indicate possible cut scores depending on which rounding rule is followed (see Table 10 for details on the rounding rule).

Listening Cut Scores

Cut scores by the standard setting panel	Cut scores based on teachers' CSE level placement of students	Cut scores derived from the TOEFL iBT–IELTS score concordance study
🟩 30	🟩 30	🟩 30
🟩 29	🟩 29	🟩 29
🟩 28	🟨 28	🟨 28
🟩 27	🟨 27	🟨 27
🟩 26	🟨 26	🟦 26
🟩 25	🟦 25	🟦 25
🟩 24	🟦 24	🟦 24
🟩 23	🟦 23	🟦 23
🟩 22	🟦 22	🟦 22
🟨 21	🟧 21	🟦 21
🟨 20	🟧 20	🟦 20
🟨 19	🟧 19	🟧 19
🟦 18	🟧 18	🟧 18
🟦 17	🟧 17	🟧 17
🟦 16	⬛ 16	🟧 16
🟦 15	⬛ 15	🟧 15
🟧 14	⬛ 14	🟧 14
🟧 13	⬛ 13	🟧 13
🟧 12	12	🟧 12
🟧 11	11	⬛ 11
🟧 10	10	⬛ 10
🟧 9	9	⬛ 9
⬛ 8	8	⬛ 8
⬛ 7	7	⬛ 7
⬛ 6	6	⬛ 6
⬛ 5	5	⬛ 5
⬛ 4	4	⬛ 4
⬛ 3	3	3
⬛ 2	2	2
1	1	1
0	0	0

Note. Green = CSE 8; yellow = CSE 7, blue = CSE 6; orange = CSE 5; gray = CSE 4; white = no reference to any level. Dotted lines in the first column of each table indicate possible cut scores depending on which rounding rule is followed (see Table 10 for details on the rounding rule).

Speaking Cut Scores

Cut scores by the standard setting panel	Cut scores based on teachers' CSE level placement of students	Cut scores derived from the TOEFL iBT–IELTS score concordance study
🟩 30	🟩 30	🟩 30
🟩 29	🟩 29	🟩 29
🟩 28	🟩 28	🟩 28
🟩 27	🟩 27	🟩 27
🟩 26	🟨 26	🟩 26
🟨 25	🟨 25	🟩 25
🟨 24	🟨 24	🟩 24
🟨 23	🟦 23	🟩 23
🟦 22	🟦 22	🟨 22
🟦 21	🟧 21	🟨 21
🟦 20	🟧 20	🟨 20
🟧 19	🟧 19	🟦 19
🟧 18	▨ 18	🟦 18
🟧 17	▨ 17	🟧 17
🟧 16	▨ 16	🟧 16
▨ 15	15	▨ 15
▨ 14	14	▨ 14
▨ 13	13	13
12	12	12
11	11	11
10	10	10
9	9	9
8	8	8
7	7	7
6	6	6
5	5	5
4	4	4
3	3	3
2	2	2
1	1	1
0	0	0

Note. Green = CSE 8; yellow = CSE 7, blue = CSE 6; orange = CSE 5; gray = CSE 4; white = no reference to any level. Dotted lines in the first column of each table indicate possible cut scores depending on which rounding rule is followed (see Table 10 for details on the rounding rule).

Writing Cut Scores

Cut scores by the standard setting panel	Cut scores based on teachers' CSE level placement of students	Cut scores derived from the TOEFL iBT–IELTS score concordance study
30	30	30
29	29	29
28	28	28
27	27	27
26	26	26
25	25	25
24	24	24
23	23	23
22	22	22
21	21	21
20	20	20
19	19	19
18	18	18
17	17	17
16	16	16
15	15	15
14	14	14
13	13	13
12	12	12
11	11	11
10	10	10
9	9	9
8	8	8
7	7	7
6	6	6
5	5	5
4	4	4
3	3	3
2	2	2
1	1	1
0	0	0

Note. Green = CSE 8; yellow = CSE 7, blue = CSE 6; orange = CSE 5; gray = CSE 4; white = no reference to any level. Dotted lines in the first column of each table indicate possible cut scores depending on which rounding rule is followed (see Table 10 for details on the rounding rule).

Suggested citation:

Papageorgiou, S., Wu, S., Hsieh, C.-N., Tannenbaum, R. J., & Cheng, M. M. (2019). *Mapping the* TOEFL iBT® *test scores to China's Standards of English Language Ability: Implications for score interpretation and use* (Research Report No. TOEFL-RR-89). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12281

Action Editor: Donald Powers

Reviewers: Brent Bridgeman and Jonathan Schmidgall