Chinese/English Journal of Educational Measurement and Evaluation | 教育测量与评估双语季刊

Volume 3 | Issue 1 Article 1

2022

Aligning Language Test Scores to Local Proficiency Levels: The Case of China’s Standards of English Language Ability (CSE)

Spiros Papageorgiou
ETS

Sha Wu
NEEA

Ching-Ni Hsieh
ETS

Richard J Tannenbaum
ETS

Mengmeng Cheng
NEEA

Recommended Citation

Papageorgiou, Spiros; Wu, Sha; Hsieh, Ching-Ni; Tannenbaum, Richard J; and Cheng, Mengmeng (2022) "Aligning Language Test Scores to Local Proficiency Levels: The Case of China’s Standards of English Language Ability (CSE)," *Chinese/English Journal of Educational Measurement and Evaluation | 教育测量与评估双语季刊*: Vol. 3: Iss. 1, Article 1.
Available at: https://www.ce-jeme.org/journal/vol3/iss1/1

Aligning Language Test Scores to Local Proficiency Levels: The Case of China's Standards of English Language Ability (CSE)

Spiros Papageorgiou [a], Sha Wu [b], Ching-Ni Hsieh [a], Richard J. Tannenbaum [a], and Mengmeng Cheng [b]

[a]Educational Testing Service
[b]National Educational Examinations Authority (Beijing, China)

Abstract
The past decade has seen an emerging interest in aligning test scores to language proficiency levels of external performance scales or frameworks, such as the Common European Framework of Reference (CEFR). 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 documented, and multiple sources of evidence should be collected. This paper demonstrates the steps in building an argument for aligning the scores of an international English language proficiency test to the levels of China's Standards of English Language Ability, or CSE, a localized language proficiency framework for English as a foreign language. Aligning an international examination to a localized framework serves to make the test score more relevant to the intended context of its use. We discuss the contextual issues that should be considered when interpreting test scores in relation to local proficiency levels, given the potential impact of score-based decisions on individuals and institutions. The implications for similar alignment research will also be presented.

Keywords
TOEFL Junior® tests;
China's Standards of English Language Ability;
Standard setting;
Cut scores;
Alignment;
Score interpretation

1 Introduction

Numerical scores typically do not convey direct information about what test takers 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 et al., 2011) for describing this procedure.

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, alignment can often make the interpretation of test scores more meaningful (Kane, 2012; Powers et al., 2017). 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 documented, and multiple sources of evidence should be collected.

The widespread use of the CEFR as a reference system 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. However, contextual issues should be carefully considered when interpreting test scores in relation to external proficiency levels, because of the consequences for both individuals and institutions when making score-based decisions within a specific

CONTACT: Spiros Papageorgiou. Educational Testing Service. 660 Rosedale Road, Princeton NJ, 08534, United States. spapageorgiou@ets.org.

educational or social context. 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 (Liu & Wu, 2019).

Recent studies linking scores of international language tests to the CSE levels have focused on adult language learners (e.g., Dunlea et al., 2019; Papageorgiou et al., 2019). This paper adds to the literature by exploring the alignment process in the context of language assessment for young learners through a collaborative research project between ETS and NEEA. Young learner populations have traditionally received less attention than adult learner populations, as often the focus of alignment studies has been on EFL tests used to inform admissions into higher education. However, young learners constitute the largest population of English learners in an educational system.

One issue explored in this study was the extent to which higher CSE levels are relevant for mapping scores of tests intended for young learners. As discussed later, the CSE presents language proficiency in nine levels. The open question for our study was how to accumulate compelling evidence regarding the upper limit of CSE-relevant proficiency for young learners. One way is by gathering evidence of construct congruence, where subject-matter experts evaluate the overlap between the proficiency levels and the test content. Evidence of construct congruence often acts as a gate-way for convening a standard setting panel as part of the alignment process. If the congruence is low, then moving forward with a standard setting study is not warranted. Although such a decision makes sense, nevertheless, when dealing with relatively unexplored populations (young English language learners in China) and relatively new language proficiency frameworks (as is the CSE), challenging evidence from the construct congruence analysis or, conversely, buttressing that evidence with the results of standard setting study seems appropriate—a form of triangulation of evidence. If the two sources converge, that is compelling evidence; if they diverge, that is important evidence establishing a warrant for deeper investigation to understand the interaction between the population, proficiency levels, and test content. In this way, convening a standard setting panel is not simply a procedural step following a construct congruence study but is a source of validation of the congruence study. This shift in the utility of standard setting panels as part of the overall alignment process, we believe, is important to bring to others' attention and consideration, especially when dealing with less explored populations and framework, and it is the path we followed in the present study.

2 Context of the Research Study

2.1 The CSE

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, the National Education Examinations Authority (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 language test development, either spoken or written. However, in comparison with their European counterparts, English learners and users in China are more prone to use English in an educational context.

The theoretical underpinnings of the CSE include models of language ability such as the Communicative Language Ability (CLA) model (Bachman, 1990; Bachman & Palmer, 1996, 2010), and an action-oriented approach to the description of language use (Council of Europe, 2001). This approach views users and learners of a language primarily as 'social agents' who have tasks (not exclusively language-related) to accomplish in certain circumstances, environments and fields of action. Language use is treated as being composed of the actions performed by individuals and social agents as they develop both general and communicative language competences. Overall, the CSE adopts 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.

Using the above theoretical basis, the CSE developers formulated a descriptive scheme, in which 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 project to develop the CSE was conducted between 2014 and 2017 in three phases. The first phase dealt with collecting descriptors from the literature but also from a database of descriptors generated by students and teachers of different educational levels. In the second phase, the CSE developers used expert and teacher judgments and removed duplicate descriptors, blended similar descriptors, and categorized descriptors into scales for the various subabilities and global language ability. The last phase was composed of two field studies for the finalization of the scaling of the descriptors. 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 or 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 easier reference, these levels are further grouped into three stages: elementary (CSE 1–3), intermediate (CSE 4–6), and advanced (CSE 7–9). More details about the CSE can be found in Papageorgiou et al. (2019).

2.2 Description of the *TOEFL Junior* Tests

The *TOEFL Junior* tests were developed by Educational Testing Service (ETS), using input from English language educators around the world. The tests measure the English communication skills of students ages 11 or older in English-medium instructional environments. The primary purpose of the *TOEFL Junior* tests is classroom placement and progress monitoring (Gu et al., 2015; Papageorgiou & Cho, 2014). The *TOEFL Junior* Design Framework points out that though some test tasks assess underlying enabling skills, such as grammatical and lexical knowledge, the main emphasis of the tests is the measurement of communicative competence, that is, the ability to use language for communicative purposes (Educational Testing Service, 2019; So et al., 2015).

The Target Language Use (TLU) domain (see Bachman & Palmer, 1996, 2010) for *TOEFL Junior* is divided into three subdomains:

- Social and interpersonal subdomain, where communication takes place in English for social and interpersonal purposes, such as having a casual conversation with classmates
- Navigational subdomain, where communication takes place in English for navigational purposes, for example reading or listening to an announcement or exchanging clarification questions about a school event
- Academic subdomain, where communication takes place in English for academic purposes, for example listening to an academic lecture, or reading academic texts

There are two *TOEFL Junior* tests: the *TOEFL Junior* Standard test, which can be delivered on paper or digitally and the digitally delivered *TOEFL Junior* Speaking test. The *TOEFL Junior* Standard test includes three sections: Listening Comprehension, Language Form and Meaning, or LFM, and Reading Comprehension. Each section is scored on a scale of 200–300, with increments of five points, resulting in a total score of 600–900. The test takes about two hours to complete. The *TOEFL Junior* Speaking test assesses the degree to which students have the speaking skills required by English-medium instructional environments. Test takers wear noise-reducing headphones and speak into a microphone to record their responses to the four speaking tasks. Each speaking task is designed to measure the test takers' ability to communicate in one of the three TLU subdomains. It should be noted that three out of the four tasks require test takers to understand language input, either written or spoken. This decision was made to measure integrated language skills for communication, thus reflecting language use in the real world. The spoken responses are digitally recorded and sent to the ETS online scoring network to be rated by human raters. Each task is scored using a 0 to 4 scoring rubric. Scores on the *TOEFL Junior* Speaking test are reported on a scale from 0 to 16. To help teachers, students, and parents better understand the meaning of the scores, the scoring rubric is available online[1].

The score reports of both *TOEFL Junior* Standard and *TOEFL Junior* Speaking contain performance descriptors to facilitate score interpretation. To help stakeholders familiarize themselves with the *TOEFL Junior* item types, sample items are available on the *TOEFL Junior* website[2].

3 Method

To collect sufficient evidence of the mapping of *TOEFL Junior* test scores on the CSE levels, our study included three main components: Investigation of construct congruence, standard setting (score mapping) meeting, and finalization of the official score mapping. The three components are described in detail next.

3.1 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 et al., 2011, p. 8); consequently, external level descriptors may not provide information that is directly relevant to test performance. 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 (see also the Specification stage in Council of Europe, 2009).

Because external levels and descriptors cannot be specific to any given test, nor can they function as the blueprint for test design, the *TOEFL Junior* tests are not necessarily a point-by-point reflection of the English language skills and expectations presented in the CSE. This lack of point-by-point correspondence is not a limitation of the tests, 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, we reviewed sample test forms and other materials related to the design and intended difficulty of the test that are available on the *TOEFL Junior* website. Based on this review, 32 CSE scales (eight per test section) were found to be the most relevant to the *TOEFL Junior* test content. These were used for the definition of borderline students during the standard setting meeting (described in the next section). In addition, the initial construct congruence analysis suggested that the focus of the score mapping study should be on levels 2 to 6, with the caveat, as discussed in the Introduction section, that the upper limit of CSE-relevant proficiency for young learners was an open question for our study. It should be noted that the Organizational Competence scales were found to be the most relevant ones for the Language Form and Meaning section of the *TOEFL Junior* Standard test. To further investigate construct congruence between the CSE levels and the *TOEFL Junior* tests, two members of the authoring team reviewed the test form to be used in the standard setting meeting and identified the CSE descriptors from levels 2 to 6 that are more closely aligned with the test content. The two researchers examined the *TOEFL Junior* Standard test form used in this study, as well as the tasks and scoring rubrics for the *TOEFL Junior* Speaking test and reached consensus on the selected descriptors.

3.2 Standard Setting Meeting

After construct congruence was established, a minimum score (cut score) on the tests needed to be identified for each CSE proficiency level. The cut score was intended to indicate the lowest point on the test score scale that test takers demonstrate performance according to a specific CSE level. Cut scores were established following a recognized standard setting process (Cizek & Bunch, 2007).

In preparation for the standard setting meeting, we collected item-level response data and score distribution information on one *TOEFL Junior* Standard test form and one *TOEFL Junior* Speaking test form, which were used by the panelists during the standard setting meeting. The *TOEFL Junior* Standard data were collected from 20,182 test takers world-wide who took the test form used for standard setting; 751 of the 20,182 test takers were located in China. The mean total scale score for the overall group was 734.29, whereas the mean total score for the test takers

[1] https://www.ets.org/toefl_junior/scores_research/speaking

[2] https://www.ets.org/toefl_junior

in China was somewhat higher (757.14). For the *TOEFL Junior* Speaking test, responses of test takers used during the meeting were selected from 5,842 test takers who took the test form around the world. The mean score was 8.02, and the most common first languages were Portuguese, Turkish and Spanish.

The standard setting meeting took place in China. The panelists were selected by the NEEA and included 16 female educators, all sufficiently fluent in English. The panelists represented a variety of Chinese institutions involved in English language teaching, representing the age groups targeted by the *TOEFL Junior* tests. Panelists completed a background questionnaire prior to the standard setting meeting. At the time of the study, two panelists indicated that they primarily worked with teachers as trainers, whereas the remaining fourteen panelists indicated that they were teaching young learners in the following contexts: Public junior high school (five panelists), public senior high school (five panelists), international junior high school (one panelist), and international senior high school (three panelists). With the exception of four panelists, all of them had over five years of experience teaching English, with half of the panelists indicating that they had more than 10 years of experience.

Prior to the standard-setting meeting, a preparation guide was prepared and sent to the panelists. The guide included information about the CSE and the *TOEFL Junior* tests presented in Section 2, as well as preparatory activities related to the 32 scales selected in the construct congruence stage. The purpose of the activities was to ensure that the panelists had a good understanding of the features that distinguished each of the five CSE levels, i.e., Levels 2 to 6, selected for the study. The panelists also brought their guide with the completed familiarization tasks to the standard-setting meeting. On each day of the meeting members of the CSE development team introduced the scales that were relevant to the test section of interest on that day, and organized short quizzes based on the above familiarization activities, to further strengthen the panelists' familiarity with the CSE levels. The creation of the preparation guide and the organization of quizzes at the beginning of each day were deemed essential because, unlike the panelists in previous CSE studies (Dunlea et al., 2019; Papageorgiou et al., 2019) who were involved at least in some aspect of the development of the CSE levels and descriptors, the panelists in this study were not familiar with the CSE levels and descriptors.

On the first day of the standard setting meeting both the *TOEFL Junior* Standard and the *TOEFL Junior* Speaking tests were presented to the panelists in the same way they are delivered to actual test takers, so that the panelists could understand the scope of what the test measures and the difficulty of the questions and tasks on the test. We also gave presentations related to the CSE, the *TOEFL Junior* tests, and the standard setting methodology.

The panelists recommended cut scores on each of the four test sections for the remainder of the four 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 2 to CSE 6). This was in a way a continuation of the pre-meeting 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 steps we followed to form the JQC definitions were identical to those reported in Papageorgiou et al. (2019).

For the three test sections of the*TOEFL Junior* Standard test which contained selected-response items (Reading Comprehension, Listening Comprehension, and Language Form and Meaning), a modified Angoff procedure was employed (Plake & Cizek, 2012). Following the development of the JQC definitions for reading, panelists were trained in the modified Angoff standard-setting process and given an opportunity to practice their judgments. The panelists first made judgments on the first three reading test items and discussed the rationale behind their judgments. The lead facilitator (one of the authors) guided this instructional discussion and provided clarification on the procedure as needed. Each panelist was asked to complete an evaluation form indicating the extent to which the training was clear and whether or not the panelist was ready to proceed. 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 2, 4, and 6 would answer each reading question correctly. They used a judgment scale from 0 to 100 with 10-point increments and entered their judgments electronically on a rating form in Excel format. 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. Panelists were also shown the percentage of the 20,182 test takers who answered each question correctly. After the group discussion concluded, panelists were asked to make Round 2 judgments again at the test question level, taking into account the panel-recommended cut score, the discussion from Round 1, as well as the empirical difficulty of the items. The Round 2 judgments were compiled, and the recommended Round 2 cut score was presented to the panel.

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 A for a sample). The transition to a holistic-level judgment places 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 20,182 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 many test takers were classified into the CSE levels. In addition to offering further opportunity for discussion, the rationale behind presenting the impact data was to capture the panelists' reaction to the reasonableness of the cut scores in a post-workshop survey, in which the panelists were asked to indicate their confidence in the recommended cut scores. The three-round process was repeated with the listening comprehension section on the following day, followed by Language Form and Meaning a day later.

For the *TOEFL Junior* Speaking test, a variation of the Performance Profile method (Hambleton et al., 2000) was followed on the last day of the standard setting meeting. This holistic standard-setting method was selected because of the constructed-response format, as 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). Prior to the meeting, the responses of 45 test takers to the four speaking prompts were selected based on their score profiles, which represented the most frequently occurring task-score patterns from the test taking population. Similar to the procedure followed for selected-response items, three rounds of judgments occurred with feedback and discussion between rounds. The audio files of the responses of 18 of the 45 test takers were played upon request by the panelists, as they refined their judgments for each cut score. The responses of the first test taker were used to also train the panelists in the standard setting method. All panelists indicated that they were ready to proceed upon completion of the standard setting task for the first test taker. In addition to listening to test taker responses, each panelist was provided with a printed student profile sheet to facilitate the judgment process (see sample of the list in Appendix B). The recommendation for the speaking cut scores were based on the final round of judgments.

To make cut score recommendations, panelists were asked to review the JQC descriptions for CSE Level 2, CSE Level 4, and CSE Level 6. The task in this method was to review the test takers' responses to the speaking tasks and decide on the test score the JQC at each CSE level would most likely receive. It should be clarified that the panelists' decision was at the test level, that this, the score the test taker receives, not at the individual prompt level. After Round 1, the panel's mean cut score, along with 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 C), the panelists decided after the Round 1 discussion that there was already a high level of agreement regarding the cut scores 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 3 and CSE 5, as originally planned for Round 3. To avoid confusion, we refer to Round 1, whereby cut scores were recommended for CSE 2, CSE 4, and CSE 6, and Round 2, whereby cut scores from Round 1 were reviewed, and cut scores for CSE 3 and CSE 5 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.

At the final debriefing on the last day, 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. 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 and Tannenbaum (2016); for a detailed discussion, see Hambleton et al. (2012).

3.3 Finalizing the Score Mapping

To finalize the mapping of the *TOEFL Junior* scores onto the CSE levels, three factors were considered. The first factor was the conversion from raw to scale scores. *TOEFL Junior* Standard test scores for each test section are reported on a 200–300 score scale, with 5-point increments. For the *TOEFL Junior* Speaking test, the reported score ranges from 0–16 with 1-point increments and is the sum of the ratings for each of four tasks, which are scored on a 0 to 4 scoring rubric. 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 then had to be converted to scale scores using the score conversion table 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 recommend 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 17.47 means that the cut score should be 18 to indicate that the minimum score is above 17.

- 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 of 17.47 means that the cut score should be 17, because the next highest score 18 was not recommended by the panel.

The second factor we considered for the finalization of the score mapping was a comparison between the CSE and the CEFR levels. *TOEFL Junior* scores were mapped onto both the CSE levels, based on the panelists' judgments, as well as the CEFR levels based on previous studies (presented on the *TOEFL Junior* website[3]). NEEA also conducted an empirical study to investigate the relationship between CEFR and CSE levels during the CSE development process (Liu & Peng, 2018). Comparing the levels of different language frameworks 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 the *TOEFL Junior* scores, the CSE, and the CEFR. The CEFR is regarded as an external criterion here, with the *TOEFL Junior* scores and the CSE levels being linked to the CEFR levels in separate studies.

The third factor we considered before finalizing the score mapping was input from members of the steering group and the working group of the project (for details see Papageorgiou et al., 2019, as these groups were the same for both studies). These experts in language education in China were asked to review the score mapping and comment on the reasonableness of the cut scores for each CSE level.

4 Results

4.1 Construct Congruence Results

The 32 CSE scales (eight per test section) that were the most relevant to the *TOEFL Junior* test content are listed in Table 1. Table 2 presents the number of descriptors for each task in the *TOEFL Junior* Standard and *TOEFL Junior* Speaking test form used in the standard setting meeting. Because a descriptor could have been selected for more than one task, Table 3 presents the number of unique descriptors selected for each of the four test sections (Reading, Listening, and LFM sections of *TOEFL Junior* Standard test and the *TOEFL Junior* Speaking test). A total of 121 descriptors were selected for all four test sections. It should be noted that to avoid unnecessary repetition in 2, all single-item listening sets were analyzed as one part of the listening section titled "Listening to classroom instructions". Moreover, the first task of the *TOEFL Junior* Speaking test (Read Aloud) presented some challenges in terms of content alignment to the CSE levels. Read-aloud tasks are common in the classroom; however, such tasks

[3] https://www.ets.org/toefl_junior/scoring_reporting/

Table 1
CSE Scales Selected for the Definition of Borderline Students

Listening	Reading
Overall listening comprehension	Overall reading comprehension
Self-assessment scale for listening comprehension	Self-assessment scale for reading comprehension
Understanding oral description	Understanding written description
Understanding oral narration	Understanding written narration
Understanding oral exposition	Understanding written exposition
Understanding oral instruction	Understanding written instruction
Understanding oral argumentation	Understanding written argumentation
Understanding oral interaction	Understanding written interaction
Speaking	Organizational competence
Overall oral expression	Overall organizational competence
Self-assessment scale for oral expression	Self-assessment scale for organizational competence
Oral description	Grammatical competence
Oral narration	Textual competence
Oral exposition	Vocabulary competence
Oral instruction	Syntactic competence
Oral argumentation	Competence of rhetorical or conversational organization
Oral interaction	Cohesion competence

lack the real-life, task-based approach typically adopted in the design of performance descriptors. Such tasks lack context of use, which increases the difficulty of mapping them to external performance levels and descriptors that are framed by context.

It should be noted that there are some caveats as to how the tables in this section should be interpreted. As we noted earlier, the description of what learners are expected to do at different framework levels is intentionally under-specified 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 only be interpreted as a justification of the decision to focus on the specific CSE levels during the subsequent standard setting meeting with the expert panel. Moreover, because only one test form of *TOEFL Junior* Standard and *TOEFL Junior* Speaking were used, the descriptor frequencies presented in this section might vary to another test form (although similarities would be expected, given that all test forms are based on the same blueprint). Finally, the descriptor frequencies should not be used as a tool to compare the content alignment of *TOEFL Junior* tests and another language test to the CSE levels.

4.2 Panel-Recommended Cut Scores for the Four Test Sections

The results of the panel's standard-setting judgments include the mean, the median, minimum, maximum and standard deviation (SD) by round of judgments for each of the test sections. 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, listening, and LFM sections of the *TOEFL Junior* Standard test, which was the metric that the panelists used. The cut scores for *TOEFL Junior* Speaking are provided on the 0-16 reported scale, as the panelists had access to that information during the standard setting process, and the judgment task involved listening to test-taker responses to the prompts. 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

Table 2
CSE Descriptors Aligned With the Content of TOEFL Junior Test Tasks

Task	CSE 2 descriptors	CSE 3 descriptors	CSE 4 descriptors	CSE 5 descriptors	CSE 6 descriptors	Total
Reading task 1	3	4	1	0	0	8
Reading task 2	1	5	3	3	0	12
Reading task 3	0	2	5	1	6	14
Reading task 4	0	2	4	3	0	9
Reading task 5	0	1	5	2	6	14
Listening classroom instructions	0	5	3	1	0	9
Listening short conversation 1	0	2	6	0	1	9
Listening short conversation 2	0	2	6	0	1	9
Listening short conversation 3	0	2	6	0	2	10
Listening academic lecture 1	0	3	3	5	2	13
Listening academic lecture 2	0	2	5	5	4	16
Listening academic lecture 3	0	3	2	5	4	14
LFM task 1	0	5	4	3	1	13
LFM task 2	0	5	5	3	1	14
LFM task 3	0	4	7	4	1	16
LFM task 4	0	5	8	4	0	17
LFM task 5	0	4	6	3	1	14
Speaking task 1	0	1	0	0	0	1
Speaking task 2	7	5	3	3	2	20
Speaking task 3	8	3	6	3	0	20
Speaking task 4	1	7	3	7	3	21

Note. LFM refers to the Language Form and Meaning test section.

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. To reduce the impact on misclassification rates (false positives and false negatives), Cohen et al. (1999) suggested that an SEJ should be no more than half the value of the standard error of measurement (SEM).

The results for the cut scores of the reading section of the *TOEFL Junior* Standard test are presented in Table 4. There was little variation in the mean cut score across the three rounds for CSE 2, CSE 4 and CSE 6. The variability in panelists' judgments decreased in general across rounds for these three levels, as can be seen by the SD, suggesting some convergence in the final round of judgments. Such convergence was particularly the case for judgments related to CSE 2, for which the SD decreased from 5.03 in Round 1 to 2.86 in Round 2 and 2.07 in Round 3. The SEJ for each cut score was within half of the raw score SEM of 2.38 for the reading section of this test form.

The results for the listening section are presented in Table 5. The mean cut score across the three rounds were similar for CSE 2, CSE 4, and CSE 6, with a slight decrease for CSE 4 and CSE 6, and a slight increase for CSE 2. The SD tended to decrease from Round 1 to Round 2, and Round

Table 3

Number of Unique CSE Descriptors Aligned With the Content of the TOEFL Junior Test Task

Task	CSE 2 descriptors	CSE 3 descriptors	CSE 4 descriptors	CSE 5 descriptors	CSE 6 descriptors	Total
Reading	3	9	11	6	7	36
Listening	0	7	11	6	5	29
LFM	0	6	9	4	1	20
Speaking	8	8	7	9	4	36
Total	11	30	38	25	17	121

Note. LFM refers to the Language Form and Meaning test section.

Table 4

Standard Setting Results for the Reading Section of the TOEFL Junior Standard Test

	Round 1			Round 2			Round 3				
	CSE 2	CSE 4	CSE 6	CSE 2	CSE 4	CSE 6	CSE 2	CSE 3	CSE 4	CSE 5	CSE 6
Mean	8.06	17.00	24.93	7.79	17.34	25.20	7.81	12.31	17.44	21.31	24.94
Median	8.10	17.15	25.25	7.90	17.30	25.45	8.00	12.00	17.00	21.00	25.00
Min.	0.70	10.30	17.90	2.50	13.30	20.10	4.00	6.00	13.00	18.00	23.00
Max.	18.50	24.40	29.60	11.70	21.80	29.00	11.00	15.00	21.00	25.00	29.00
SD	5.03	3.74	2.94	2.86	2.31	2.16	2.07	1.99	2.03	1.66	1.57
SEJ	1.26	0.93	0.73	0.72	0.58	0.54	0.52	0.50	0.51	0.42	0.39

Table 5

Standard Setting Results for the Listening Section of the TOEFL Junior Standard Test

	Round 1			Round 2			Round 3				
	CSE 2	CSE 4	CSE 6	CSE 2	CSE 4	CSE 6	CSE 2	CSE 3	CSE 4	CSE 5	CSE 6
Mean	6.18	17.43	26.25	6.03	17.58	26.46	6.25	11.94	17.00	21.56	25.88
Median	6.40	17.80	26.40	6.25	17.70	26.50	6.00	12.00	17.00	21.50	26.00
Min.	2.40	11.90	22.30	2.80	15.30	23.10	4.00	10.00	15.00	20.00	24.00
Max.	7.90	24.20	29.10	8.00	21.00	29.10	8.00	13.00	19.00	24.00	27.00
SD	1.53	2.55	1.67	1.36	1.42	1.42	1.00	0.85	0.97	1.15	0.96
SEJ	0.38	0.64	0.42	0.34	0.36	0.35	0.25	0.21	0.24	0.29	0.24

3, in particular for CSE 4 (from 2.55 in Round 1 to 0.97 in Round 3. The SEJ for each cut score was within half of the raw score SEM of 2.37 for the listening section of this test form.

The results for the LFM section are presented in Table 6. The mean cut score decreased slightly from Round 1 to Round 3 for CSE 2, CSE 4, and CSE 6. The SD tended to decrease, suggesting convergence in the judgments, with the exception of CSE 2, for which SD for Round 2 was somewhat lower than the SD for Round 3 (1.08 and 1.20 respectively). The SEJ for each cut score was within half of the raw score SEM of 2.43 for the LFM section of this test form.

The results for the *TOEFL Junior* Speaking test are presented in Table 7. The mean cut score across Round 1 and Round 2 was similar for CSE 2, CSE 4, and CSE

Table 6

Standard Setting Results for the LFM Section of the TOEFL Junior Standard Test

	Round 1			Round 2			Round 3				
	CSE 2	CSE 4	CSE 6	CSE 2	CSE 4	CSE 6	CSE 2	CSE 3	CSE 4	CSE 5	CSE 6
Mean	5.88	18.98	27.68	5.69	19.07	27.76	5.31	11.88	18.44	22.94	27.31
Median	5.60	18.25	28.00	5.65	18.80	28.05	5.00	12.00	18.00	22.50	27.00
Min.	3.20	15.40	24.70	3.40	16.50	25.10	3.00	10.00	16.00	20.00	26.00
Max.	11.60	26.70	30.00	8.30	24.00	30.00	9.00	15.00	23.00	26.00	30.00
SD	1.78	3.15	1.47	1.08	2.12	1.38	1.20	1.31	1.75	1.53	1.01
SEJ	0.45	0.79	0.37	0.27	0.53	0.35	0.30	0.33	0.44	0.38	0.25

Table 7

Standard Setting Results for the TOEFL Junior Speaking Test

	Round 1			Round 2				
	CSE 2	CSE 4	CSE 6	CSE 2	CSE 3	CSE 4	CSE 5	CSE 6
Mean	6.06	10.19	14.19	6.00	8.00	10.00	12.00	14.06
Median	6.00	10.00	14.00	6.00	8.00	10.00	12.00	14.00
Min.	4.00	10.00	11.00	6.00	8.00	10.00	12.00	14.00
Max.	8.00	11.00	15.00	6.00	8.00	10.00	12.00	15.00
SD	1.00	0.40	0.54	0.00	0.00	0.00	0.00	0.25
SEJ	0.25	0.10	0.14	0.00	0.00	0.00	0.00	0.06

6. In Round 2, the panelists selected the same cut score for each CSE level, except for the cut score for CSE 6, for which one panelist recommended 15, when all other panelists recommended 14. As a result of this agreement, the SEJ for four cut scores was 0. The SEJ for CSE 6 was within half of the raw score SEM of 1.4 for this test form of the *TOEFL Junior* Speaking test.

Table 8 summarizes the panel's feedback in the evaluation survey regarding the general process followed in the standard setting meeting. The majority of the 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; 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 selected the option strongly disagree for any statement. Panelists were also asked to indicate their level of comfort with the final cut score recommendations (Table 9). There was one panelist that selected "very uncomfortable". The panelist wrote in the survey: "Probably next time some activities can be designed to encourage participants to express their opinions about why they have disagreement on the cutting scores". This panelist might have felt that she did not have enough opportunities for discussion of the proposed cut scores, even though there was always some whole-panel discussion between rounds of judgments. It is interesting to note that this panelist's feedback on the standard setting process was positive, including the item "The opportunity for feedback and discussion between rounds was helpful" (Table 8). Irrespective of the reason for low confidence ratings, the fact that the panelist had doubts about the recommended cut scores warrants exclusion of her judgments from the calculation of the panel's final recommendation presented in the next sections. Overall, there was no meaningful change in the panel's recommended cut score after excluding this panelist's judgments (the cut score varied from 0 to 0.12 raw score points), and the group's overall confidence was satisfactory.

Table 8
Panelists' Feedback on the Standard Setting Process

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

Table 9
Panelists' Reported Confidence in the Recommended Cut Scores

Test section	Very uncomfortable	Somewhat uncomfortable	Somewhat comfortable	Very comfortable
Listening	1			15
Reading	1		1	14
LFM	1		1	14
Speaking	1			15

Note. LFM refers to the Language Form and Meaning test section.

4.3 Final Score Mapping

Table 10 provides the conversion from raw to scale scores, using both rounding approaches described in the Method section. The panelist indicating low confidence in the cut score recommendation was excluded from the calculation of the recommended cut scores, as discussed earlier. Rounding for the *TOEFL Junior* Speaking test is only relevant for the CSE 6 cut scores, as the panelists recommended the same cut score for each of the other levels.

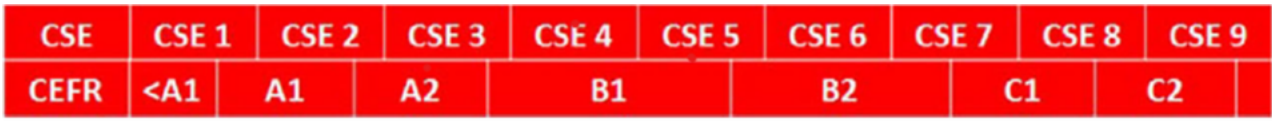
The link between the levels of the CEFR and CSE based on the results of the NEEA study is presented in Figure 1 and can be summarized as follows:

- 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.
- CSE 3 covers most of CEFR Level A2 and the lower CEFR Level B1.
- CSE 2 covers most of CEFR Level A1 and the lower CEFR Level A2.

The mapping of the *TOEFL Junior* test scores to the CEFR levels, is presented in Table 11.

Upon consideration of the two rounding rules and the CEFR mapping information, as well as feedback from members of the steering and working groups, we finalized the CSE mapping as shown in Table 12.

Figure 1
Link Between the Proficiency Levels of CSE and the CEFR (Liu & Peng, 2018; reproduced from Papageorgiou et al., 2019)



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

Table 10
Panel-Recommended Raw and Scale Cut Scores for the TOEFL Junior Tests

	Panel-recommended cut scores (raw)					Cutscores converted on the reporting scale using two rounding approaches				
Test section	CSE 2	CSE 3	CSE 4	CSE 5	CSE 6	CSE 2	CSE 3	CSE 4	CSE 5	CSE 6
Reading	7.93	12.33	17.47	21.27	24.87	205	225	245/250	260/265	270/275
Listening	6.27	11.87	16.93	21.53	25.87	200	220	240/245	260/265	280/285
LFM	5.33	11.87	18.47	22.93	27.33	200	225/230	255/260	270/275	290/295
Speaking	6	8	10	12	14.07	6	8	10	12	14/15

Note. When two scale scores are shown, the lower value is the results of the "round down" rule, whereas the highest values are the result of the "round up" rule. In some cases, rounding up or rounding down did not make a difference. LFM refers to the Language Form and Meaning test section.

Table 11
Score Mapping of the TOEFL Junior Tests to the CEFR Levels

Test section	CEFR A2	CEFR B1	CEFR B2
Reading	210−240	245−275	280−300
Listening	225−245	250-285	290−300
LFM	210−245	250-275	280−300
Speaking	8−10	11−13	14−16

Note. LFM refers to the Language Form and Meaning test section.

The final score mapping is based on the following rationale for each level:

- **CSE 2.** We chose to remove this level from the proposed mapping for all test sections. Cut scores at the bottom of the score scale are typically of little use to test takers and score users, as practically anyone taking the test can obtain the lowest scale score. A cut score at the bottom of the scale also suggests a lack of construct congruence between the test and the standards, in this case the CSE. As can be seen in Table 10, the panelists recommended the lowest score (200) for two out of three sections of *TOEFL Junior* Standard. Also, the content analysis in the project report identified only a few CSE descriptors that were relevant to the content of *TOEFL Junior*. Finally, the lowest CEFR level for *TOEFL Junior* includes A2, which, as shown in Figure 1, is primarily linked to CSE 3, rather than CSE 2.

- **CSE 3.** We treated this level as generally similar to CEFR A2, but we also took into account that cut scores for CSE 3 might need to be somewhat higher than the corresponding cut score for CEFR A2, based on Figure 1. For both reading and listening we chose 220. The cut scores for CEFR A2 are 210 and 225 respectively. For LFM we chose 225, the lower panel-based recommendation, which is closer to the cut score for CEFR A2 (210). For Speaking, the panel-based recommendation was used, which is identical to the cut score for CEFR A2 (score of 8).

- **CSE 4.** Figure 1 shows that this level is approximately equivalent to CEFR B1. For both Reading and Listening we opted for the highest option, depending on the rounding rule, of the panel-based recommendations (250 and 245 respectively). For LFM we opted for the lower cut scores (255). By choosing these cut scores for the three sections of *TOEFL Junior* Standard, cut scores for CSE 4 and CEFR B1 are either identical or adjacent (difference of one 5-point increment). For Speaking, the panel-based recommendation was used, which was also adjacent to the cut score for CEFR B1 (difference of one-point increment).

- **CSE 5.** Figure 1 suggests that CSE 5 mostly covers the upper half of CEFR B1 and the lower half of CEFR B2 and might need to be higher than the corresponding cut score for CEFR B1. For both reading and listening we chose 265 (the highest recommendation) and 270 for LFM (the lowest), which are approximately halfway between the cut scores for CSE 4 and the cut score for CEFR B2. For Speaking, the panel-based recommendation was used (score of 12).

- **CSE 6.** Upon recommendation by members of the working and steering groups, we removed CSE 6 from the final score mapping. Group members noted that CSE 6 mainly captures the language use activities typical in tertiary education. Given that the target population of the *TOEFL Junior* tests is younger, linking the *TOEFL Junior* test scores to CSE 6 might result in misinterpretation of test scores.

5 Discussion and Conclusion

In this paper, we provided a detailed rationale behind the mapping of the *TOEFL Junior* test scores onto the CSE levels, building on several sources of data to support the final score mapping. Although the different sources of data support the score mapping we presented, policymakers in the educational context where the CSE levels are used might want to further investigate the relationship between *TOEFL Junior* test scores and the CSE levels. Such exploration should focus on the relevance and usefulness of the score mapping to facilitate score-based decision-making in the context of assessing the language proficiency of young learners. As we discussed in the Introduction section, the upper limit of the score mapping was unclear at the outset. Ultimately, we decided to remove CSE Level 6 from the final score mapping, a decision which reaffirms that this level is more applicable to learners at a higher educational level and also reaffirms the importance of attending to contextual factors when designing, implementing, and interpreting results from these types of alignment studies. This may be especially true when the outcomes of such studies will inform and guide educational policy (Wu, 2019).

We believe that because of the focus on young learners, our study makes a useful contribution to the growing literature on the mapping of test scores onto language proficiency levels. Similar to previous research (Dunlea et al., 2019; Papageorgiou et al., 2019), our study demonstrates how evidence should be collected to support a claim about the alignment of test scores to the proficiency levels of the CSE. In addition, our research underscores the importance of considering contextual factors and how test scores might be interpreted or misinterpreted in a specific educational context, as a result of the score mapping. In the case of our study, we decided to remove CSE 6 from the final mapping not because the panelists were not able to set a cut score for that level, but because we were concerned about the mismatch between the intended population for that CSE level and the population of young learners targeted by the test. The post-standard setting adjustments to the recommended cut scores are consistent with good practice (Geisinger & McCormick, 2010), and reinforces the widely accepted view that standard setting is closely related to policy formation (Kane & Tannenbaum, 2013). We also believe that this study is a useful addition to the literature in the sense that it stressed the importance of multiple perspectives to ensure responsible and meaningful interpretation of the score mapping.

The practical implications of our study are twofold. First, we demonstrated an operational process for future test alignment studies, especially on tests or populations that are less studied. Second, our study might facilitate comparison and discussion on young learners' language proficiency across countries, given that the *TOEFL Junior* tests are delivered in several countries and their scores are used to make placement and progress monitoring decisions

Table 12
Score Mapping of the TOEFL Junior Tests to the CSE Levels

Test section	CSE 3	CSE 4	CSE 5
Reading	220−245	250−260	265−300
Listening	220−240	245−260	265−300
LFM	225−250	255−265	270−300
Speaking	8−9	10−11	12−16

Note. LFM refers to the Language Form and Meaning test section.

in secondary school contexts.

We conclude this paper by emphasizing the need for language testers to be cautious about potential issues in the context of aligning test scores to the CSE levels. As other have noted (Council of Europe, 2009; Papageorgiou et al., 2019; Tannenbaum & Cho, 2014), policymakers should not consider alignment to be sufficient evidence of the quality of a language test or sufficient support for score interpretation and use. Also, different 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. It is also possible that as additional data are collected, revisions to the alignment of test scores to the proficiency levels might be required, based on improved understanding of how the language test and the language framework operationalize the underlying language ability construct (see discussion in Papageorgiou et al., 2015).

References

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

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

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

Beaton, A., & Allen, N. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics*, *17*, 191–204.

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

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

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

Council of Europe. (2009). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR). A manual.* https://www.coe.int/en/web/common-european-framework-reference-languages/relating-examinations-to-the-cefr

Dunlea, J., Spiby, R., S., W., Zhang, J., & Cheng, M. M. (2019). *China's Standards of English Language Ability (CSE): Linking UK exams to the CSE* (Research Report No. VS/2019/0003). https://www.britishcouncil.org/sites/default/files/linking_cse_to_uk_exams_5_0.pdf

Educational Testing Service. (2019). *TOEFL Junior® framework and test development* (TOEFL® Research Insight Series, Volume 7). https://www.ets.org/s/toefl/pdf/toefl_test_framework_and_development.pdf

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

Geisinger, K. F., & McCormick, C. M. (2010). Adopting cut scores: Post-standard-setting panel considerations for decision makers. *Educational Measurement: Issues and Practice*, *29*(1), 38–44.

Gu, L., Lockwood, J., & Powers, D. (2015). *Evaluating the TOEFL Junior® Standard test as a measure of progress for young English language learners* (ETS Research Report RR-15-22). Educational Testing Service.

Haberman, S. J., Sinharay, S., & Lee, Y.-H. (2011). *Statistical procedures to evaluate quality of scale anchoring* (ETS Research Report RR-11-02). Educational Testing Service.

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

Hambleton, R. K., Pitoniak, M. J., & 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* (pp. 47–76). Routledge.

Kane, M. (2012). Validating score interpretations and uses. *Language Testing*, *29*(1), 3–17.

Kane, M., & Tannenbaum, R. (2013). The role of construct maps in standard setting. *Measurement: Interdisciplinary Research & Perspectives*, *11*, 177–180.

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* (pp. 201–224). Routledge.

Liu, J., & Peng, C. (2018). *Aligning CSE with CEFR* [Paper presentation]. The 4th International Conference on Language Testing and Assessment, Beijing, China.

Liu, J., & Wu, S. (2019). *Research on China's Standards of English Language Ability*. Higher Education Press.

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

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.

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

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). Educational Testing Service.

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). Educational Testing Service. https://doi.org/10.1002/ets2.12281

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* (pp. 181–199). 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.

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). Erlbaum.

So, Y., Wolf, M. K., Hauck, M. C., Mollaun, P., Rybinski, P., Tumposky, D., & Wang, L. (2015). *TOEFL Junior® design framework* (Research Report No. RR-15-13). Educational Testing Service. https://doi.org/10.1002/ets2.12058

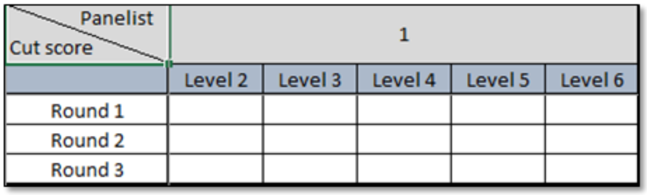
Tannenbaum, R. J., & Cho, Y. (2014). Criteria for evaluating standard-setting approaches to map english language test scores to frameworks of english language proficiency. *Language Assessment Quarterly*, *11*(3), 233–249.

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

Appendix A

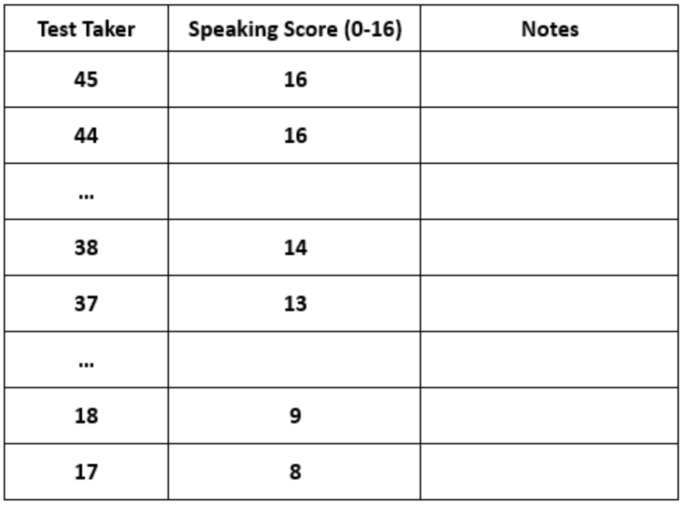
Sample Panelist Rating Form for Reading (Round 3)



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

Appendix B

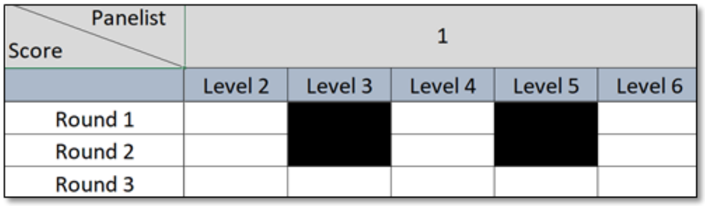
Sample of Test Taker Scores Used for Setting Speaking Cut Scores



Test Taker	Speaking Score (0-16)	Notes
45	16	
44	16	
...		
38	14	
37	13	
...		
18	9	
17	8	

Appendix C

Sample Panelist Rating Form for Speaking (All Rounds)



Panelist Score	1				
	Level 2	Level 3	Level 4	Level 5	Level 6
Round 1		■		■	
Round 2		■		■	
Round 3					