Research Memorandum
ETS RM–24-04

Mapping the TOEFL ITP® Speaking Scores to the Levels of the Common European Framework of Reference

Michael Suhan
Spiros Papageorgiou
Larry Davis
Molly Palmer

February 2024

ETS Research Memorandum Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Memorandum series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Memorandum series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

Mapping the TOEFL ITP® Speaking Scores to the Levels of the Common European Framework of Reference

Michael Suhan, Spiros Papageorgiou, Larry Davis, and Molly Palmer
ETS, Princeton, New Jersey, United States

February 2024

Corresponding author: M. Suhan, E-mail: msuhan@ets.org

Suggested citation: Suhan, M., Papageorgiou, S., Davis, L., & Palmer, M. (2024). *Mapping the TOEFL ITP® Speaking scores to the levels of the Common European Framework of Reference* (Research Memorandum No. RM-24-04). ETS.

Find other ETS-published reports by searching the
ETS ReSEARCHER database.

To obtain a copy of an ETS research report, please visit
https://www.ets.org/contact/additional/research.html

Action Editor: Heather Buzick

Reviewers: Kathryn Hille and Rick Tannenbaum

Abstract

In this research memorandum, we report on a study to map the scores of the TOEFL ITP® Speaking test to the language proficiency levels of the Common European Framework of Reference (CEFR). We conducted our study prior to the operational launch of the test, which was developed as an optional component of the digitally delivered version of the TOEFL ITP test. The TOEFL ITP Speaking test measures English speaking abilities in everyday situations and in academic contexts. It is intended for use by educational institutions for purposes such as placement, monitoring student progress, and exiting language programs. Mapping TOEFL ITP Speaking test scores to the CEFR levels allows stakeholders to interpret test results in reference to a widely used language framework, providing additional support for score interpretation. The score mapping process involved establishing recommended minimum test scores (cut scores), informed by the judgments of expert panelists, to classify test takers into the CEFR levels.

Keywords: TOEFL ITP®, CEFR, standard setting, speaking assessment, English as a foreign language, score mapping

Mapping (linking or aligning) test scores to national or international proficiency levels is a common practice that facilitates the interpretations and intended use of test scores by score users (Papageorgiou, 2016). The proficiency levels of the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001) and its companion volume (Council of Europe, 2020) are perhaps the most widely used ones, not only by the Council of Europe and its member states, but also worldwide. Following the widespread adoption of the CEFR, a manual was published to provide guidance to test developers on linking test scores with the CEFR proficiency levels (Council of Europe, 2009). Given its extensive application to teaching, learning, and assessment following its development, there is presently an expectation for language test scores to be linked to the CEFR proficiency levels (Deygers et al., 2018).

Following the paradigmatic shift toward communicative approaches in language teaching, several publications by the Council of Europe (e.g., van Ek & Trim, 1991, 1998, 2001) sought to describe language proficiency levels in terms of language activities and competences. These publications then provided the impetus for further research, aimed at linking such descriptions of language ability to proficiency levels on a common scale (North, 2000), leading to the publication of the CEFR (Council of Europe, 2001). The CEFR describes language proficiency in terms of six main levels (from the lowest, A1, to A2, B1, B2, C1, and C2, the highest). The CEFR provides dozens of different scales describing language proficiency not only in terms of the social context it in which it occurs, but also in terms of the competences associated with respective levels. The illustrative descriptors of the CEFR scales describe learners' performance and abilities in a positive way, expressing what learners can—rather than cannot—do. Certain scales have been supplemented with additional "plus" levels (i.e., A2+. B1+, B2+), providing more detailed descriptors for the thresholds between the three main levels. The companion volume (Council of Europe, 2020) similarly provides descriptors for a pre-A1 level on select scales.

Recently, a new test was added to the TOEFL® family of assessments: The TOEFL ITP® Speaking test (https://www.ets.org/toefl/itp). This research memorandum reports on an effort to map the scores of this new test to the CEFR levels using recognized standard setting procedures.

The TOEFL ITP Speaking Test

The TOEFL ITP Speaking test measures the ability to speak English in everyday situations and in basic campus and academic contexts. It is used by educational institutions where assessment of oral skills is required for a variety of purposes, including placement, monitoring student progress, and exiting language programs. The TOEFL ITP Speaking test is used as an optional addition to the digitally delivered TOEFL ITP Level 1 or Level 2 tests, which include three sections: Listening, Structure and Written Expression, and Reading. Oral skills are measured using three task types: (a) Read Aloud, (b) Independent Speaking, and (c) Integrated Speaking. All three task types are scored automatically, based on scoring rubrics developed as part of a field test conducted prior to launch in January 2022. Scores are reported on a scale of 31–68, in 1-point increments. TOEFL ITP Speaking test scores are reported separately from TOEFL ITP Level 1 and Level 2 test scores. Table 1 provides an overview of the task types included on the test, which simulate tasks in everyday and academic contexts. A sample TOEFL ITP Speaking test is available at https://www.ets.org/toefl/itp/prepare.

Table 1. Task Types on the TOEFL ITP Speaking Test

Task Type	Number of Tasks
Read Aloud	1
Independent Speaking	2
Integrated Speaking	1

Methodology for the Standard Setting Study

General Procedures for Setting Cut Scores

When mapping test scores onto external proficiency levels, such as the CEFR ones, a decision must be made regarding the range of test scores that corresponds to each level. This decision-making process, known as *standard setting* (Cizek & Bunch, 2007), entails establishing a minimum score (cut score) that distinguishes each level from the adjacent level below it. To facilitate standard setting, it is common practice to assemble a panel of experts tasked with recommending cut scores based on the test content and the knowledge and skills of test takers at the targeted proficiency level. Study facilitators provide the panelists with statistical information about the test (e.g., item difficulty estimates, distributions of test scores) to inform

their decisions. A standard setting meeting most often consists of two or three rounds so that the panelists may make preliminary cut score recommendations, consider statistical information, and then discuss and consider the implications of their individually recommended cut scores before choosing to keep or revise them. While considering the recommended cut scores of the panel as the primary determining factor, the test provider or score user then makes the final decision regarding recommended cut score use (i.e., accepting, raising, or lowering the cut scores).

Overview of the Score Mapping Process

This section provides an overview of the standard setting study, which was conducted from August 3 to August 6, 2021, shortly after the administration of the field test of the TOEFL ITP Speaking test (see Appendix A for the meeting schedule). The project team comprised a senior assessment developer and two senior research scientists. Seventeen panelists were recruited to participate in the study: 10 panelists were ETS scoring leaders, and seven were ETS assessment developers. To prepare for the standard setting meetings, panelists were required to complete a set of activities intended to familiarize them with the TOEFL ITP Speaking test, the CEFR levels, and the standard setting method. All meetings were conducted remotely through Microsoft Teams and facilitated by the project team following recommended standard setting methodology. Upon the conclusion of the final meeting, cut scores were recommended by the panel for the purposes of classifying TOEFL ITP Speaking test takers in CEFR levels A2 to C1. These CEFR levels are the target levels for the other TOEFL ITP tests, as well as the TOEFL iBT® Speaking test, which has content relevant to the TOEFL ITP Speaking test. Therefore, CEFR levels A2 to C1 were the goal for this study. The panel-recommended cut scores were then evaluated, and the final mapping of the TOEFL ITP Speaking test scores to the CEFR levels was produced.

Selection of Panelists

Seventeen panelists in total were recruited internally from ETS for the study. Seven panelists were assessment developers who had prior experience working on both speaking assessments and the TOEFL family of assessments. Ten panelists were scoring leaders who had

previously worked with the TOEFL iBT Speaking test. As part of their regular work for ETS, scoring leaders provide support and assistance to raters during operational scoring sessions.

Similar to the reasons provided by Davis et al. (2023), the panel was of appropriate size and expertise but not necessarily representative of the broad range of stakeholders who would be ideally included in such a panel. Although recruiting internal staff from the test provider as panelists could lead to "insider bias" (Papageorgiou, 2010), it was considered a justifiable measure as the content of the preoperational test had not yet been released to the public. Furthermore, the study had to be completed online because of the COVID-19 pandemic, so including only internal staff on the panel meant that all panelists would already be experienced using ETS's remote platforms and that most panelists would have prior experience participating in an online standard setting study, as nine of the 17 panelists had been included in the score-mapping study conducted by Davis et al. (2023). This arrangement was preferable, as it likely reduced the cognitive load of the standard setting activities and eliminated concerns about the feasibility of training external panelists to participate in the study remotely.

Panelists' Preparation Prior to the Study

The panelists received a preparation guide before the first standard setting meeting. The content of the guide included information about the standard setting study, the CEFR, and the TOEFL ITP Speaking test. The guide also included two familiarization activities for the panelists to complete prior to the first meeting, intended to orient them with the CEFR scales relevant to the assessment, including (a) overall oral production, (b) overall spoken interaction, (c) overall phonological control, and (d) overall general linguistic range.

In the first activity, panelists were asked to first review all levels of the CEFR scales. They were then provided with 24 representative descriptors from different CEFR scales that had been reordered so that they were no longer presented in order of difficulty. Based on their knowledge of the provided overall scales, panelists were asked to put the descriptors into three groups sorted by difficulty: advanced, intermediate, and elementary. This first stage of the activity was intended to provide support for the second stage, in which panelists then put the same descriptors into a table sorted by CEFR levels A1 to C2. Finally, panelists completed an activity in which they were asked to list three to five distinguishing speaking features that

separate each CEFR level from adjacent levels (e.g., the features of CEFR level B2, which distinguish that level from CEFR level B1 and CEFR level C1). The materials used for the preparation activities are provided in Appendix B.

To familiarize panelists with the TOEFL ITP Speaking test, detailed descriptions of each task type were included. Screenshots were included in the guide, showing the tasks as they were experienced by test takers on the online testing platform during the administration of the field test. For the Integrated Speaking task, a transcript of the audio used in the stimulus was included. The guide also included a link to a SharePoint site where panelists could access the scoring rubrics used for each task as well as video demonstrations of each task on the testing platform. Time was allotted during the first standard setting meeting to show panelists the video demonstrations and scoring rubrics and to answer their questions about the test. To ensure that panelists understood standard setting procedures, the study facilitators provided a demonstration of the steps to be followed when making cut score recommendations, posed comprehension check questions about the procedures to the panelists, and fielded their questions.

Discussion of Familiarization Activities and Definition of the Just Qualified Candidate

During the first meeting, panelists reviewed the familiarization activities as the first step toward making cut score recommendations and were provided the opportunity to discuss the challenges they faced in associating descriptors with CEFR levels. This activity, along with the second familiarization activity in which panelists listed the distinguishing speaking features for each CEFR level, was used to facilitate a discussion aimed at reaching a consensus among panelists on what speaking skills were characteristic of a just qualified candidate (JQC) for CEFR levels A2, B1, B2, and C1. The JQC is defined as a test taker possessing the minimally acceptable skills, as identified by the panel, for a respective level. The discussion was guided by the study coordinators, who prompted panelists to specify the speaking skills characteristic of a JQC for each level and summarized the panelists' responses on screen during the discussion, allowing time for clarification and debate. The key distinguishing features of CEFR levels A1 to C2 specified by the panelists, which were used to help define the JQC, are listed in Appendix C.

Although panelists were asked to describe C2 level features, they were informed that the TOEFL ITP Speaking test does not attempt to provide classifications at the C2 level.

Standard Setting Method

A variation of the performance profile method (Fleckenstein et al., 2020) was used in this standard setting study, allowing panelists to review a set of performance samples from test takers' responses to the tasks on the test. The review process drew on the panelists' professional expertise as assessment developers and score leaders, as it required making holistic judgments on responses given by test takers (Kingston & Tiemann, 2011).

The responses reviewed by panelists were selected from the TOEFL ITP Speaking field test. A total of 34 test takers were selected from 859 participants who had fully completed the field test, providing a response to all four tasks across the three task types. Individuals were selected to represent raw test scores ranging from 2.5–19, in 1-point increments from 3–19, with two individuals selected per score point greater than 3. The profiles, comprising all responses given by each test taker, included one response to the Read Aloud task, two responses to the Independent Speaking task, and one response to the Integrated Speaking task.

Cut scores were set through two rounds of judgments, with feedback and discussion between rounds. At the beginning of Round 1, panelists reviewed the key distinguishing features for each CEFR level and were asked to judge which test taker(s) best exemplified the characteristics of the JQC at a given CEFR level. The overall score earned by this individual was then taken as the raw cut score for that CEFR level. Panelists independently reviewed the test response profiles and entered their judgments on a rating form, which they then submitted to the project team at the conclusion of Round 1 (see the sample rating form in Appendix D).

For Round 2, panelists were provided with descriptive statistics on their Round 1 judgments (i.e., the mean, median, mode, minimum, and maximum of cut scores for each CEFR level). They were then shown what percentage of test takers from the field test would be categorized into each CEFR level if the recommended cut scores from the first round were used. Panelists then had the opportunity to explain the rationale for their recommendations to the group and to discuss and listen to specific profiles in relation to the JQC descriptions. Following the discussion of the Round 1 judgments, panelists were asked to review the complete set of

test-taker profiles once again to make a final judgment about what scores the JQC for each CEFR level would receive.

Results of the Standard Setting Study

This section summarizes the two rounds of the panel's standard setting judgments. The mean, median, mode, minimum, maximum, and standard deviation (*SD*) of Round 1 and Round 2 judgments are reported in Table 2. The mean cut scores in Round 2 represent the panel's recommended cut scores. Scores are presented as raw totals on a scale from 0–19, the same scale presented to the panel. To show the level of uncertainty of the panel, the standard error of judgment (SEJ) is also reported. SEJ, which is the standard deviation of judgments divided by the square root of the number of panelists (Cizek & Bunch, 2007), provides an indication of how close a cut score recommended by a panel of similar experts is likely to be to that of the current panel. Providing that both panels underwent similar training on standard setting methods, both panels would be expected to set recommended cut scores within 1 SEJ of each other about 68% of the time and within 2 SEJ about 95% of the time. To reduce the impact of classification errors (i.e., false positive and false negative misclassifications), the variability of the panelists' judgments should not exceed the measurement error of the test itself. Cohen et al. (1999) suggests that SEJ within half of the standard error of measurement (SEM) of the test meets this standard.

Between the two rounds, mean cut scores decreased by no more than 0.5, with no change in the mean C1 cut score between rounds. The only median cut score to change was the C1 level, which along with the C1 minimum cut score, increased by 1 in Round 2. Decreased standard deviations were observed in Round 2, with between-round differences ranging from 0.5–0.9 *SD*. As the SEM for the TOEFL ITP Speaking test form used in the standard setting study was 0.98, the SEJ for every cut score across rounds was within acceptable bounds of half the SEM (i.e., no more than 0.49).

Table 2. Standard Setting Results for the TOEFL ITP Speaking Test

	Round 1				Round 2			
	CEFR A2	CEFR B1	CEFR B2	CEFR C1	CEFR A2	CEFR B1	CEFR B2	CEFR C1
n	17	17	17	17	17	17	17	17
Mean	7.0	10.1	13.4	16.6	6.8	9.6	13.0	16.6
Maximum	11	15	17	18	8	10	14	17
Minimum	2.5	8	12	15	5	8	12	16
Median	7	10	13	16	7	10	13	17
Mode	7	10	13	16	7	10	13	17
*S*D	1.8	1.6	1.3	1.0	0.9	0.7	0.5	0.5
SEJ	0.44	0.39	0.32	0.24	0.22	0.17	0.12	0.12

Note. CEFR = Common European Framework of Reference; SD = standard deviation; SEJ = standard error of judgment.

TOEFL ITP Speaking test scores are reported on a scale of 31–68 in increments of 1. To simplify the standard setting process, panelists recommended cut scores based on raw scores on a scale of 0–19 without considering the conversion of raw scores to scale scores. For raw cut scores to be converted to the TOEFL ITP Speaking test scale, the panel-recommended cut scores must be a multiple of 0.5. As the mean is used to determine the panel-recommended cut score, rounding is at times necessary. There are two options for rounding raw cut scores:

- The raw score is rounded up to the next multiple of 0.5. This follows the rationale that a cut score that is between two achievable score points indicates ability beyond the lower of the two score points. In the case of a raw score of 6.8, because the minimum score is above 6.5, the cut score would be set as 7.0.
- The raw score is rounded down to the previous multiple of 0.5. This follows the rationale that even though there may be evidence of ability beyond the lower of the two score points, the higher score point has not been achieved. In the case of a raw cut score of 6.8, because the minimum score is less than the next achievable score point of 7.0, the cut score would be rounded down to 6.5, as a cut score of 7.0 was not recommended by the panel.

Table 3 provides the results of the two approaches to rounding raw cut scores. When considering which approach to use, it is important to consider how rounding may impact false negative and false positive classifications (see discussion in Papageorgiou et al., 2015). For

example, rounding the cut score up represents a relatively conservative approach to decision making in that there will be greater confidence that the test taker actually satisfies the requirements for placement into the higher level.

Table 3. Two Approaches to Rounding TOEFL ITP Speaking Test Cut Scores

Cut score	CEFR A2	CEFR B1	CEFR B2	CEFR C1
Panel-recommended cut scores	6.8	9.6	13.0	16.6
Cut scores rounded down	6.5	9.5	13.0	16.5
Cut scores rounded up	7.0	10.0	13.0	17.0

Note. CEFR = Common European Framework of Reference. Rounding did not make a difference in the case of the CEFR B2 level, where the panel-recommended cut score was a round number.

Final Score Mapping

Table 4 summarizes the recommended mapping of TOEFL ITP Speaking test scores onto the CEFR levels. Based on the panel-recommended cut scores, the following decisions were made:

- A cut score of 14 for CEFR level B2 was selected by the study facilitators, which is 1 point above the panel-recommended cut score. This was deemed necessary because a cut score of 13 would allow for a scenario in which a test taker may be classified as B2 with task scores of 3 (on a scale of 0–5) on the two Independent Speaking tasks and the Integrated Speaking task. Such individual task scores are unlikely to collectively represent the performance expected at the B2 level with a score of 13, because a score of 3 on the Independent Speaking task and Integrated Speaking task rubrics primarily comprises descriptors addressing abilities below the B2 JQC definition. However, a scenario of such individual task scores was not present in the test-taker sample responses provided to the panelists. Although other higher individual task scores could result in a total raw score of 13, thus collectively demonstrating B2 performance, the study facilitators decided that a B2 classification with so many individual task scores of 3 was not acceptable for a level commonly used as a standard in high-stakes decision making (e.g., university admissions).

- Cut scores for CEFR levels A2, B1, and C1 were rounded up to the next achievable score point, further reducing the likelihood of false positive classifications. This rounding decision was made in light of two additional factors from the standard setting process and analysis of the results. First, aside from the profile of the test taker receiving the lowest score (2.5), panelists reviewed responses by test takers who only received whole score points. Second, the recommended cut scores for CEFR levels A2, B1, and C1 in Table 4 represent the majority of the panelists' judgments, as indicated by the mode in Table 2.

Table 4. Recommended Mapping of TOEFL ITP Speaking Test Cut Scores onto the CEFR Levels

CEFR level	Cut scores (raw score scale)	Cut scores (reported score scale)
C1	17	64
B2	14	58
B1	10	48
A2	7	41

Note. CEFR = Common European Framework of Reference.

Discussion and Conclusion

This standard setting study was conducted to recommend cut scores for the TOEFL ITP Speaking test that correspond to the A2, B1, B2, and C1 levels of the CEFR. Although the panelists were responsible for recommending cut scores, it is the responsibility of policymakers to make the final determination on cut score use (Kane, 2002). As part of the TOEFL ITP assessment series, the TOEFL ITP Speaking test is intended for institutional use (e.g., by colleges and universities) for purposes such as placement, monitoring student progress, and exiting language programs. Because institutional needs may vary depending on contextual factors, it is therefore not possible that the panel-recommended cut scores represent the optimal final cut score for every institution. To set final cut scores, policymakers are consequently obligated to consider other pieces of evidence (Geisinger & McCormick, 2010), accepting or adjusting the recommended cut scores as necessary (i.e., by raising or lowering). When cut scores are adjusted, the extent that classification errors will impact score use should be considered. Lowering a cut score minimizes the risks of a false negative classification while increasing the odds of a false positive classification. When a false positive classification occurs (i.e., when a

test taker is classified into a level above their actual level) the abilities represented by the performance of a test taker are overestimated. Raising the cut score will minimize the likelihood of false positive classifications at the expense of an increased likelihood of false negative classifications. A false negative classification occurs when a test taker possesses the abilities being tested but is assigned to a lower level. In the case of a false positive, the test taker may be placed into a situation for which they are not ready, whereas in the case of a false negative, test takers may be denied opportunities for which they are actually qualified. To make the decisions that best represent their institutional needs, policymakers should evaluate the impact of false positive and false negative classifications and set final cut scores accordingly.

We note that the justification for a standard setting study and the validity of the resulting cut scores rests on an assumption of sufficient content alignment (or construct congruence) between the assessment and the relevant language framework (Tannenbaum & Cho, 2014). Alignment of test content to the CEFR poses a particular challenge in that the CEFR descriptors are designed to be underspecified to allow for application in a variety of contexts. Because of time constraints, conducting a detailed construct congruence analysis was not possible for the current study. However, the content relevance between the speaking scoring rubrics of the TOEFL iBT test and the CEFR levels, described in Papageorgiou et al. (2015), suggests reasonable construct congruence between the TOEFL ITP Speaking test and the CEFR levels because of the shared approach to task design between the TOEFL iBT and TOEFL ITP tests (e.g., use of similar independent and integrated speaking tasks). In addition, language from relevant CEFR speaking descriptors was consulted during the development of the scoring rubric for the field test (https://www.ets.org/pdfs/toefl/toefl-itp-speaking-descriptors.pdf). A further limitation of this study was the lack of a poststudy panelist evaluation survey, because of logistical issues.

We also note that at the time of this study the TOEFL ITP Speaking test had recently been field tested. The results of this study are based on the field test population. In the future, following several administrations of the TOEFL ITP Speaking test, it would be desirable to review the proposed CEFR score mapping with reference to individuals taking the operational test, as well as other tests in the TOEFL Family, in particular the TOEFL iBT test, given the

content similarities between the two speaking tests. However, the field test sample was crafted to be representative of the expected population of test takers, which allowed us to set cut scores with reasonable confidence that the results will be valid for the operational test.

References

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

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press. https://assets.cambridge.org/052180/3136/sample/0521803136ws.pdf

Council of Europe. (2009). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR). A manual.* http://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680667a2d

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment. Companion volume.* https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4&usg=AOvVaw3GG5_eUIXiPP8OTr2H0CHy&opi=89978449

Davis, L., Garcia Gomez, P., Li, S., & Manna, V. F. (2023). Mapping TOEFL® Essentials™ Speaking and Writing scores to the CEFR levels. In S. Papageorgiou & V. F. Manna (Eds.), *Meaningful language test scores: Research to enhance score interpretation* (pp. 120–140). John Benjamins. https://doi.org/10.1075/illa.1.07dav

Deygers, B., Zeidler, B., Vilcu, D., & Carlsen, C. H. (2018). One framework to unite them all? Use of the CEFR in European university entrance policies. *Language Assessment Quarterly*, *15*(1), 3–15. https://doi.org/10.1080/15434303.2016.1261350

Fleckenstein, J., Keller, S., Krüger, M., Tannenbaum, R. J., & Köller, O. (2020). Linking TOEFL iBT® writing rubrics to CEFR levels: Cut scores and validity evidence from a standard setting study. *Assessing Writing*, *43*, Article 100420. https://doi.org/10.1016/j.asw.2019.100420

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. https://doi.org/10.1111/j.1745-3992.2009.00168.x

Kane, M. T. (2002). Conducting examinee-centered standard-setting studies based on standards of practice. *The Bar Examiner*, *71*(4), 6–13.

Kingston, N. M., & Tiemann, G. C. (2011). Setting performance standards on complex assessments. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 201–224). Routledge. https://www.taylorfrancis.com/chapters/edit/10.4324/9780203848203-14/setting-performance-standards-complex-assessments-neal-kingston-gail-tiemann

North, B. (2000). *Theoretical studies in second language acquisition: Vol. 8. The development of a common framework scale of language proficiency.* Peter Lang.

Papageorgiou, S. (2010). Investigating the decision-making process of standard setting participants. *Language Testing*, *27*(2), 261–282. https://doi.org/10.1177/0265532209349472

Papageorgiou, S. (2016). Aligning language assessments to standards and frameworks. In D. Tsagari & J. Banarjee (Eds.), *Handbook of second language assessment* (pp. 327–340). De Gruyter Mouton. https://doi.org/10.1515/9781614513827-022

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

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

van Ek, J. A., & Trim, J. L. M. (1991). *Waystage 1990.* Cambridge University Press.

van Ek, J. A., & Trim, J. L. M. (1998). *Threshold 1990.* Cambridge University Press. https://doi.org/10.1017/CBO9780511667176

van Ek, J. A., & Trim, J. L. M. (2001). *Vantage.* Cambridge University Press. https://doi.org/10.1017/CBO9780511667114

Appendix A. Schedule for the Standard Setting Meeting

Wednesday, July 28, 2021 (prior to meeting)

- Receive preparation materials and familiarization exercise

Tuesday, August 3, 2021, 1:30–3:30 PM (EDT)

- Welcome and overview of the standard setting meeting
- Introduction to the TOEFL ITP Speaking test
- Introduction to the standard setting methodology
- Developing just qualified definitions for relevant CEFR levels
- Review of the familiarization exercise
- Practice for the judgment task
- Assignment of Round 1 judgment task

Wednesday, August 4, 2021 (between meeting sessions)

- Submit Round 1 judgments

Friday, August 6, 2021, 11:30 AM–1:30 PM (EDT)

- Review of Round 1 judgments
- Round 1 discussions
- Round 2 revision and finalization of cut scores

Appendix B. Preparation Activity Materials Provided to Panelists

Part 1. Please put the descriptors (numbers 1 to 24) into three groups (advanced, intermediate, elementary) according to your judgment.

Advanced	Intermediate	Elementary
Descriptors	Descriptors	Descriptors

Part 2. Please put the descriptors from each group (advanced, intermediate, elementary) into CEFR levels based on your understanding. The correct answers will be provided to you during the standard setting workshop.

CEFR levels	Descriptors	Groups
C2		Advanced
C1		
B2		Intermediate
B1		
A2		Elementary
A1		

Selected Descriptors From CEFR Speaking Subscales

No.	Descriptors
1	Can argue a case on a complex issue, formulating points precisely and employing emphasis effectively.
2	Has a good command of a very broad lexical repertoire including idiomatic expressions and colloquialisms; shows awareness of connotative levels of meaning.
3	Can manage very short, isolated, mainly pre-packaged utterances, with much pausing to search for expressions, to articulate less familiar words/signs, and to repair communication.
4	Prosodic features (e.g., word stress) are adequate for familiar everyday words and simple utterances.
5	Can articulate virtually all the sounds of the target language with clarity and precision.
6	Can construct phrases on familiar topics with sufficient ease to handle short exchanges, despite very noticeable hesitation and false starts.
7	Is generally intelligible throughout, despite regular mispronunciation of individual sounds and words they are less familiar with.
8	Can exploit a comprehensive and reliable mastery of a very wide range of language to formulate thoughts precisely, give emphasis, differentiate and eliminate ambiguity. No signs of having to restrict what they want to say.
9	Can produce appropriate collocations of many words/signs in most contexts fairly systematically.
10	Can correct slips and errors that they become conscious of, or that have led to misunderstandings.
11	Can communicate what they want to say in a simple and direct exchange of limited information on familiar and routine matters, but in other situations they generally have to compromise the message.
12	Can produce smooth, intelligible spoken discourse with only occasional lapses in control of stress, rhythm and/or intonation, which do not affect intelligibility or effectiveness.
13	Has enough language to get by, with sufficient vocabulary to express themselves with some hesitation and circumlocutions on topics such as family, hobbies and interests, work, travel and current events, but lexical limitations cause repetition and even difficulty with formulation at times.
14	Can develop a clear description or narrative, expanding and supporting their main points with relevant supporting detail and examples.
15	Can communicate basic information about personal details and needs of a concrete type in a simple way.
16	Has a basic vocabulary repertoire of words/signs and phrases related to particular concrete situations.
17	Can briefly give reasons and explanations for opinions, plans and actions.
18	Has a limited repertoire of short, memorised phrases covering predictable survival situations; frequent breakdowns and misunderstandings occur in non-routine situations.
19	Can express themselves fluently and spontaneously, almost effortlessly. Only a conceptually difficult subject can hinder a natural, smooth flow of language.

20	Can express themselves at length with a natural, effortless, unhesitating flow. Pauses only to reflect on precisely the right means to express their thoughts or to find an appropriate example or explanation.
21	Can articulate a limited number of sounds, so that speech is only intelligible if the interlocutor provides support (e.g. by repeating correctly and by eliciting repetition of new sounds).
22	Can produce clear, smoothly flowing, well-structured language, showing controlled use of organisational patterns, connectors and cohesive devices.
23	Can produce stretches of language with a fairly even tempo; although they can be hesitant as they search for patterns and expressions, there are few noticeably long pauses.
24	Can keep going comprehensibly, even though pausing for grammatical and lexical planning and repair is very evident, especially in longer stretches of free production.

For CEFR levels A1 through C2, please list what you think are some of the distinguishing speaking features that separate each level from the level above and below (e.g., the features of CEFR level B2 which distinguish that level from CEFR level B1 and CEFR level C1). Please consult the scales presented earlier in this section (Overall Oral Production and Overall Spoken Interaction, Phonological Control, General Linguistic Range) as well as the speaking scales in the Appendix. List between 3 and 5 distinguishing features in your own words. You may write a few key words or 1–2 sentences for each feature. ***Please do not exceed two pages.***

CEFR level	Distinguishing features for speaking
C2	• • • • •
C1	• • • • •
B2	• • • • •
B1	• • • • •

CEFR level	Distinguishing features for speaking
A2	• • • • •
A1	• • • • •

Appendix C. Borderline Student Definitions Made by the Panelists

CEFR level	Key distinguishing features of the CEFR levels
C2	• Finer shades of meaning • Consistent precision • Broad range • Degree of complexity • Able to reformulate • No need for circumlocution • Without restriction • Fluency with complex subjects • Organizational structure helps listener • Exploits prosodic features • L1 unnoticeable • Well-structured discourse • Repairs without interlocutor noticing • Detailed • Effortless • Memorability of utterance • Idiomatic expression, colloquial • Effective use of grammar and vocabulary • Involves and engages audience • Connotations utilized
C1	• Holds audience attention • Almost effortless expression • Can circumlocute effectively • Minor L1 influence, but no impact on clarity • Can adjust register and maintain consistent register • Wide range of vocabulary • Complex grammatical structure • Few or rare errors, which are difficult to spot • Notices own errors and self-corrects • Difficult concept may hinder speech somewhat • Detailed descriptions, but may search • Discourse competence with integrated sub-themes and appropriate conclusion • Expresses complex ideas cogently, but with occasional hesitation • Intelligible throughout with effective use of intonation • Errors may be noticeable but don't interrupt
B2	• Generally clear but noticeable accent • Slightly restricted, but communicates ideas effectively • Some fluency and spontaneity in interaction, with little strain • Moves into abstract topics

	• Errors don't lead to misunderstanding • Some repetitions • Sufficient linguistic resources with some complex structures • Few limitations in expression • Mastery of basic constructions, but may be hesitant or less successful with complex constructions • Elaboration and supported topic related to field of interest • Little obvious searching for words • Some complex grammar • Self-corrects a word or phrase, but may not always notice error • Some effort but no strain on either party • Generally well organized with clear detailed descriptions
B1	• Linear organization ("and then…") • Can discuss familiar topics and daily routine fairly successfully • Sufficient range for unpredictable situations but with hesitation (B1+?) • Lexical limitations result in significant hesitation, repetition, and obvious lexical planning • Straightforward descriptions within topics of their own expertise • Can sustain speech on straightforward topics, but may lose traction on more complex/conceptual topics • Conspicuous searching for words • Generally intelligible; some pronunciation, intonation, and stress errors that do not interfere with communication • Command of shorter utterances, but unsuccessful with longer ones • Neutral register — "one speed" — lack of audience awareness • Slow delivery, some hesitations, pausing for lexical/grammatical planning, and filler words • Restarts when communication breaks down
A2	• Can communicate simple ideas, but assistance needed to sustain • Formulaic chunks, simple expressions, basic sentence patterns/structures about self and immediate/familiar contexts • Short, simple descriptions • Memorized phrases • Familiar words are intelligible but listener may need to repeatedly ask for clarification • Pausing, hesitation, and repair evident • Some clarification by interlocutor is required/needed to steer conversation • Social exchanges are very short and not sustainable • Attempts to utilize language structures instead of relying on memorized phrases • Everyday polite forms

	• May tend to use lists • Strong L1 influence • Pronunciation may not be clear • Sufficient vocabulary to express basic needs • Short, simple, and predictable exchanges are manageable • A2+ — "like," "for example"
A1	• Few words/phrases understandable without listener effort • Produces short, isolated phrases with basic information • Repair, repetition, rephrasing, and slow rate • Limited connectors • Vocabulary limited to everyday, basic information • Can communicate some personal, concrete information • Brief and simple • Speech is rehearsed, simple, short phrases. • May need preparation time • Relies on signs/gestures • Resorts to occasional use of L1; breakdowns • Disconnected speech • Omissions • Basic, polite forms ("thank you," "sorry") • Able to express concrete needs • Can manage simple utterances with high frequency words and phrases • Significant interlocutor support needed • Can express concrete needs

Appendix D. Cut Score Rating Form Used by Panelists

Round	Minimum score CEFR A2	Minimum score CEFR B1	Minimum score CEFR B2	Minimum score CEFR C1
Round 1				
Round 2				