Research Memorandum
ETS RM–24-03

Mapping the Scores of the TOEFL Primary® Writing Test to the Common European Framework of Reference Levels

Michael Suhan
Spiros Papageorgiou
Mikyung Kim Wolf

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 Scores of the TOEFL Primary® Writing Test to the Common European Framework of Reference Levels

Michael Suhan, Spiros Papageorgiou, and Mikyung Kim Wolf
ETS, Princeton, New Jersey, United States

February 2024

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

Suggested citation: Suhan, M., Papageorgiou, S., & Wolf, M. K. (2024). *Mapping the scores of the TOEFL Primary® Writing test to the Common European Framework of Reference levels* (Research Memorandum No. RM-24-03). ETS.

Action Editor: Larry Davis

Reviewers: Ching-Ni Hsieh and Saerhim Oh

Abstract

This research memorandum reports on a study to map the scores of the TOEFL Primary® Writing test to the language proficiency levels of the Common European Framework of Reference for Languages (CEFR). Developed as a standalone module in the TOEFL Primary Test program, the TOEFL Primary Writing test is intended to measure young English as a foreign or additional language learners' computer-based English writing abilities to communicate about familiar topics related to their daily lives. As the TOEFL Primary Writing test is used globally, mapping its scores to the levels of the CEFR helps relevant stakeholders to interpret test results in relation to a globally used language framework as a point of reference. This report details the process undertaken, utilizing the judgment of expert panelists to establish recommended minimum test scores (cut scores) to classify test takers into CEFR levels. It also discusses study limitations and future research for continuous validation efforts to support appropriate interpretations and use of test results.

Keywords: TOEFL Primary®, Common European Framework of Reference (CEFR), standard setting, writing assessment, young learners, English as a foreign language, English as an additional language

To facilitate the interpretations and intended use of test scores for users, it is a common practice to link test scores to national or international proficiency levels (Papageorgiou, 2016). The Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001) and its companion volume (Council of Europe, 2020), which were developed to inform language curriculum development, teaching, and learning, provide a useful basis for score mapping. Following the widespread use of the CEFR in education systems around the world, the Council of Europe (2009) published a manual to guide test developers in linking test scores to the CEFR proficiency levels. Consequently, the worldwide adoption of the CEFR has led to the expectation for alignment of language test scores to CEFR proficiency levels (Deygers et al., 2018).

The purpose of the present study was to map the TOEFL Primary® Writing scores to the CEFR language proficiency levels. The score mapping was achieved by following recognized standard setting procedures to determine minimum test scores (or cut scores) indicating performance at specific CEFR proficiency levels. Curricula and instruction for young learners (YLs) in many parts of the world are informed by the CEFR. Therefore, mapping the TOEFL Primary Writing test scores to the CEFR levels can help score users understand test results in a way that is relevant to their local programs.

Use of CEFR Levels for Young Learners

The development of the CEFR followed several publications by the Council of Europe (e.g., van Ek & Trim, 1991, 1998, 2001) and North (2000) aimed at providing language proficiency scales for describing a wide variety of language activities and competencies. Additionally, the CEFR conceptualizes language learners as "social agents" who must complete tasks occurring within "a wider social context" (Council of Europe, 2001, p. 9), utilizing at times resources outside the construct of language ability. Language proficiency is described in the CEFR using illustrative descriptors placed on a common scale at six main levels, ranging from A1 (the lowest) to A2, B1, B2, C1, and C2 (the highest). Certain scales include additional descriptors for thresholds between criterion levels in the form of "plus levels" (i.e., A2+, B1+, B2+). With the 2020 publication of the companion volume, more scales with new descriptors were added, as was a Pre-A1 level in relevant scales (Council of Europe, 2020).

Although many of the CEFR's descriptors are applicable to the teaching, learning, and assessment of YLs, its descriptors were not deliberately designed for use with YLs nor was there a consideration of the social contexts or cognitive abilities relevant to YLs during the development process. In fact, the Council of Europe recognizes that current CEFR descriptors do not provide a comprehensive inventory of YL language activities and competences (Council of Europe, 2020). To facilitate use of the CEFR descriptors in YL contexts, the Council of Europe published two resource volumes grouped by the age range of YLs: the first encompassing the "primary age group" (ages 7–10; Council of Europe, 2018a, p. 12), and the second encompassing the "post-primary age group" (ages 11–15; Council of Europe, 2018b, p. 12). These resource volumes include CEFR descriptors with an indication of their relevance to YL groups.

The two volumes of CEFR descriptors related to YLs were not available when previous CEFR mapping studies were conducted for other tests in the TOEFL® Young Students Series (Baron & Papageorgiou, 2014; Tannenbaum & Baron, 2015). In these earlier studies, the researchers modified relevant CEFR descriptors to better correspond to the test constructs and content for the assessment of YLs. For instance, the descriptor including "common everyday or job related topics" was modified to "common everyday and school-related topics" (Papageorgiou & Baron, 2017). In the present study, the CEFR descriptors from the Council of Europe (2018a) pertinent to the primary age range were employed.

The TOEFL Primary Writing Test

The TOEFL Primary test measures the English language proficiency of YLs, ages 8 and above, who are learning English as a foreign/additional language by assessing students' knowledge, skills, and abilities to fulfill core social and academic communication goals in English. The TOEFL Primary test includes a computer-delivered speaking test, the recently added TOEFL Primary Writing test, and a two-step series of reading and listening tests. TOEFL Primary Step 1 is intended for students at the beginning stages of learning English; TOEFL Primary Step 2 is designed for students with some communicative English skills. Both tests are available on paper or computer. With the addition of the TOEFL Primary Writing test, the TOEFL Primary test has evolved into a four-skill assessment. However, each component may be delivered individually, allowing for the tests to be administered according to the specific needs of institutions and test takers.

The TOEFL Primary Writing test is intended to measure young English as a foreign/additional language learners' computer-based English writing abilities to communicate about familiar topics related to their daily lives. The main communication goals involved in the test tasks are to describe familiar topics and to narrate a simple story in a linear sequence (Wolf et al., 2024). Test results are intended for low-stakes use to guide instruction and learning, monitor student progress over time, and, if appropriate, use as a criterion in making placement decisions in English-language programs or local schools that teach English as part of the curriculum. Score reports for the writing test provide a variety of information including (a) band levels, (b) total scores from 0–17, (c) CEFR levels from below A1 to B1, and (d) proficiency descriptors with information for each section, including recommendations on next steps that students can take to improve their English language abilities. (For an overview of the TOEFL Primary Writing score report, see ETS, 2023a). Table 1 provides an overview of the four tasks included on the test.

Table 1. Overview of TOEFL Primary Writing Task Types

Task type	Response format	Points per item	No. of items
Write a Word	Typing a word	0 or 1	5
Build a Sentence	Ordering words	0 or 1	5
Edit a Text	Multiple choice	0 or 1	4
Write a Story	Typing sentences	0 to 3	1

Note. The scoring rubric for the Write a Story task is provided in Appendix A.

Considering YLs' characteristics, the test is designed to be delivered in a relatively short period of time (approximately 30 minutes) while measuring language skills relevant to primary school contexts across the range of proficiency levels from below A1 to B1 on the CEFR scale. The four tasks comprising the test were designed based on language activities selected from the written production and written interaction scales in the CEFR collated descriptors for YLs (Council of Europe, 2018a) with respect to (a) the cognitive abilities of the young students in the intended age range and (b) the target language use domain of the test. The first task (Write a Word) requires students to write a missing word using accurate forms in a provided sentence that describes a picture of people in familiar situations. The second task (Build a Sentence) presents pictures of people in familiar situations but requires the use of knowledge of English syntax and vocabulary to place given words in correct order to create a sentence describing each picture. In the third task (Edit a Text), students read a paragraph and select correct language forms, which tests the ability to review written texts and use knowledge of English grammar and usage to make texts meaningful and accurate. The fourth and final task (Write a Story)—aligned with the CEFR B1 Creative Writing descriptor, "Can narrate a story"—requires students to write a story based on a sequence of events presented in four pictures, measuring the ability to write a coherent story with appropriate details using knowledge of English vocabulary, grammar, and mechanics.

The scoring rubric for the Write a Story task was designed to cover the key subconstructs of content development, coherence, and language use on a holistic 4-point scale. In efforts to ensure construct congruence between the TOEFL Primary Writing test and CEFR levels, the CEFR descriptors were also taken into consideration in the development of the rubric. For instance, the CEFR A2 Overall Written Production descriptor, "Can write a series of simple phrases and sentences linked with simple connectors like 'and,' 'but' and 'because,'" guided the development of specific descriptors referencing the use of cohesive devices.

Methodology for the Standard Setting Study

General Procedures for Setting Cut Scores

To map test scores onto external proficiency levels, such as the CEFR levels, it is necessary to establish minimum scores (cut scores) on the test that correspond to the boundary between each level. The function of a cut score is to establish a minimum threshold of performance on a test that distinguishes test takers performing within a specific level from those who have not yet reached that level. In the body of literature surrounding score mapping, this process is referred to as *standard setting* (Cizek & Bunch, 2007). The standard setting process entails assembling a panel of experts who are led to make judgments about the questions, items, or tasks on the test together with test takers' knowledge and skills at the targeted proficiency levels that are then used as a basis to establish recommended cut scores. To inform the panelists' judgment, the study facilitators typically provide and explain statistical information about the test (e.g., item difficulty estimates and distribution of test scores). The process in most cases consists of two or three rounds through which panelists may revise their individual cut score recommendations based on new statistical information showing the implications of the recommendations made in prior rounds. After the panel recommends cut scores, the examination provider or score user considers the panel-recommended cut scores as the main factor in making the final decision about how to use the recommended cut scores, opting to either accept, raise, or lower each cut score.

Overview of the Score Mapping Process for the TOEFL Primary Writing Test

A series of standard setting meetings took place in December 2021 using data from the TOEFL Primary Writing field test, with the intent of making the CEFR cut scores available when the test was officially launched. The operational field-test form used in this study was administered to 602 students in intact classes in 11 countries where other TOEFL Primary tests are in wide use. The study team, consisting of two ETS senior research scientists and a research assistant, recruited a panel of 13 ETS assessment developers with backgrounds in English language teaching and assessment. Prior to the standard setting meetings, the panelists completed a series of familiarization activities to help ensure familiarity with the CEFR scales and descriptors that informed the development of the TOEFL Primary Writing test. The standard setting meetings were conducted remotely via Microsoft Teams and facilitated by the project team following typical standard setting methodology (i.e., Fleckenstein et al., 2020). The outcome of the standard setting study was a recommended set of cut scores indicating the minimum test scores needed to classify a test taker in CEFR levels A1, A2, and B1. This range of target levels was selected to align with the CEFR levels considered when designing the tasks on the test. Furthermore, the range corresponds with the target CEFR levels of the TOEFL Primary Reading and Listening test, which has a similar level of content. Following the standard setting meetings, the panel-recommended cut scores were evaluated and used to determine a final score mapping of TOEFL Primary Writing test scores to the CEFR levels.

Qualifications of Panelists

Potential panelists were asked to complete a background questionnaire and only individuals with both English language teaching experience and extensive test development experience were invited to participate in the study. All panelists had extensive professional experience in language assessment development and possessed a variety of educational backgrounds and professional experiences related to language teaching and assessment. They had prior experience working with YLs as an English as a second or foreign language teacher and on the development of standardized English language assessments. On average, the panelists had 13.2 years of experience in developing standardized English language assessments. In terms of educational background, all panelists had a language-related master's or doctoral degree. With respect to familiarity with CEFR levels and descriptors, all panelists indicated their familiarity except one panelist. Although most panelists were moderately to very familiar with CEFR levels and descriptors, everyone was provided with preparation materials including relevant CEFR level information prior to the standard setting meeting (discussed in the next section). Regarding the familiarity with standard setting procedures, all panelists had prior experience participating in standard setting meetings as a panelist.

Despite the expertise of the panelists, it is unlikely that the panel represented the ideal range of stakeholders who might be included in such a panel. Moreover, being composed solely of ETS staff, "insider bias" was a potential risk (Papageorgiou, 2010). *Insider bias* refers to the tendency of panelists who are aware of the CEFR levels targeted by the test designers to unintentionally make decisions about cut scores affected by their knowledge of the intended level coverage. For example, a panelist who knew that B1 is one of the intended CEFR levels would decide to set a B1 cut score, even if the test was too easy for learners at that level, and thus no such cut score should be set. To address the potential insider bias issue, the project team implemented a series of steps including premeeting preparation activities and detailed discussion of the minimum skills and abilities at each CEFR level (see the next two sections). The goal of these activities was to help panelists reach adequate understanding of each CEFR level and also to emphasize that the standard setting task, discussed in detail later, should be about skills and abilities at each CEFR level, not about confirming the CEFR levels targeted by the test development team. The panel members were not involved in the development of the TOEFL Primary Writing test. It should also be pointed out that convening a panel of ETS staff with the requisite expertise allowed for the study to be conducted efficiently in advance of the public release of the operational test, particularly given the challenges of recruiting qualified panelists around the globe during the pandemic period. Overall, given the expertise of these panelists and the implementation of the above measures to mitigate potential insider bias, the standard setting meeting proceeded with their involvement.

Panelist Preparation Prior to the Standard Setting Meeting

Prior to the standard setting study, a preparation guide was created and sent to the panelists. The guide included information about the CEFR and the TOEFL Primary Writing test as well as familiarization activities related to the CEFR scales for overall written production and overall written interaction (see Appendix B). All panelists were asked to individually complete two familiarization activities to help ensure that they had a good understanding of the features that distinguished CEFR levels pre-A1 to B1 on the two overall scales.

In the first activity, panelists were asked to review the two overall scales and then, based on their understanding of the scales, to sort the CEFR writing subscale descriptors used to inform test development into the appropriate scale level. In the second activity, panelists were asked to list what they believed were three to five key distinguishing writing features that separated adjacent CEFR levels based on how they sorted the descriptors in the first activity. The familiarization activities were completed online, and upon completion, panelists received a copy of their responses and the answer keys for both activities. An example of the familiarization activity is provided in Appendix C.

The preparation guide additionally included a detailed overview of the test, the scoring rubric, and examples of individual test items for each task. The panelists were also asked to review a document illustrating the test-taking experience. This document contained screenshots of the test on the testing platform so that panelists could gain an understanding of (a) the testing interface and navigation, (b) the test composition and content, and (c) the difficulty of the tasks and items.

Discussion of Familiarization Activities and Definition of the Just Qualified Candidate

The first familiarization activity was reviewed at the beginning of the first standard setting meeting, allowing panelists to discuss any challenges in distinguishing the descriptors across CEFR levels. Afterward, panelists participated in further discussion until they were able to reach an agreement on what language skills were needed to reach CEFR levels A1, A2, and B1. The second familiarization activity, where panelists had noted distinguishing features at each level, was used to support this activity. A test taker with the identified minimally acceptable skills was defined as the just qualified candidate (JQC) for the given level. The panel's definitions for JQCs at the three CEFR levels are listed in Appendix D.

Standard Setting Method

The standard setting method used for this study is a variation of the performance profile method (Fleckenstein et al., 2020; Hambleton et al., 2000). Following this method, panelists make holistic judgments of profiles of student performance. The responses of each test taker were presented to the panelists as they were entered on the test platform. Only the overall writing test score, and not individual test task scores, were made available to panelists. The use of the overall writing test scores was in line not only with the type of score reported to test takers, but also with the holistic approach of the standard setting judgment task. The written instructions for the judgment task provided to the panelists were as follows: "Imagine one JQC CEFR A1. Read responses of actual TOEFL Primary test takers. What writing score would the JQC CEFR A1 earn?" The panelists were asked to perform this judgment task for CEFR levels A2 and B1 and to indicate the score for these two levels as well. Another reason for selecting the performance profile method was that its holistic review approach was relevant to panelists' professional expertise as educators and assessment developers, where it is common to make such judgments (Kingston & Tiemann, 2011).

In this study, the panelists reviewed the responses of individuals who had participated in the TOEFL Primary Writing field test. A sample of 38 test takers was drawn from 602 participants who had all completed the same version of the test form at the time of sampling. Individuals were selected to represent the full range of raw scores (0–17) for the field test. For each test taker, a portfolio was created that included responses to every item for each of the four tasks (five items for the Write a Word task, five items for the Build a Sentence task, four items for the Edit a Text task, and one item for the Write a Story task). Panelists were provided with the responses exactly as they were entered by test takers without information indicating if they were scored as correct or incorrect because the focus of the judgment task was holistically on the overall performance across all tasks of the writing test. For the Write a Word task, panelists were shown how test takers spelled the word. Similarly, for the Build a Sentence task, panelists were provided with the actual sentences constructed by test takers. The Edit a Text task included the option (either the key or one of two distractors) selected by test takers for each of the four items. The response written for the Write a Story task was provided without a task score. To identify test takers who would receive the writing score expected by a JQC for CEFR levels A1, A2, and B1, panelists were instructed to read the scoring rubric and performance descriptors included in the preparation guide together with the test takers' responses.

Two rounds of judgments took place within 3 days of each other with feedback and discussion between rounds (see the sample of the rating form in Appendix E). To make cut score recommendations for Round 1, panelists were asked to review JQC descriptions for CEFR levels A1, A2, and B1. Before panelists made cut score recommendations, the study facilitators provided a demonstration of the process. Time was allotted for panelists to ask questions about the process, and the panel was asked questions intended to check their comprehension of the instructions. Next, panelists reviewed test takers' complete set of responses across the 38 portfolios to decide the scores that a JQC at each CEFR level would receive (i.e., the cut scores). To begin Round 2, the mean, median, mode, minimum, and maximum of the Round 1 cut scores were presented to the panel, and panelists shared their judgment rationales. Panelists were also shown impact data to inform them of what percentage of students from the field test would be classified into each CEFR level. Panelists were then asked to review the complete set of responses again and to make a final decision as to what scores the JQCs would receive.

At the end of the second meeting, the research team asked the panelists to complete a questionnaire about their perceptions of the standard setting process, the influence of the materials used in the study, and the meeting format. Panelists were also asked to rate their comfort level with the panel-recommended cut scores and provided with an opportunity to comment on the study.

Results of the Standard Setting Study

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

Recommended Cut Scores for the TOEFL Primary Writing Test

The panel recommendations for cut scores corresponding to CEFR levels are presented in Table 2. The results include the mean, median, minimum, maximum, standard deviation, and standard error of judgment (SEJ) of each round of judgments. The SEJ is included as an estimate of the uncertainty in the panelists' judgments. The SEJ is computed by dividing the standard deviation of the judgments by the square root of the number of panelists (Cizek & Bunch, 2007) and can be interpreted as an indication of how close each recommended cut score is likely to be to a cut score recommended by other panels of experts similar in composition to the current panel and similarly trained in the same standard setting methods. A comparable panel's cut score would be within one SEJ of the cut score 68% of the time and within two SEJs 95% of the time. In order to reduce the impact on misclassification rates (false positives and false negatives), Cohen et al. (1999) suggested that an SEJ should be no more than half the value of the standard error of measurement.

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 as the raw total writing test score on a scale from 0–17, which is the metric that the panelists used. The mean cut score for all CEFR levels across the two rounds changed by less than 1 point, while the median cut score for all levels did not change. The variability in panelists' judgments decreased in general across rounds, as indicated by the standard deviation, suggesting convergence in the final round of judgments. The SEJ for all cut scores across rounds was less than half of the standard error of measurement for the test (1.80 for the test form used in the standard setting study).

Table 2. Standard Setting Results for the TOEFL Primary Writing Test

	Round 1			Round 2		
Description	CEFR A1	CEFR A2	CEFR B1	CEFR A1	CEFR A2	CEFR B1
Mean	6.85	10.54	14.62	7.15	11.00	15.15
Median	7	11	15	7	11	15
Mode	8	12	15	6, 7, 8	11	15
Minimum	4	8	12	6	10	15
Maximum	9	12	16	9	12	16
SD	1.63	1.51	1.39	0.99	0.58	0.38
SEJ	0.45	0.42	0.38	0.27	0.16	0.10

Note. N = 13. CEFR = Common European Framework of Reference for Languages; SEJ = standard error of judgment.

TOEFL Primary Writing test results are reported in terms of a raw total score, ranging from 0–17 in 1-point increments. However, when a recommended (mean) cut score is not a whole number, the conversion process first requires a decision to be made about rounding the panel's recommended cut scores, given that only whole numbers, rather than decimals, are reported. There are two options for rounding the recommended cut scores for the TOEFL Primary Writing test:

- The recommended raw cut score is rounded down to the previous whole number, the justification being that although the decimal values indicate ability beyond a given score point, the next higher score has not been achieved. Using the previous example, a raw score of 15.15 means that the cut score should be 15, because the next higher score, 16, was not recommended by the panel.
- The recommended raw cut score is rounded up to the next whole number, the justification being that the decimals indicate ability beyond a given score point. For example, a raw cut score of 15.15 means that the cut score should be 16 to indicate that the minimum score is above 15.

Table 3 shows the results of both conversion rules applied to the panel-recommended raw cut scores. The rounding of cut scores should also be informed by the implications for false positive and false negative classifications (see discussion in Papageorgiou et al., 2015).

Table 3. Raw and Rounded Cut Scores for the TOEFL Primary Writing Test Recommended by the Panel

Cut score	CEFR A1	CEFR A2	CEFR B1
Panel-recommended cut scores (raw)	7.15	11.00	15.15
Cut scores converted to total scores by rounding down	7	11	15
Cut scores converted to total scores by rounding up	8	11	16

Note. CEFR = Common European Framework of Reference for Languages. In the case of the CEFR A2 cut score, rounding up or rounding down did not make a difference.

Final Score Mapping

To arrive at the final cut scores for each CEFR level, further discussions were conducted among the panelists, researchers, and test developers based on the panelists' cut scores. The following is a summary of the discussion that led to final cut scores (Table 4 presents the recommended cut scores).

- The panel's recommendation for the A1 cut score was decreased from 7.15 to 6. This decision was made to reduce the chances for false negative classifications (i.e., test takers being classified as below the A1 level when they are actually at the A1 level). Considering that the target test takers are YLs who are at a relatively early stage of developing their English writing abilities, the panel recommended reducing the number of possible false negative classifications to avoid negative impacts on beginning writers' motivation.

- The panel's recommendation for the B1 cut score was rounded from 15.15 to 16 to eliminate the possibility of a test taker being classified as B1 without demonstrating adequate ability to construct a short piece of writing. However, with a cut score of 15, it would be possible for a learner to receive a B1 classification with a score of 1 on the Write a Story task, while a score of 2 or higher must be attained on this task in order to achieve a total score of 16. A score of 1 on the Write a Story task is clearly much lower than the type of writing expected by a learner at the B1 level, even from a learner who has just crossed the line from A2 to B1. To safeguard scores against false positive classifications for the B1 level, 16 was set as the B1 cut score.

Table 4. Recommended Cut Scores of TOEFL Primary Writing Test for Each CEFR Level

CEFR level	TOEFL Primary Writing test total score
B1	16
A2	11
A1	6

Note. CEFR = Common European Framework of Reference for Languages.

Results of the Meeting Evaluation Survey

Panelists' evaluation of the standard setting process provides important procedural validity evidence for the score mapping produced by the study (Tannenbaum & Cho, 2014). Table 5 summarizes the panel's feedback. The evaluation form was completed by 10 of the 13 panelists, the majority of whom expressed an understanding of the study and its procedures. No panelists who provided feedback disagreed with any statement.

Table 5. Panelists' Perceptions of Clarity and Helpfulness of Instructions and Feedback

Question	Strongly disagree	Disagree	Agree	Strongly agree
I understood the purpose of this study.	0	0	1	9
I understood the descriptors to distinguish CEFR levels.	0	0	4	6
The instructions and explanations provided by the facilitators were clear.	0	0	1	9
The explanation of the standard setting method and procedure was adequate to give me the information I needed to complete my assignment.	0	0	1	9
The explanation of how the recommended cut score is computed was clear.	0	0	3	7
The opportunity for feedback and discussion between the two judgment task rounds was helpful.	0	0	1	9
The information about the percentage of test takers placed into CEFR levels was helpful.	0	0	5	5

Note. CEFR = Common European Framework of Reference for Languages.

Regarding a question as to which particular factors influenced their standard setting judgments (see Appendix F, Table F1 for panelists' responses), the majority of panelists found that the definition of the JQC, the discussion of the descriptors to distinguish CEFR levels, and the description of the knowledge/skills required to answer each test task type were very influential. The between-round discussion was thought to be influential or very influential by all panelists who responded. The majority of panelists found that the cut scores of other panel members were influential, while two did not find them influential. Six panelists did not find the percentage of test takers placed into CEFR levels influential, with one finding it not applicable and three finding it influential. The majority of panelists found their own professional experiences either influential or very influential, whereas two did not consider this to be an influential factor.

The evaluation form also asked panelists to rate the degree to which the meeting process was (a) efficient, (b) coordinated, (c) understandable, and (d) satisfying (Appendix F, Table F2). Panelists gave generally high ratings for all aspects of the meeting process. In an open-ended response, one panelist commented, “I felt the facilitators did an excellent job of explaining the purpose and methods involved in this standard-setting activity.”

Panelists were also asked to indicate their level of confidence with standard setting results (Appendix F, Table F3). The majority of panelists were either very comfortable or comfortable with the recommended cut scores. However, two panelists indicated that they were either very uncomfortable or uncomfortable with the recommended cut scores. Reasons for discomfort with the recommended cut scores were not addressed when panelists were asked to provide comments on the standard setting process. However, one panelist who was comfortable with the cut scores commented, “I would have liked more samples so that I could feel very comfortable with my cut scores. I enjoyed the discussion and the experience very much. It was challenging and interesting.”

Discussion

The purpose of this study was to establish cut scores for the TOEFL Primary Writing test aligned with the levels of the CEFR. As described earlier, the CEFR levels and descriptors served as one of the guiding documents for the development of the TOEFL Primary Writing test and its scoring rubric. Mapping TOEFL Primary Writing test scores to CEFR levels allows educators to relate TOEFL Primary Writing test scores to an international benchmark. It is particularly useful in educational contexts where teaching and learning is also aligned to the CEFR.

It is important to note that the construct measured by the TOEFL Primary Writing test is the young English as a foreign/additional language students' ability to communicate in written English on familiar topics in social and academic contexts within a computer-mediated environment. The major communication goals represented in the test focus on describing situations and narrating stories with four task types. Thus, the present standard setting study utilized CEFR level descriptors that are both relevant to YLs and aligned to the construct and content of the TOEFL Primary Writing test. To adequately interpret score levels in relation to CEFR levels, it is important for test users to refer to additional ETS publications supporting score interpretation, particularly *TOEFL Primary®: Understanding Your Writing Score Report* (ETS, 2023a) and *TOEFL Primary®: Writing Score Level Descriptors* (ETS, 2023b).

It is also important for score users to be aware that the CEFR classifications specified in this study represent probabilities that test takers' writing skills are aligned with CEFR descriptors within the domains of written production and written interaction. This distinction may become less meaningful in the case of the JQC or in other borderline cases where a test taker scores just below a cut score. To that end, TOEFL Primary Writing test scores and aligned CEFR levels should be regarded as only one source of information and used together with other available information related to English language proficiency (e.g., teacher observations, formative assessment) and the teaching and learning context within which the TOEFL Primary Writing test scores and aligned CEFR levels are used.

Despite the low SEJ of this study's panel of experts, it should also be noted that recommended cut scores were based on a sample of student performance on a field test prior to operational delivery. As mentioned earlier, the panelists involved in this study had extensive experience in both instruction and assessment of English language learning but were exclusively ETS assessment developers. Their affiliation could be a concern for insider bias, as discussed earlier. Despite the measures taken to address this concern, it is crucial to maintain ongoing monitoring of TOEFL Primary Writing test data and cut scores to ensure valid interpretations and use of the test scores and to treat the results of this study as an initial mapping. Research to validate the CEFR score mapping for the TOEFL Primary Writing test is planned for the near future. This research will entail a close examination of test takers' performances in relation to CEFR descriptors and gather additional criterion measures (e.g., teachers' ratings, students' English writing samples from an additional measure). The results of such research will provide additional evidence to confirm the score mapping reported here or to inform further refinements.

References

Baron, P. A., & Papageorgiou, S. (2014). *Mapping the TOEFL Primary test onto the Common European Framework of Reference* (Research Memorandum No. RM-14-05). ETS. https://www.ets.org/Media/Research/pdf/RM-14-05.pdf

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

Cohen, A. S., Kane, M. T., & Crooks, T. J. (1999). A generalized examinee-centered method for setting standards on achievement tests. *Applied Measurement in Education*, *12*(4), 343–366. https://doi.org/10.1207/S15324818AME1204_2

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. (2018a). *Collated representative samples of descriptors of language competences developed for young learners. Resource for educators: Vol. 1. Ages 7–10*. https://rm.coe.int/collated-representative-samples-descriptors-young-learners-volume-1-ag/16808b1688

Council of Europe. (2018b). *Collated representative samples of descriptors of language competences developed for young learners. Resource for educators*: *Vol. 2. Ages 11–15.* https://rm.coe.int/collated-representative-samples-descriptors-young-learners-volume-2-ag/16808b1689

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment. Companion volume.* http://bit.ly/42YaANi

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

ETS. (2023a). *TOEFL Primary®: Understanding your writing score report.* https://www.ets.org/pdfs/toefl/toefl-primary-understand-writing-score-reports.pdf

ETS. (2023b). *TOEFL Primary®: Writing score level descriptors.* https://www.ets.org/pdfs/toefl/toefl-primary-writing-score-descriptors.pdf

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

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

Hambleton, R. K., Pitoniak, M. J., & Copella, J. M. (2012). Essential steps in setting performance standards on educational tests and strategies for assessing reliability of results. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 47–76). Routledge. https://bit.ly/49NlgAm

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., & Baron, P. (2017). Using the Common European Framework of Reference to facilitate score interpretations for young learners' English language proficiency assessments. In M. K. Wolf & Y. G. Butler (Eds.), *English language proficiency assessments for young learners* (pp. 136–152). Routledge. https://doi.org/10.4324/9781315674391-8

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

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

Tannenbaum, R. J., & Baron, P. A. (2015). *Mapping scores from the TOEFL Junior® comprehensive test onto the Common European Framework of Reference (CEFR)* (Research Memorandum No. RM-15-13). ETS. https://www.ets.org/Media/Research/pdf/RM-15-13.pdf

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

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

Wolf, M. K., Suhan, M., Ginsburgh, M., Futagi, Y., & Li, F. (2024). *Design framework for the TOEFL Primary® Writing test* (Research Memorandum No. RM-24-02). ETS. https://www.ets.org/Media/Research/pdf/RM-24-02.pdf

Appendix A. Write a Story Scoring Rubric

Score	Descriptors
3	The test taker achieves the communication goal. A typical response at this level is characterized by the following: The response is complete with appropriate details. For items with a required word list, all of the words are used. The response maintains coherence with the support of cohesive devices (e.g., pronouns, transition words). The language demonstrates accuracy and/or variety in word choice, grammar, and mechanics (e.g., capitalization, punctuation, spelling), though a few errors may be present.
2	The test taker partially achieves the communication goal. A typical response at this level is characterized by the following: The response is partially complete, with some appropriate details. For items with a required word list, some of the words are used. Parts of the response are coherent. Limitations or inaccuracies in the use of cohesive devices weaken the overall coherence. The language demonstrates a lack of variety or control of sentence structures and may include multiple errors in word choice, grammar, and mechanics (e.g., missing punctuation or inaccurate spelling).
1	The test taker attempts to achieve the communication goal. A typical response at this level is characterized by the following: The response is incomplete, perhaps addressing only one picture beyond the given sentence or one aspect of the descriptive prompt. Appropriate details may be expressed in single words, short phrases, or even a single sentence. For items with a required word list, few, if any, of the words are used. The response is mostly incoherent. The word choice is basic and/or repetitive, and the grammar and mechanics are mostly inaccurate. Major errors are present throughout the response, or the response is too short to evaluate language use.
0	A typical response at this level may be: Off-topic (e.g., a memorized response to a different question) Entirely in another language Random strings of letters No response (i.e., blank) A copy of the prompt or provided scaffolding language (with no attempt to modify or create new language) Contains only “I don’t know”

Appendix B. CEFR Scales Used in Familiarization Activity

Table B1. CEFR Scale for Overall Written Production

C2	Can produce clear, smoothly flowing, complex texts in an appropriate and effective style and a logical structure which helps the reader identify significant points.
C1	Can produce clear, well-structured texts of complex subjects, underlining the relevant salient issues, expanding and supporting points of view at some length with subsidiary points, reasons and relevant examples, and rounding off with an appropriate conclusion. Can employ the structure and conventions of a variety of genres, varying the tone, style and register according to addressee, text type and theme.
B2	Can produce clear, detailed texts on a variety of subjects related to their field of interest, synthesising and evaluating information and arguments from a number of sources.
B1	Can produce straightforward connected texts on a range of familiar subjects within their field of interest, by linking a series of shorter discrete elements into a linear sequence.
A2	Can produce a series of simple phrases and sentences linked with simple connectors like “and,” “but” and “because.”
A1	Can give information about matters of personal relevance (e.g. likes and dislikes, family, pets) using simple words/signs and basic expressions. Can produce simple isolated phrases and sentences.
Pre-A1	Can give basic personal information (e.g. name, address, nationality), perhaps with the use of a dictionary.

Note. CEFR = Common European Framework of Reference for Languages. Source: Council of Europe (2020, p. 66).

Table B2. CEFR Scale for Overall Written Interaction

C2	Can express themselves in an appropriate tone and style in virtually any type of formal and informal interaction.
C1	Can express themselves with clarity and precision, relating to the addressee flexibly and effectively.
B2	Can express news and views effectively in writing, and relate to those of others.
B1	Can compose personal letters and notes asking for or conveying simple information of immediate relevance, getting across the point they feel to be important.
A2	Can compose short, simple formulaic notes relating to matters in areas of immediate need.
A1	Can ask for or pass on personal details.
Pre-A1	Can convey basic information (e.g. name, address, family) in short phrases on a form or in a note, with the use of a dictionary.

Note. CEFR = Common European Framework of Reference for Languages. Source: Council of Europe (2020, p. 82).

Appendix C. Materials Used by Panelists in Familiarization Activity

Table C1. Table Used by Panelists to Sort CEFR Descriptors in Familiarization Task 1

CEFR levels	Descriptors
B1	
A2	
A1	
Pre-A1	

Note. CEFR = Common European Framework of Reference for Languages.

Table C2. Selected Descriptors from CEFR Writing Sub-Scales Used in Familiarization Task 1

1	Can write with reasonable phonetic accuracy (but not necessarily fully standard spelling) short words that are in their oral vocabulary
2	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.
3	Can use isolated words/signs and basic expressions in order to give simple information about themselves.
4	Uses some simple structures correctly, but still systematically makes basic mistakes; nevertheless, it is usually clear what they are trying to say.
5	Can use some basic structures in one-clause sentences with some omission or reduction of elements.
6	Can write an introduction to a story or continue a story, provided he/she can consult a dictionary and references (e.g. tables of verb tenses in a course book).
7	Can communicate very basic information about personal details in a simple way.
8	Can form longer sentences and link them together using a limited number of cohesive devices, e.g. in a story.
9	Shows only limited control of a few simple grammatical structures and sentence patterns in a learnt repertoire.
10	Can employ very simple principles of word/sign order in short statements.
11	Can use basic punctuation (e.g. full stops, question marks).
12	Spelling, punctuation and layout are accurate enough to be followed most of the time
13	Has a basic vocabulary repertoire of words/signs and phrases related to particular concrete situations
14	Has sufficient vocabulary for the expression of basic communicative needs.
15	Can write accounts of experiences, describing feelings and reactions in simple connected text.

Table C3. Table Provided to Panelists for Listing Key Distinguishing Features in Familiarization Task 2

CEFR level	Distinguishing features for writing
B1	
A2	
A1	
Pre-A1	

Note. CEFR = Common European Framework of Reference for Languages.

Appendix D. Distinguishing Features of JQCs Selected Defined by Panelists

CEFR level	Distinguishing features expected for the JQC
B1	Connected texts (e.g., storytelling, range of familiar topics) Wide vocabulary for familiar topics Can use circumlocution Spelling and punctuation reasonably accurate
A2	Formation of communicative ability Can relate an immediate need Can link simple sentences Can use more connectors (e.g., but, because) at clause and sentence level Communication not totally impeded by errors Production is simple but understandable / can tell what they’re trying to communicate Can write with reasonable phonetic accuracy Variety of words and phrases dealing with concrete and familiar contexts Emerging mastery of tense Light / mild paraphrasing More facility with basic social language (e.g., politeness) May have trouble communicating about non-routine situations
A1	Use of connectors to link words Basic topics outside of personal info (e.g., expressing likes and dislikes) Formulaic words and phrases / simple sentences / concrete vocabulary Not completely reliant on dictionary Emerging grammatical control Use of basic punctuation Can provide simple descriptions (e.g., color of object)

Note. JCQ = just qualified candidate; CEFR = Common European Framework of Reference for Languages.

Appendix E. Panelist Cut Score Rating Form

Instructions

- Read the scoring rubric
- Read the performance descriptors in your preparation guide
- Read responses by test takers who received different writing scores
- Identify test takers who would receive the writing score expected by JQC at CEFR A1, A2, B1
- Enter the score these test takers received (not the test taker ID number)

Round	Minimum score CEFR A1	Minimum score CEFR A2	Minimum score CEFR B1
Round 1			
Round 2			

Appendix F. Panelists’ Responses to the Evaluation Form

Table F1. Panelists' Opinion About the Influence of Study Materials in Making Standard Setting Judgments

Question	Not applicable	Not influential	Influential	Very influential
The definition of the just qualified candidate (JQC)	0	0	0	10
The discussion of the descriptors to distinguish CEFR levels	0	0	1	9
The description of the knowledge/skills required to answer each test task type	0	0	2	8
The between-round discussion	0	0	6	4
The cut scores of other panel members	0	2	8	0
The percentage of test takers placed into CEFR levels	1	6	3	0
My own professional experience	0	2	6	2

Note. CEFR = Common European Framework of Reference for Languages.

Table F2. Panelists' Evaluation of the Meeting Process

Question	1	2	3	4	5
Inefficient (1) — Efficient (5)	0	0	2	2	6
Uncoordinated (1) — Coordinated (5)	0	0	0	0	10
Confusing (1) — Understandable (5)	0	0	0	2	8
Dissatisfying (1) — Satisfying (5)	0	0	0	3	7

Table F3. Panelists' Comfort With Standard Setting Results

Question	Very uncomfortable	Uncomfortable	Comfortable	Very comfortable
Please indicate the degree to which you were comfortable with the recommended cut scores.	1	1	5	3