Quality Beyond Measure.

TOEFL® Research Report
TOEFL–RR-100
ETS Research Report No. RR–22-16

Mapping *TOEFL*® *Essentials*™ Test Scores to the Canadian Language Benchmarks

Spiros Papageorgiou

Larry Davis

Renka Ohta

Pablo Garcia Gomez

December 2022

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

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

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

Current (2022–2023) members of the TOEFL COE are:

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

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

ETS is an Equal Opportunity/Affirmative Action Employer.

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

TOEFL Research Report Series and ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Mapping *TOEFL®* *Essentials*™ Test Scores to the Canadian Language Benchmarks

Spiros Papageorgiou, Larry Davis, Renka Ohta, & Pablo Garcia Gomez

ETS, Princeton, NJ

In this research report, we describe a study to map the scores of the *TOEFL® Essentials*™ test to the Canadian Language Benchmarks (CLB). The TOEFL Essentials test is a four-skills assessment of foundational English language skills and communication abilities in academic and general (daily life) contexts. At the time of writing this report, the test was the most recent addition to the *TOEFL®* Family of Assessments. TOEFL Essentials test scores are intended to provide academic programs and other users with reliable information regarding the test taker's ability to understand and use English. Mapping of scores to widely used language frameworks such as the CLB provides additional support for interpreting test results and for making inferences regarding test-taker abilities. The score mapping process consisted of the following steps, as recommended in the literature: (a) establishing construct congruence between the test content and the performance descriptors of the CLB; (b) establishing recommended minimum test scores (cut scores) required to classify language learners into CLB levels, based on the judgments of local experts; and (c) providing evidence of procedural, internal, and external validation of the recommended cut scores.

Keywords *TOEFL® Essentials*™ test; Canadian Language Benchmarks (CLB); standard setting; cut scores; score interpretation

doi:10.1002/ets2.12357

Mapping (aligning or linking) test scores to external proficiency levels and descriptors is a common approach to facilitate the interpretation of test scores (Tannenbaum & Cho, 2014). The foremost example of a language proficiency framework being used to interpret test scores is arguably the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001). The CEFR was introduced in 2001 and expanded with a companion volume (Council of Europe, 2020) in order to promote the development of language learning curricula and provide an orientation for language teaching and learning. Because of the worldwide adoption of the CEFR, language tests are often expected to provide scores that can be interpreted in reference to the CEFR proficiency levels (Deygers et al., 2018). The widespread use of the CEFR in educational systems around the world led its developer to publish a manual to guide test developers in linking test scores to the CEFR levels (Council of Europe, 2009).

In addition to the CEFR, there is a growing literature on the mapping of language proficiency scores to local proficiency frameworks (e.g., Dunlea et al., 2019; Papageorgiou et al., 2019). As is the case with the CEFR, mapping of test scores to local frameworks can inform decisions about language proficiency requirements within a particular geographic, educational, or social context. One such example is the Canadian Language Benchmarks (CLB), which comprise a framework for understanding language ability that is used by a wide variety of educational, workplace, and governmental institutions in Canada to document and measure second language proficiency in English and French in adult immigrants (www.language.ca). The CLB, published by the Centre for Canadian Language Benchmarks (2012), incorporates 12 levels (or benchmarks) extending from very basic to highly advanced language ability, as relevant to adult ESL learners. The 12 levels are further organized into three superordinate stages: Stage I-Basic, Stage II-Intermediate, and Stage III-Advanced. These stages differ broadly in terms of the complexity of language the learner is expected to comprehend and produce and the types of tasks and contexts that the learner can successfully navigate. The descriptions of ability found in the CLB provide a reference to inform teaching, learning, and assessment and also serve as a standard for setting language proficiency requirements across a broad range of settings, such as academic admissions, employment, professional licensing, and immigration.

In this research report, we describe the application of a series of steps in mapping the scores of the *TOEFL® Essentials*™ test to the CLB. At the time of writing this report, the TOEFL Essentials test was the most recent addition to

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

the *TOEFL*® Family of Assessments, and mapping scores to the CLB was seen as an important step to facilitate use of this new test in Canada. The score mapping process consisted of the following steps recommended in the literature: (a) establishing construct congruence between the test content and the performance descriptors of the CLB; (b) establishing recommended minimum test scores (cut scores) required to classify language learners into CLB levels, based on the judgments of local experts; and (c) providing evidence of procedural, internal, and external validation of the recommended cut scores.

The TOEFL Essentials Test

The TOEFL Essentials test measures the English language proficiency of older adolescents and adults and was added to the TOEFL Family of Assessments in August of 2021. The test measures the four language skills of listening, reading, writing, and speaking, in contexts characteristic of both daily life and academic study. Test design and delivery are intended to emphasize both measurement quality and test-taker access, with the latter achieved through at-home test administration, relatively low cost, and a relatively short testing time (approximately 90 min). TOEFL Essentials test scores are reported in the form of band scores from 1 to 12, with an overall band score reported in 0.5 point increments and four section scores reported in 1-point increments. Test scores are intended for use by learners who need to demonstrate their English language proficiency for various purposes, such as admission to academic programs. A detailed description of the design of the TOEFL Essentials test, score reporting, and intended uses can be found in Papageorgiou et al. (2021).

The TOEFL Essentials test is designed to efficiently assess learners and provide trustworthy measurement of language skills across a wide range of proficiency, extending from levels A1 to C2 on the CEFR (Council of Europe, 2001). A combination of efficiency, trustworthiness, and proficiency range is accomplished through a variety of means. First, a multistage adaptive test (MST) design is used; the test sections for listening, reading, and writing include a first stage with tasks of average difficulty, followed by a second stage where the difficulty of the test tasks is determined by performance on the first stage. This design makes it possible to quickly and precisely measure ability across a broad range.

Additionally, the test incorporates different types of language tasks that assess both foundational and communicative skills in English. Tasks targeting foundational skills allow for highly reliable measurement of general language proficiency across a broad proficiency range. Such tasks include identifying synonyms (reading section), arranging words and phrases in a grammatical sequence (writing section), and listening to and repeating sentences of increasing length (speaking section). These tasks of foundational abilities are accompanied by communicative tasks to help ensure that test scores will support trustworthy inferences regarding the ability to use language in real life. Such communicative tasks require understanding the format, content, and purposes of written texts (reading section) and audio input (listening section), writing short texts such as emails to accomplish various communicative goals (writing section), and responding to an interviewer's questions regarding one's experiences and opinions (speaking section). The task types used in the TOEFL Essentials test are listed in Table 1 and discussed further in the following section; additional information is available in Papageorgiou et al. (2021).

Overview of the Score Mapping Process

In this section, we provide an overview of the key components of the score mapping project, which was conducted between October 2021 and February 2022. The mapping project began with a construct congruence study by ETS staff to compare the content of four sections of the TOEFL Essentials test to the CLB descriptors for listening, reading, speaking, and writing. This effort was necessary because external levels and descriptors tend to describe ability in general terms and so are likely to suffer from what has been called "descriptional inadequacy" (Fulcher et al., 2011, p. 8). That is, descriptions of abilities are made at a high level of abstraction and therefore miss much of the detail of real-life language use. One consequence of this abstraction is that external level descriptors do not provide information that fully describes performance on any particular test. Given this limitation of external levels and descriptors, evidence of "construct congruence" (Tannenbaum & Cho, 2014) is needed first to establish that a test measures language skills in a manner consistent with the way the external levels describe language proficiency.

Upon establishing construct congruence, a standard setting study was conducted that included all four sections of the test. The project team, consisting of two ETS senior research scientists and a research project manager, worked with ETS Canada staff to recruit individuals for the study. Ultimately, a panel of 17 educators was selected who represented a

Table 1 Content Overview of the TOEFL Essentials Test

Section	Task name	Description
Listening	Listen and Reply	Listen to an input sentence and select an appropriate response
	Listen to a Conversation	Understand and draw inferences from a brief conversation between two speakers
	Academic Listening: Announcements	Understand and draw inferences from a brief informational announcement in an academic setting
	Academic Listening: Talks	Understand and draw inferences from an expository monologue on an academic topic
Reading	Vocabulary	Select the synonym to a given word
	General Reading: Daily Life	Understand and draw inferences in brief and/or nonlinear texts from daily life
	Academic Reading: Tables	Understand and draw inferences regarding academic information in a table
	Academic Reading: Passages	Understand and draw inferences from an academic text
Writing	Build a Sentence	Given an input sentence, combine words/phrases to make a grammatical sentence in response
	Describe a Photo	Describe the contents of a photo for a social media post
	Write an Email	Given a scenario, write an email to accomplish a specified purpose
	Write for an Academic Discussion	Contribute to an online class discussion on an academic topic
Speaking	Read Aloud	Read aloud a part in a multiturn dialog
	Listen and Repeat	Repeat a sequence of sentences of increasing length
	Virtual Interview	Answer a video-recorded interviewer's questions regarding personal views and experiences

diversity of educational and geographic settings; these individuals were experts in the Canadian educational context and the use of the CLB and had extensive teaching experience. Prior to standard setting meetings, the panelists completed a series of homework activities to help ensure familiarity with the descriptions of all 12 CLB levels as well as the content of the TOEFL Essentials test. The standard setting meetings were conducted remotely via Microsoft Teams and facilitated by the project team following recommended standard setting methodology. The outcome of the standard setting study was a recommended set of cut scores, indicating the minimum test score needed to classify a test taker at each CLB level.

Following the standard setting study, we compared the resulting panel-recommended cut scores to the findings of the construct congruence study. Cut scores were also considered in light of the correspondence between the CEFR and CLB levels established by North and Piccardo (2019), given that TOEFL Essentials test scores had already been mapped to the CEFR levels (Papageorgiou et al., 2021). The project team also analyzed data to provide three types of validity evidence for the panel-recommended cut scores: procedural, internal, and external. (For a detailed discussion of the different types of validity evidence for standard setting, see Council of Europe, 2009; Hambleton et al., 2012; and Tannenbaum & Cho, 2014). Upon analysis of all data, the project team established the final score mapping presented in this report.

Construct Congruence Study

To inform interpretations regarding the relationship between the TOEFL Essentials test and the CLB, we carried out an evaluation of the extent to which the test content and scoring reflected the descriptions of language proficiency provided in the CLB, that is, the degree of construct congruence (Tannenbaum & Cho, 2014). However, given inherent differences in scope and detail, this evaluation of construct congruence should not be taken as a point-by-point comparison of TOEFL Essentials and the content of the CLB. Rather, in this instance we made holistic judgments regarding the degree of match between the abilities tested by TOEFL Essentials and those specified in the CLB.

The CLB provides several different types of information regarding abilities expected at each proficiency level. The three CLB sections oriented toward broad descriptions of language ability were used for evaluation, in keeping with the design of TOEFL Essentials as a test of general English language proficiency in academic and general (daily life) contexts. Specifically, content was evaluated from the following CLB sections:

- profiles of ability,

- features of communication, and
- knowledge and strategies.

Descriptions in the profiles of ability are further divided into three elements, which were considered separately in the analysis:

- General statement of ability, prefaced with "the (listener/speaker/reader/writer) can"
- Descriptions of communicative context, prefaced with "when the communication is"
- Descriptions of language performance, prefaced with "demonstrating these strengths and limitations"

The features of communication provide additional information regarding the types of language, texts, and situations a learner at a given level is able to navigate. Finally, the section for knowledge and strategies describes some of the language knowledge and strategies that an individual might need to learn within one of three stages, where each stage is defined as progression of four CLB levels, that is, Stage I covers CLB levels 1–4.

Judgments took into account both the number of descriptors to which test content could be aligned and the degree of coverage of individual descriptors. The congruence of TOEFL Essentials content to CLB descriptors was coded on a four-category scale consisting of *extensive*, *substantial*, *partial*, and *negligible* coverage. The extensive category was assigned when test content covered most or all of the description, the substantial category was used when a majority of description was covered, the partial category indicated a minority of the description was covered, and negligible was used when there was little coverage. Two ETS staff carried out the review, with judgments determined through consensus. One individual was a senior research scientist and the other was an assessment designer; both were involved in the development of the test and were highly familiar with the content of the test and the language constructs measured. The results of this coding exercise are reported in the following sections, organized by language skill (speaking, writing, listening, and reading).

Speaking

In the speaking section, the Virtual Interview task provides direct evidence of the ability to communicate and therefore serves as the primary basis for mapping TOEFL Essentials content to CLB descriptors regarding communicative contexts and functions. The other speaking tasks, Read Aloud and Listen and Repeat, elicit evidence of underlying aspects of language ability and provide supporting evidence of the ability to efficiently and accurately process language and produce intelligible speech across a range of language proficiency levels.

Profile of Ability and Features of Communication

Coverage of the TOEFL Essentials Speaking section across CLB levels for profile of ability and features of communication is summarized in Table 2. In the Virtual Interview task, test takers are initially asked about their concrete personal experiences and then answer questions on increasingly abstract topics, expressing their opinions, suggestions, predictions, and so forth. The lowest levels of the CLB require the ability to communicate very basic information in limited language; such performance is expected on the initial interview questions, where responses of this type would receive scores of 1 or 2. The Read Aloud and Listen and Repeat tasks also provide scaffolding for beginning-level speakers in the form of input to be read or repeated, respectively, in keeping with expectations that learners at CLB Stage I will require support to communicate.

Higher CLB levels require the ability to produce descriptions of personal experiences and familiar topics in routine social situations, which aligns relatively closely with the initial questions in the Virtual Interview task. In Stage II of the CLB (Levels 5–8), there is an expectation of dealing with increasingly abstract and challenging topics, in more challenging situations. Questions on abstract topics occurring near the end of the Virtual Interview task correspond well to these expectations.

The highest CLB levels (CLB 9–12, Stage III) add an increasing range of communicative situations and content, including complex and specialized topics. Although it is not possible to fully assess this range and depth of situations in a single assessment, strong performance on the Virtual Interview task is expected to be predictive of a general ability to communicate in demanding situations. In addition, the Listen and Repeat task provides evidence of an underlying ability for online (immediate) language processing, simultaneously engaging the formulation, articulation, and self-monitoring steps of the speech production process (Levelt, 1989). Research also supports the use of the Listen and Repeat task (also known as

Table 2 Coverage of the TOEFL Essentials Speaking Section for Canadian Language Benchmarks (CLB) Profiles of Ability and Features of Communication

	Profiles of ability			
Level	The speaker can:	When the communication is:	Demonstrating these strengths and limitations:	Features of communication
CLB 1	Substantial	Substantial	Extensive	Substantial
CLB 2	Extensive	Substantial	Extensive	Substantial
CLB 3	Extensive	Extensive	Extensive	Extensive
CLB 4	Extensive	Extensive	Extensive	Extensive
CLB 5	Extensive	Extensive	Extensive	Extensive
CLB 6	Extensive	Extensive	Extensive	Extensive
CLB 7	Extensive	Extensive	Extensive	Substantial
CLB 8	Extensive	Extensive	Extensive	Substantial
CLB 9	Substantial	Substantial	Extensive	Partial
CLB 10	Substantial	Substantial	Substantial	Partial
CLB 11	Partial	Substantial	Substantial	Partial
CLB 12	Partial	Substantial	Substantial	Partial

Table 3 Coverage of the TOEFL Essentials Speaking Section for Canadian Language Benchmarks (CLB) Knowledge and Strategies

CLB stage	Grammatical knowledge	Textual knowledge	Functional knowledge	Sociolinguistic knowledge	Strategic competence
Stage I (1–4)	Extensive	Extensive	Substantial	Partial	Partial
Stage II (5–8)	Extensive	Extensive	Partial	Partial	Partial
Stage III (9–12)	Extensive	Partial	Negligible	Partial	Partial

elicited imitation) as a measure of global oral language ability (Davis & Norris, 2021; Kostromitina & Plonsky, 2021), and strong performance should be predictive of advanced speaking ability as defined in the CLB.

Knowledge and Strategies

In terms of the knowledge and strategies specified in the CLB, the responses elicited by TOEFL Essentials Speaking tasks elicit evidence of grammatical knowledge across the full CLB spectrum, evident in both spontaneous language production and the ability to understand, process, and accurately reproduce written and spoken input (Table 3). Evidence of the ability to effectively organize discourse (textual knowledge) is elicited in the interview task, where test takers must create cohesive accounts of their experiences and opinions. Speaking tasks provide partial coverage of certain aspects of functional and sociolinguistic knowledge. Although all of the tasks in TOEFL Essentials Speaking section are embedded in communicative scenarios, it is the Virtual Interview task that highlights the ability to communicate appropriately and spontaneously in a specific situation.

Writing

The writing section of the TOEFL Essentials test includes tasks with differing language and pragmatic demands. Test takers first complete a series of short, scaffolded writing tasks, where supplied words or phrases must be placed in the correct order to form a grammatical and appropriate response to a prompt by a simulated interlocutor (the *Build a Sentence* task). They then proceed to open-ended writing tasks, which may include writing a brief description of a picture for a social media post (*Describe a Photo*), producing an email within a daily life or academic setting (*Write an Email*), or contributing to a classroom discussion carried out in an online forum (*Write for an Academic Discussion*). The Build a Sentence task indicates the ability to use a variety of structures to communicate meaning. Scoring of spontaneous writing in the other tasks is based on the ability of the test taker to produce texts that are clear, cohesive, elaborated, and appropriate for the communicative purpose.

Table 4 Coverage of the TOEFL Essentials Writing Section for Canadian Language Benchmarks (CLB) Profiles of Ability and Features of Communication

	Profiles of ability			
Level	The writer can:	When the communication is:	Demonstrating these strengths and limitations:	Features of communication
CLB 1	Substantial	Substantial	Substantial	Substantial
CLB 2	Substantial	Substantial	Substantial	Extensive
CLB 3	Extensive	Extensive	Extensive	Extensive
CLB 4	Extensive	Extensive	Extensive	Extensive
CLB 5	Extensive	Extensive	Extensive	Substantial
CLB 6	Extensive	Extensive	Extensive	Substantial
CLB 7	Extensive	Extensive	Extensive	Substantial
CLB 8	Extensive	Extensive	Extensive	Substantial
CLB 9	Extensive	Substantial	Extensive	Partial
CLB 10	Substantial	Partial	Substantial	Partial
CLB 11	Partial	Partial	Partial	Partial
CLB 12	Negligible	Negligible	Partial	Partial

Profile of Ability and Features of Communication

The coverage of CLB levels by TOEFL Essentials Writing tasks by profile of ability and features of communication is shown in Table 4. Learners at the lower CLB levels (Stage I, CLB 1–4) are expected to use words, phrases, and simple sentences to briefly communicate personal information (CLB 1–2) and produce sentences and brief texts related to personal experiences and everyday situations (CLB 3–4). The less difficult examples of the scaffolded writing task (Build a Sentence) require knowledge of simple sentence structures, and the Describe a Photo task elicits descriptions of everyday situations or (imagined) personal experiences. These tasks are perhaps somewhat more advanced than the very basic communicative demands of the lowest CLB levels, but it is expected that beginning learners should be able to attempt a response. A response consisting of the isolated words and phrases expected at CLB 1–2 would receive a score of 1 on the 0–5 scale used to score the task. Performance showing developing control of simple language (CLB 3–4) would be characteristic of responses receiving scores of 2 or 3.

At intermediate levels (Stage II, CLB 5–8), writers are expected to create texts of increasing complexity, abstraction, appropriateness, and linguistic sophistication, on familiar topics relevant to daily life and academic study. The Write an Email and Write for an Academic Discussion tasks provide an opportunity to demonstrate such abilities, with the former task oriented toward concrete topics and socially appropriate communication and the latter task oriented toward abstract topics and academic register.

At the highest levels (Stage III, CLB 9–12), writers are expected to effectively communicate through an expanding range of demanding, specialized, and lengthy genres. Due to practical limitations, it is not possible for the writing section to include written texts of the length or breadth that characterize the highest CLB levels. However, production of syntactically complex sentences in the scaffolded writing task and production of clear, appropriate, and well-elaborated spontaneous writing under timed conditions provides evidence of the critical writing skills required to create complex and effective texts. Performance on these tasks is therefore expected to be predictive of writing ability at the higher CLB levels.

Knowledge and Strategies

In regard to the knowledge and strategies expected for writing, TOEFL Essentials tasks elicit considerable evidence of the ability to use a range of vocabulary, grammar, and textual conventions to effectively communicate (grammatical knowledge), as well as the ability to produce well-organized and cohesive texts (textual knowledge, Table 5). Writing tasks also elicit evidence of sociolinguistic knowledge, in that test takers must produce texts appropriate for a variety of different contexts and communicative purposes. Functional knowledge related to the content of basic written genres (e.g., email) is also covered in the writing section. However, at higher stages, functional knowledge is defined in terms of a broad array of genres and contexts (e.g., taking messages, completing forms) that are not possible to fully cover in a language proficiency assessment such as the TOEFL Essentials test. In the CLB framework, strategic competence is largely defined in terms

Table 5 Coverage of the TOEFL Essentials Writing Section for Canadian Language Benchmarks (CLB) Knowledge and Strategies

CLB stage	Grammatical knowledge	Textual knowledge	Functional knowledge	Sociolinguistic knowledge	Strategic competence
Stage I (1–4)	Extensive	Extensive	Extensive	Partial	Negligible
Stage II (5–8)	Extensive	Extensive	Partial	Substantial	Partial
Stage III (9–12)	Extensive	Substantial	Partial	Substantial	Partial

Table 6 Coverage of the TOEFL Essentials Listening Section, for Canadian Language Benchmarks (CLB) Profiles of Ability and Features of Communication

	Profiles of ability			
Level	The listener can:	When the communication is:	Demonstrating these strengths and limitations:	Features of communication
CLB 1	Substantial	Substantial	Substantial	Substantial
CLB 2	Extensive	Substantial	Substantial	Substantial
CLB 3	Extensive	Substantial	Substantial	Extensive
CLB 4	Extensive	Extensive	Extensive	Extensive
CLB 5	Extensive	Extensive	Extensive	Substantial
CLB 6	Extensive	Extensive	Extensive	Substantial
CLB 7	Extensive	Extensive	Extensive	Substantial
CLB 8	Extensive	Extensive	Extensive	Substantial
CLB 9	Substantial	Substantial	Substantial	Partial
CLB 10	Substantial	Substantial	Substantial	Substantial
CLB 11	Substantial	Substantial	Partial	Partial
CLB 12	Partial	Partial	Partial	Partial

of process-writing features and use of writing aids such as dictionaries or word processors. TOEFL Essentials Writing tasks will provide evidence of the ability to generate ideas and create a draft under timed conditions; individuals who are more effective in these aspects of the writing process are expected to perform better on the test tasks, but the test does not evaluate the writing process per se.

Listening

In the listening section of TOEFL Essentials, test takers must be able to select an appropriate response to a short conversational turn (the Listen and Reply task) and identify or infer information provided in a variety of contexts, including short conversations (Listen to a Conversation), announcements (Academic Listening: Announcements), and extended expository monologues (Academic Listening: Talks). The conversational tasks take place in daily life contexts, whereas the academic tasks mimic communications related to academic life (Academic Listening: Announcements) or discuss general topics of an academic nature (Academic Listening: Talks).

Profile of Ability and Features of Communication

At the lowest levels (CLB 1–2), listeners are expected to comprehend a limited range of words and expressions, in brief input of a few words to a few sentences. The conversational listening tasks provide such input, and it is expected that low proficiency individuals will be able to answer easier items through recognition of single words or phrases. Overall, conversational listening tasks fit well with the expectations for Stage I (CLB 1–4), where leaners can understand simple, brief, spoken input on everyday topics (Table 6).

At Stage II (CLB 5–8), learners are expected to understand input of increasing length and complexity on an increasing range of topics related to daily life and general knowledge. Learners progressing through CLB levels 5–8 are also expected to increasingly comprehend main ideas, details, and implied meanings; understand an expanding range of concrete and abstract language; and decode more complex structures. These language demands align relatively closely with the design of the academic speaking tasks of the TOEFL Essentials test, which vary in language demands and also feature a range of common situations and general topics.

Table 7 Coverage of the TOEFL Essentials Listening Section, for Canadian Language Benchmarks (CLB) Knowledge and Strategies

CLB stage	Grammatical knowledge	Textual knowledge	Functional knowledge	Sociolinguistic knowledge	Strategic competence
Stage I (CLB 1–4)	Extensive	Extensive	Substantial	Substantial	Partial
Stage II (CLB 5–8)	Extensive	Extensive	Substantial	Partial	Substantial
Stage III (CLB 9–12)	Extensive	Substantial	Substantial	Negligible	Substantial

At Stage III (CLB 9–12), the learner is expected to understand ever more complex language, identify bias or other subtle aspects of meaning, and comprehend texts from a range of complex genres of increasing length. TOEFL Essentials academic listening tasks of higher difficulty require comprehension of increasingly complex vocabulary and grammar as well as the ability to identify a variety of implied meanings. However, it was not feasible to include the range or length of listening tasks listed at the highest CLB levels. Excellent performance on the more difficult listening test tasks is expected to be predictive of a general ability to understand complex language, which will support comprehension of the types of listening texts described in CLB Stage III. However, the TOEFL Essentials test does not directly measure performance on such tasks.

Knowledge and Strategies

In terms of the knowledge and strategies specified in the CLB, TOEFL Essentials Listening tasks provide relatively robust coverage of the ability to understand complex language (grammatical knowledge) and identify devices for producing cohesive text (textual knowledge, Table 7). The CLB categories of functional knowledge and sociolinguistic knowledge cover the ability to understand the conventions of an increasing range of communicative genres and registers, including culturally specific references; norms for interaction; or texts such as jokes, songs, and stories. An understanding of conventions is required for successful performance on the tasks used in the test, but again, it is not feasible to assess the full range of genres suggested at the higher levels. For sociolinguistic knowledge, coverage is further limited by fairness concerns, given that the TOEFL Essentials test is administered to an international candidature with widely varying opportunities to acquire Canadian norms for communication. Strategic competence is defined in terms of the ability to correct misunderstandings (Stage I) or infer and interpret meaning (Stages II and III). Repair strategies are sometimes noted in the speaking section of the test, whereas the ability to infer meaning is required throughout the listening section.

Reading

In the reading section of the TOEFL Essentials test, test takers must identify stated and implied meanings from a variety of written texts. Test takers also demonstrate vocabulary knowledge by identifying synonyms of words (the Vocabulary task). Written texts cover content from both daily life (General Reading: Daily Life) and academic domains (Academic Reading: Tables and Academic Reading: Passages). Daily life passages simulate a variety of commonly encountered texts, including notices, labels, forms, instructions, schedules, advertisements, social media posts, emails, and so on. Formatting and graphical elements in these readings also mimic real-life texts and provide support for navigating the information presented. The passages in the Academic Reading: Tables task consist of informational statements presented as bullets in a table, oriented toward developing readers. The Academic Reading: Passages task consists of short expository texts (approximately 200 words) with topics and language that might be found in secondary and higher education.

Profile of Ability and Features of Communication

As shown in Table 8, reading tasks in the TOEFL Essentials test, specifically Vocabulary; General Reading: Daily Life; and Academic Reading: Tables, correspond well with the descriptions of reading ability at lower CLB levels (Stage I, CLB 1–4), where beginning learners are expected to recognize words and simple phrases (CLB 1), eventually progressing to an ability to "get most information from short, simple texts related to familiar, routine everyday topics" (CLB Level 4, Centre for Canadian Language Benchmarks, 2012, p. 74). Easier items of the General Reading: Daily Life task, such as finding information on a schedule, require only recognition of limited words and phrases, whereas intermediate difficulty items

Table 8 Coverage of the TOEFL Essentials Reading Section for Canadian Language Benchmarks (CLB) Profiles of Ability and Features of Communication

	Profiles of ability			
Level	The reader can:	When the communication is:	Demonstrating these strengths and limitations:	Features of communication
CLB 1	Extensive	Extensive	Extensive	Extensive
CLB 2	Extensive	Extensive	Extensive	Extensive
CLB 3	Extensive	Extensive	Extensive	Extensive
CLB 4	Extensive	Extensive	Extensive	Extensive
CLB 5	Extensive	Extensive	Extensive	Extensive
CLB 6	Extensive	Extensive	Extensive	Extensive
CLB 7	Extensive	Extensive	Extensive	Extensive
CLB 8	Extensive	Extensive	Extensive	Extensive
CLB 9	Extensive	Extensive	Extensive	Substantial
CLB 10	Extensive	Extensive	Extensive	Substantial
CLB 11	Substantial	Substantial	Substantial	Substantial
CLB 12	Substantial	Substantial	Substantial	Substantial

Table 9 Coverage of the TOEFL Essentials Reading Section for Canadian Language Benchmarks (CLB) Knowledge and Strategies

CLB stage	Grammatical knowledge	Textual knowledge	Functional knowledge	Sociolinguistic knowledge	Strategic competence
Stage I (1–4)	Extensive	Extensive	Extensive	Extensive	Substantial
Stage II (5–8)	Extensive	Extensive	Extensive	Substantial	Substantial
Stage III (9–12)	Substantial	Extensive	Partial	Substantial	Substantial

require broader vocabulary and the ability to extract information from longer and more complex texts such as recipes or magazine articles.

Learners at CLB Stage II (CLB 5–8) are expected to understand a range of increasingly complex factual texts of short to moderate length from a variety of work, study, or social situations. Learners also develop the ability to comprehend a range of information from the text, including the writer's purpose, main idea, details, and implied meanings, and at higher levels, manipulate this content in various ways, such as through integration of information or comparison and contrast. TOEFL Essentials daily life and academic reading tasks measure these skills using multiple choice items that target these different elements of comprehension.

At the highest CLB levels (Stage III, CLB 9–12), readers are expected to comprehend increasingly long, complex, and specialized texts; identify various types of stated and implied meanings; recognize the author's purpose and stance; critically evaluate content; and understand idiomatic and figurative language. The most difficult daily life and academic reading items from the TOEFL Essentials test make similar demands, although using a narrower range of shorter texts. Successful performance provides evidence of the general literacy skills that support performance at CLB Stage III and should be predicative of the ability to comprehend longer and more varied texts. However, for reasons of practicality it is not possible to directly assess the range and length of text types mentioned at the highest levels of the CLB.

Knowledge and Strategies

For the knowledge and strategies portion of the CLB framework, the TOEFL Essentials Reading section provides good evidence of the ability to decode vocabulary, syntax, and mechanics relevant to the category of grammatical knowledge (Table 9). The reading section also assesses the ability to interpret strategies for organizing information and establishing cohesion of content (textual knowledge), as well as recognizing the author's purpose and how information is formatted in common genres (functional knowledge). However, coverage of functional knowledge is somewhat more limited at the highest level (Stage III), where knowledge of an increasingly wide range of texts and cultural content is required. Similarly, reading tasks are consistent with expectations for sociolinguistic knowledge and strategic competence at the lower CLB levels (Stage I), where recognition of social meanings and basic strategies for comprehending text are considered. Coverage is somewhat less extensive at higher levels, where individuals are expected to comprehend an expanding range of culturally specific writing and use a variety of tools to decode and manipulate text.

Standard Setting Study

The evaluation of construct congruence presented in the previous section indicated that the abilities measured by the TOEFL Essentials test were generally consistent with language proficiency as described in the CLB. This finding provided justification for carrying out a standard setting study to establish the relationship between test scores and the levels of the CLB. This relationship is operationalized through cut scores, which define the boundaries of specific CLB levels.

General Procedures for Setting Cut Scores

A cut score indicates the point on the test score scale that separates examinees who have demonstrated a specific level of performance from those who have not. Cut scores are established with a well-researched process called *standard setting* (Cizek & Bunch, 2007). During standard setting, a panel of experts is typically required, under the guidance of one or more meeting facilitators, to make judgments about the difficulty of test questions (items or tasks). The outcome of the standard setting meeting is a set of cut score recommendations to the examination provider. Statistical information about the test (e.g., item difficulty estimates and distribution of test scores) is also used to help panelists with their judgment task. A fairly common practice in standard setting meetings is to conduct more than one round of judgments. Between rounds, the panel reviews and discusses individual judgments from the previous round and may receive statistical information about items as well as the distribution of current cut scores. Following this review, the panel then repeats the judgments. After the standard setting exercise, the recommendations of the panel are evaluated in light of other relevant evidence that might be available, and final cut scores are set. Validity evidence is also collected during the standard setting meeting, as we discuss later in this report.

In the current study, standard setting workshops were conducted over 2 weekends, with 2 weeks in between, during the second half of January 2022. A full-day session was devoted to each section of the test, with the speaking and writing sections covered on Friday and Saturday of the first weekend, and listening and reading covered on Friday and Saturday of the second weekend (see Appendix A for the full schedule). In the week prior to each weekend workshop, the panelists individually completed preparation activities, as described in the section on panelist preparation below. Standard setting meetings were conducted online using Microsoft Teams, where text and materials could be presented to the panel and panelists could engage with each other through video and text comments. Details of each standard setting meeting are provided in subsequent sections.

Empirical Data from the TOEFL Essentials Test

In preparation for specific steps during the standard setting meetings described in subsequent sections, members of the project team collaborated with assessment developers and psychometricians at ETS to collect item-level response data and score distribution information. For the reading and listening sections, item difficulty measures were based on the empirical ability estimates derived from item-response theory, for over 5,500 students who took the field test (Papageorgiou et al., 2021). The distribution of test scores from the field test was also obtained for all four test sections (5,599 test takers for listening, 5,998 test takers for reading, 4,599 test takers for writing, and 4,244 test takers for speaking). Responses were obtained from the field test administration because this source of data was felt to be most representative of the expected test-taking population for the TOEFL Essentials test, where a great deal of effort was made to recruit participants that would reflect the demographic characteristics of likely test takers. Moreover, at the time of the study, the operational test was in the initial stages of administration, and the composition of the TOEFL Essentials candidature had yet to stabilize.

Selection of Panelists

A total of 17 educators who were based in Canada served as panelists, including nine females and eight males. T he panelists represented a variety of institutions and provinces (see Appendix B). All potential panelists responded to a background questionnaire prior to the standard setting meetings, and the project team selected panelists based on their experience teaching English as a second or foreign language as well as their familiarity with the CLB. At the time of the study, the panelists indicated that they were working at the following educational institutions:

- College or university (14 panelists)
- English language training institution (three panelists)

All panelists were experienced English language teachers with at least 7 years of teaching experience. Eight of the 17 panelists had more than 10 years of teaching experience, and another eight had taught more than 20 years. Aside from teaching, 16 panelists had experience developing learning materials or assessments for learners of English as a second or foreign language.

The panelists indicated some level of familiarity with the CLB in the background questionnaire. Nine panelists said they were very familiar with it, four indicated they were familiar, and another four were somewhat familiar. Regardless of their familiarity level with the CLB, the project team provided all panelists with the same preparation materials prior to the standard setting meeting.

Regarding the panelists' familiarity with the TOEFL Essentials test, only one panelist indicated that they were very familiar with the test content. Nine panelists said that they were familiar with the test, with three saying they were somewhat familiar, and two panelists indicating they were a little bit familiar. Two panelists were not familiar with the test at all. It should be noted that although three of the 17 panelists had prior experience with standard setting, most of the panelists were not familiar with either standard setting or the specific methods used in this study. Therefore, all panelists went through the same training activities.

Panelist Preparation Prior to the Standard Setting Workshop

Prior to each weekend of standard setting meetings, a preparation guide was sent to the panelists; one guide was for speaking and writing, and a second guide was for listening and reading. T he guide included information about the CLB and the TOEFL Essentials test as well as familiarization activities targeting the CLB levels. All panelists were asked to complete two familiarization activities to help ensure that they had a good understanding of the features that distinguished each of the 12 CLB levels. In the first activity, panelists were presented with descriptors drawn from the CLB and asked to sort the descriptors into the appropriate stage, and then level. In the second activity, panelists were asked to consult the full set of CLB descriptors provided in the preparation guide and list three to five distinguishing features for each CLB level that separated the level from the adjacent levels. 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.

Additionally, prior to the online standard setting meetings, all panelists signed a nondisclosure–confidentiality agreement and watched videos of all four sections of the test. The videos simulated the test-taking experience so that the 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. Familiarizing the panelists with the test content prior to the standard setting meetings was deemed necessary because several panelists indicated in their background questionnaire that they were not very familiar with the test content.

Borderline Student Definition

Each standard setting meeting started with a review of the first familiarization activity, where panelists had a chance to discuss any challenges in distinguishing the descriptors across CLB stages and levels. Following this discussion, panelists worked together to define the minimum language skills needed to reach each of the CLB Levels 1–12. This effort was informed by the second familiarization activity, where panelists had noted distinguishing features at each level. A student (test taker) with the identified minimally acceptable skills was defined as the *just qualified candidate* (JQC) for the given level. These JQC descriptions served as the frame of reference for subsequent standard setting tasks for each test section.

Given the large number of levels in the CLB, the descriptions of JQCs were broken into several steps. First, the entire panel compiled key features for the JQCs of the Intermediate stage (Stage II, CLB 5–8) in a group discussion facilitated by the project team. Two subpanels were then formed, with one panel focused on the JQCs for the Basic stage (Stage I, CLB 1–4) and the other panel on the JQCs for the Advanced stage (Stage III, CLB 9–12). Each subpanel was asked to present the suggested JQCs to the other subpanel to help ensure that agreement was reached on the definitions. The panelists were resorted into subpanels on each day to encourage collaboration across all individuals over the course of the standard setting exercise. Although the panel was split into two groups to identify JQCs for CLB Stage I and Stage III, all subsequent discussions of cut scores and other activities were done as a single group.

The above steps were considered an efficient way to develop the JQC definitions for such a large number of levels while allowing for adequate discussions of the JQC features among the panelists. The borderline student definitions developed by the panelists can be found in Appendix D.

Standard Setting Method for Constructed-Response Items

For the test sections containing constructed-response items (speaking and writing), a variation of the Performance Profile method (Fleckenstein et al., 2020; Hambleton et al., 2000) was selected because it allows panelists to review a set of student performance samples. Such a review was relevant to the panelists' professional expertise as educators, where it is common to make judgments about samples of actual student work in a holistic fashion (Kingston & Tiemann, 2012). Standard setting for constructed-response items was completed during the first weekend of meetings, first for the speaking section and then for the writing section.

For speaking, panelists reviewed the responses of individuals who had participated in the TOEFL Essentials field test. A sample of 33 test takers was drawn from a subgroup of 1,300 field test participants who had completed the test at the time of sampling and had produced a full set of 16 scorable responses. Although this sample was not drawn from the full set of field test participants, it was used because the same sample had been used in an earlier study to map the speaking section of the TOEFL Essentials test to the CEFR levels. Using the same sample meant that raw cut scores could be directly compared to those identified for the CEFR levels, produced using the same data. Individuals were selected to represent even-numbered raw scores for the speaking section of the field test, including one additional individual who achieved a perfect score of 75 and excluding scores of 10 or less, where no test taker completed a full set of scorable responses. For each test taker, a portfolio was created that included two responses to the Read Aloud task, three responses to the Listen & Repeat task, and two responses to the Virtual Interview task.

Three rounds of judgments occurred with feedback and discussion between rounds (see sample of the rating form in Appendix E). The judgment task was presented as follows: "What speaking score would a JQC at a given CLB level earn?" Prior to Round 1, the panel practiced the standard setting technique by attempting to identify the JQC for CLB Level 7. A researcher played audio recordings of responses, moving up and down the list of test takers as requested by the panelists, who attempted to identify the test taker who best fit the characteristics of the JQC at CLB 7 as identified earlier in the meeting. At this time, panelists were also shown how to enter their cut scores in the spreadsheet used for reporting. Following training, each panelist then completed an evaluation form indicating the extent to which the procedure for making judgments was clear and whether the panelist was ready to proceed; all panelists indicated readiness to go on. Round 1 followed the same process of playing responses as directed by the panelists, considering each JQC in turn. Although responses from all test takers were played, this round did not necessarily include every response for each test taker. Specifically, responses to the Virtual Interview task were reviewed more thoroughly, given the communicative focus of this task type. In Rounds 2 and 3, various responses were replayed when panelists indicated they wanted to review a particular level or task. The recommendations for the speaking and writing cut scores were based on the final round of judgments (Round 3).

For writing, a total of 39 responses were selected from a subgroup of 1,963 participants in the field test. As was done for the speaking section, a subgroup of field test participants was used so that the sample of responses presented to panelists would be identical to the sample used in the previous CEFR mapping study. It should also be noted that field test participants completed four spontaneous writing items (Describe a Photo, Write for an Academic Discussion, and both brief and extended versions of Write an Email), unlike the operational TOEFL Essentials test where a given test taker completes only two of these tasks. For each possible whole-numbered score for the field test writing section (0–20, four items scored 0–5), two individuals were selected. Only one participant with a total score of 1 was available, to make a final sample of 39 individuals. Panelists were provided with a document containing the four written responses produced by each test taker.

As for speaking, panelists had the opportunity to practice the standard setting method, which followed a similar procedure except that panelists could independently access the test takers' written responses. Following standard setting training, the panelists individually reviewed this document to make their initial cut score judgments (Round 1) and could refer to the document during subsequent group discussion and decision-making (Rounds 2 and 3). The Build a Sentence task of the writing section was not reviewed by the panelists because of its selected-response format, but performance on this task was considered at a later point by the project team when finalizing the score mapping.

To make cut score recommendations for Round 1, panelists were asked to review the JQC descriptions for CLB 1, CLB 3, CLB 5, CBL 7, CLB 9, CLB 10, and CLB 12. The task in this method was to review the test takers' spoken or written responses and decide the speaking and writing section scores a borderline student at each of these CLB levels would receive. After Round 1, the panel's mean, median, and mode and the minimum and maximum cut scores recommended for each cut score were presented, and panelists shared their judgment rationales. Impact data were also shown to inform panelists about the percent of students from the field test who would be classified into each CLB level. The process was repeated for Round 2, but this time the panelists entered cut scores for all 12 CLB levels. Panelists had the opportunity to further review responses, as desired, to inform their Round 2 judgments. Following discussion of the Round 2 cut scores, a final, third round of cut score judgments were made.

At the end of the second day of meetings, panelists completed an evaluation form that collected their perceptions of the standard setting process for speaking and writing, the importance of various factors in the process, and which factors influenced their judgments. Panelists were also asked to indicate their level of confidence in the final set of recommended cut scores for speaking and writing.

Standard Setting Method for Selected-Response Items

For the two test sections consisting of selected-response items (listening comprehension and reading comprehension), a variation of the Item-descriptor Matching method (Ferrara et al., 2008), known as the *basket method* (Council of Europe, 2009), was followed. The basket method was chosen because panelists simply identify the level that would correspond to the minimum ability required to successfully answer the item. This approach was seen as an efficient way to navigate the sizable number of cut score decisions to be made, given the large number of CLB levels. Standard setting for selected-response items was completed during the second weekend of meetings, with the listening section on the first day and the reading section on the second day.

For listening and reading, panelists reviewed items taken from a single full operational form of the TOEFL Essentials test, including the first stage router and the three panels of items used in Stage Two, which were designed to be of low, medium, and high difficulty, respectively. (In operational use, a given test taker encounters only one of the Stage Two panels.) Item content was presented in PDF format using screenshots from test administration, with items presented in the order that they would appear on the test. Listening items also included a script of the audio used a listening input, and audio/video of each listening item was shown to panelists. Items used to establish overlap between panels, which appear more than one time in the full test form, were judged only once. This omission of duplicate items resulted in a total raw score scale of 0–54 for listening and 0–53 for reading.

Following the development of the JQC definitions as described in the Borderline Student Definition subsection, panelists were trained in the basket method and given an opportunity to practice the standard setting task. During this practice, the panelists were asked to individually identify the minimum CLB level required to successfully answer each of three listening items. They then discussed the rationale behind their judgments. The project team provided clarification on the procedure as needed. Each panelist then completed an evaluation form indicating the extent to which the training was clear and whether the panelist was ready to proceed; all panelists indicated their readiness and so went on to independently review the test items and record their judgments on a rating form.

The basket method was implemented in three rounds of judgments informed by feedback and discussion between rounds. In Round 1, panelists were asked to decide the minimum CLB level a test taker would need to attain in order to answer a test item correctly. The question for the panelists was stated as follows: "At what CLB level can a test taker answer each test question correctly?" The panelists selected one level for each test item and entered their item-level judgments on a rating form created in Microsoft Excel, which automatically computed a recommended cut score for each CLB level (see Appendix F for a sample). This recommended cut score was calculated as the sum of the items requiring knowledge at that CLB level along with all items requiring knowledge at lower CLB levels. The panelists were instructed to focus only on the alignment between the English language skills demanded by the test item and the English language skills possessed by the JQCs at each level, and not to factor random guessing into their judgments.

Following the first and the second round of judgments, the results were summarized and shown to the panelists. The number of panelists who chose a specific CLB level for each test item was presented, followed by discussion. In addition, an item difficulty measure was provided for each item, presented as the percent of test takers expected to answer the item correctly. Although presented in terms of percentage of test takers, this metric was based on the item difficulty as estimated using item response theory (IRT), to account for the fact items were spread across multiple forms used in the field test and so were taken by differing groups of test takers. Panelists were instructed to use the item difficulty values as a guide when considering the relative difficulty of the test items, not as an indicator of the CLB level needed to get an item correct. In Round 3, panelists were asked to make holistic judgments, that is, to provide one cut score recommendation for each CLB level instead of item-level judgments (see Appendix F for a sample of the Round 3 rating form). To facilitate final cut score recommendations, panelists were asked to review the automatically calculated Round 2 cut scores and then adjust these scores as needed to produce their final Round 3 judgments. The transition to a holistic-level judgment placed emphasis on the overall language skill of interest (i.e., reading comprehension or listening comprehension). Upon completion of Round 3, panelists were shown a summary of the results and were asked to discuss the reasonableness of the average cut score for each level.

As with the speaking and writing meetings, at the end of the second day panelists completed a final survey of their perceptions of the standard setting process, the importance of various factors in the process, and which factors influenced their judgments. Panelists were also asked to indicate their level of confidence in the final set of recommended cut scores.

Results of the Standard Setting Judgments

In this section, we summarize the panel's cut score recommendations by round of judgments for each of the test sections. The results include the mean, median, minimum, maximum, and standard deviation (*SD*) of each round of judgments. The mean cut scores in the final round of judgments for each test section are considered the panel's final recommendations. The results are presented in raw scores because the project team did not want to add to the panelists' cognitive load by introducing technical information about the conversion from raw scores to reported band scores.

The results for the speaking section are presented in Table 10. The mean cut score for each CLB was similar across rounds. On the raw score scale of 0–75 (16 responses, each scored on either a 0–4 or a 0–5 scale), the largest change was 1.6 score points for the cut score of CLB 1. The standard deviation tended to decrease, suggesting convergence in the judgments.

The results for the writing section are presented in Table 11. The mean cut scores across rounds were very similar, as there were only 20 raw score points available (four responses, each scored using a 0–5 scoring rubric). The variability in panelists' judgments decreased in general across rounds, as can be seen by the standard deviation.

Table 10 Standard Setting Results for the Speaking Section of the TOEFL Essentials Test

Round	Statistic	CLB 1	CLB 2	CLB 3	CLB 4	CLB 5	CLB 6	CLB 7	CLB 8	CLB 9	CLB 10	CLB 11	CLB 12
1	Mean	13.2	—	26.7	—	39.1	—	47.1	—	60.9	66.8	—	74.2
	Median	12	—	26	—	40	—	47	—	60	66	—	75
	Mode	10	—	24	—	40	—	48	—	60	68	—	75
	Minimum	10	—	22	—	36	—	44	—	58	62	—	72
	Maximum	22	—	30	—	41	—	52	—	64	70	—	76
	SD	4.4	—	2.7	—	1.5	—	2.1	—	2.4	2.0	—	1.4
	n	13	—	17	—	17	—	17	—	17	17	—	15
2	Mean	12.2	16.9	25.6	32.1	39.6	43.6	48.0	53.8	60.8	66.6	70.4	73.8
	Median	10	14	24	32	40	44	48	54	60	66	70	74
	Mode	10	12	24	32	40	44	48	54	60	66	70	75
	Minimum	8	12	22	28	36	40	44	52	58	64	68	72
	Maximum	20	25	30	36	42	46	50	56	64	68	72	75
	SD	3.8	4.6	2.7	2.1	1.5	1.4	1.7	1.4	1.6	1.3	1.5	1.4
	n	14	17	17	17	17	17	17	17	17	17	17	17
3	Mean	11.6	16.0	25.2	31.8	39.5	43.8	48.2	53.9	60.8	66.3	70.1	73.6
	Median	10	14	24	32	40	44	48	54	60	66	70	74
	Mode	10	12	24	32	40	44	48	54	60	66	70	75
	Minimum	8	12	24	30	36	42	46	52	58	64	68	72
	Maximum	16	25	30	35	42	46	50	56	62	68	72	75
	SD	2.6	4.1	2.1	1.5	1.3	1.0	1.2	1.3	1.2	1.1	1.1	1.3
	n	11	17	17	17	17	17	17	17	17	17	17	17

Note. *n* = number of panelists providing a cut score. CLB = Canadian Language Benchmarks.

Table 11 Standard Setting Results for the Writing Section of the TOEFL Essentials Test

Round	Statistic	CLB 1	CLB 2	CLB 3	CLB 4	CLB 5	CLB 6	CLB 7	CLB 8	CLB 9	CLB 10	CLB 11	CLB 12
1	Mean	1.4	—	3.8	—	7.2	—	11.6	—	15.1	17.0	—	19.7
	Median	1	—	4	—	7	—	11	—	15	17	—	20
	Mode	1	—	4	—	7	—	11	—	17	17	—	20
	Minimum	1	—	2	—	5	—	9	—	13	15	—	18
	Maximum	2	—	5	—	10	—	15	—	17	20	—	20
	SD	0.5	—	0.7	—	1.2	—	1.7	—	1.5	1.4	—	0.6
	n	8	—	17	—	17	—	17	—	17	16	—	11
2	Mean	1.2	2.0	3.9	5.4	7.2	9.3	11.4	13.5	15.3	17.2	18.5	19.8
	Median	1	2	4	5	7	9	11	13	15	17	18	20
	Mode	1	1	4	5	7	9	11	13	15	17	18	20
	Minimum	1	1	3	4	6	8	10	11	13	16	17	19
	Maximum	2	3	5	6	8	13	14	16	17	19	20	20
	SD	0.4	0.9	0.4	0.6	0.5	1.1	0.9	1.4	1.2	0.9	0.8	0.4
	n	6	17	17	17	17	17	17	17	17	17	17	12
3	Mean	1.0	1.9	3.9	5.4	7.2	9.2	11.4	13.6	15.4	17.3	18.5	19.8
	Median	1	2	4	5	7	9	11	13	15	17	18	20
	Mode	1	1	4	5	7	9	11	13	15	17	18	20
	Minimum	1	1	3	4	6	8	10	12	13	16	17	19
	Maximum	1	3	5	6	8	11	13	16	17	19	20	20
	SD	0.0	0.9	0.5	0.6	0.5	0.7	0.7	1.2	1.1	0.8	0.8	0.4
	n	3	17	17	17	17	17	17	17	17	17	17	12

Note. *n* = number of panelists providing a cut score. CLB = Canadian Language Benchmarks.

Table 12 Standard Setting Results for the Listening Section of the TOEFL Essentials Test

Round	Statistic	CLB 1	CLB 2	CLB 3	CLB 4	CLB 5	CLB 6	CLB 7	CLB 8	CLB 9	CLB 10	CLB 11	CLB 12
1	Mean	—	1.9	5.8	12.1	20.8	28.6	34.8	40.5	45.4	50.7	52.7	54.0
	Median	—	2	4.5	9	20	30	35	39	45	51	53	54
	Mode	—	2	8	7	8	36	29	50	52	54	54	54
	Minimum	—	1	1	2	8	11	17	24	34	44	49	54
	Maximum	—	3	17	32	42	44	47	50	53	54	54	54
	SD	—	0.6	4.5	8.1	8.6	8.5	7.9	7.1	5.4	3.2	1.6	0.0
	n	0	8	14	17	17	17	17	17	17	17	12	6
2	Mean	—	1.4	3.4	8.9	20.2	30.1	35.9	40.8	45.5	51.1	53.2	54.0
	Median	—	1	3	7	21	32	36	40	46	51	54	54
	Mode	—	1	1	7	21	32	33	40	47	50	54	54
	Minimum	—	1	1	2	9	21	30	33	39	49	52	54
	Maximum	—	2	8	15	29	36	43	50	52	54	54	54
	SD	—	0.5	2.4	4.1	5.2	4.5	3.6	4.5	3.4	1.8	1.0	0.0
	n	0	7	16	17	17	17	17	17	17	17	13	6
3	Mean	—	1.4	4.1	9.9	20.8	30.3	35.5	40.9	46.3	50.9	53.2	54.0
	Median	—	1	4	9	21	31	36	40	46	50	54	54
	Mode	—	1	4	7	20	24	35	40	45	50	54	54
	Minimum	—	1	1	3	10	24	30	36	44	49	52	54
	Maximum	—	3	10	20	30	36	40	46	52	54	54	54
	SD	—	0.7	2.5	4.3	4.3	3.7	2.5	3.0	2.2	1.7	1.0	0.0
	n	0	10	16	17	17	17	17	17	17	17	14	6

Note. *n* = number of panelists providing a cut score. CLB = Canadian Language Benchmarks.

The results for the listening section are presented in Table 12. The mean cut scores across the three rounds were very similar. On the raw score scale of 0–54, the largest changes were seen for the cut scores of CLB 3 and CLB 4, where cut scores decreased by 2.4 points and 3.2 points, respectively, from Round 1 to Round 2. The standard deviation tended to decrease across rounds, although in the case of the cut scores for CLB 3 and CLB 4, it went up by 0.1 and 0.2 points, respectively, between Round 2 and Round 3.

The results for the cut scores of the reading section are presented in Table 13. The mean cut scores across the three rounds were very similar. On the raw score scale of 0–53, the largest change was in the boundary between Stage I and

Table 13 Standard Setting Results for the Reading Section of the TOEFL Essentials Test

Round	Statistic	CLB 1	CLB 2	CLB 3	CLB 4	CLB 5	CLB 6	CLB 7	CLB 8	CLB 9	CLB 10	CLB 11	CLB 12
1	Mean	—	1.3	4.7	9.8	18.0	25.8	33.4	39.5	45.7	51.0	52.8	53.0
	Median	—	1	4	9	17	28	32	40	45	52	53	53
	Mode	—	1	4	7	22	20	31	41	45	53	53	53
	Minimum	—	1	1	4	9	18	23	31	40	45	51	53
	Maximum	—	3	9	18	32	39	43	52	53	53	53	53
	SD	—	0.8	2.4	4.1	5.5	5.9	5.5	5.3	3.9	2.2	0.6	0.0
	n	0	6	15	16	17	17	17	17	17	17	12	2
2	Mean	—	1.0	4.5	8.6	16.9	24.9	32.8	38.8	46.1	51.6	52.9	53.0
	Median	—	1	4	9	17	26	34	38	46	52	53	53
	Mode	—	1	3	7	17	21	37	36	46	53	53	N/A
	Minimum	—	1	2	4	12	20	24	33	40	48	52	53
	Maximum	—	1	7	12	25	31	37	46	53	53	53	53
	SD	—	0.0	1.6	2.4	3.4	3.8	3.7	3.9	3.4	1.7	0.3	N/A
	n	0	2	16	17	17	17	17	17	17	17	9	1
3	Mean	—	1.0	4.6	8.5	16.5	24.5	32.8	38.8	46.0	51.4	52.8	53.0
	Median	—	1	4.5	9	16	25	33	39	46	51	53	53
	Mode	—	1	4	10	15	25	30	38	46	53	53	53
	Minimum	—	1	3	6	12	21	30	35	40	48	52	53
	Maximum	—	1	6	10	21	28	37	43	51	53	53	53
	SD	—	0.0	0.8	1.5	2.4	2.1	2.4	2.5	2.4	1.7	0.4	0.0
	n	0	4	16	17	17	17	17	17	17	17	10	2

Note. n = number of panelists providing a cut score. CLB = Canadian Language Benchmarks.

Stage II, Levels CLB 4–CLB 6, where cut scores varied from 1.3 to 1.5 points across the three rounds. The variability in panelists' judgments decreased in general across rounds, as can be seen by reductions in standard deviation, suggesting convergence in the final round of judgments. (One exception was CLB 11, for which there was a minor increase in standard deviation from 0.3 in Round 2 to 0.4 in Round 3).

Results of the Meeting Evaluation Survey

Panelists completed the end-of-meeting evaluation survey twice during the standard setting study: once at the end of the 2-day meeting for speaking and writing and a second time at the end of the 2-day meeting for listening and reading. The survey was conducted (a) to understand the panelists' perceived overall satisfaction with the meeting process, (b) to determine the factors that influenced their standard setting judgments, and (c) to assess their level of confidence in the recommended cut scores. The two surveys were identical except for two questions regarding the helpfulness of familiarization activities completed prior to the standard setting meeting, which were included in the second survey only. The information collected from the survey provides critical evidence for procedural validity, providing the panelists' point of view regarding whether the standard setting procedures were clear, the degree to which different types of information influenced their judgments, and their confidence in the resulting cut scores (Hambleton et al., 2012; Papageorgiou & Tannenbaum, 2016; Tannenbaum & Cho, 2014).

Table 14 summarizes the panel's feedback on the standard setting process for each of the two meetings. Most of the panelists strongly agreed or agreed that they understood the distinguishing features of the 12 CLB levels, although more panelists strongly agreed for listening and reading than for speaking and writing. This difference could be partly explained by the fact that listening and reading were explored in the second 2-day meeting, allowing panelists to become more familiar with the CLB. The majority of the panelists also strongly agreed or agreed that the types of instructions and feedback that they received were clear and helpful (Table 14).

Table 15 shows that most panelists found the completion of the familiarization activities helpful, including both the activity asking them to sort the CLB descriptors into their appropriate levels and the activity asking them to summarize the distinguishing features of CLB levels in their own words (see Appendix C).

Table 16 summarizes the panelists' opinions about the extent to which particular factors influenced their standard setting judgments. The majority of panelists indicated that the definitions of JQCs produced by the group were very influential in guiding their decisions. The discussion of the descriptors to distinguish the 12 CLB levels and the between-round

Table 14 Panelists' Perceptions of the Clarity and Helpfulness of Instructions and Feedback

Question	Meeting	Strongly disagree	Disagree	Agree	Strongly agree
I understood the distinguishing features of the 12 CLB levels.	L & R			4	13
	S & W			6	11
The instructions and explanations provided by the facilitators were clear.	L & R				17
	S & W			3	14
The explanation of the process for the judgment task helped me complete my assignment.	L & R				17
	S & W			3	14
The explanation of how the recommended cut scores are computed was clear.	L & R			4	13
	S & W			4	13
Feedback and discussion between judgment task rounds was helpful.	L & R				17
	S & W			4	13
The statistical information presented between rounds was helpful.	L & R			2	15
	S & W			1	16

Note. CLB = Canadian Language Benchmarks; L = listening; R = reading; S = speaking; W = writing.

Table 15 Panelists' Opinion About Helpfulness of the Familiarization Activities (Listening and Reading Meeting Only)

	Strongly disagree	Disagree	Agree	Strongly agree
Part 1: Sorting CLB descriptors			4	13
Part 2: Summarizing the distinguishing features of CLB levels			4	13

Note. CLB = Canadian Language Benchmarks.

Table 16 Panelists' Opinion About the Influence of Study Materials in Making Standard Setting Judgments

Question	Meeting	Not influential	Influential	Very influential
The definition of the just qualified candidate (JQC)	L & R		2	15
	S & W		2	15
The discussion of the descriptors to distinguish the 12 CLB levels	L & R	1	7	9
	S & W		7	10
The between-round discussion	L & R		6	11
	S & W	1	8	8
The summary statistics presented after each round of judgments	L & R		4	13
	S & W	2	4	11
My own professional experience	L & R		2	15
	S & W		3	14

Note. L = listening; R = reading; S = speaking; W = writing; CLB = Canadian Language Benchmarks.

discussion were both thought to be influential or very influential. Most panelists found the summary statistics presented after each round of judgments and their own professional experience to be influential or very influential, except for two panelists claiming no influence of the summary statistics for speaking and writing.

The survey also asked panelists to rate the degree to which the meeting process was (a) efficient, (b) coordinated, (c) understandable, and (d) satisfying (Table 17). Panelists gave overwhelmingly positive ratings for the meeting process for listening and reading, although the speaking and writing received generally high ratings for all aspects of the meeting process.

Panelists were also asked to indicate their level of confidence with the standard setting results (Table 18). Almost all panelists were either very comfortable or comfortable with the cut scores for all sections, but more panelists showed a higher level of confidence for reading and speaking compared to listening and writing. Yet, one panelist indicated a lower level of confidence in the speaking section cut score. Although the panelists' level of confidence in the cut scores is satisfactory across all sections, we note that the higher level of confidence in the cut scores of the reading section is consistent with the observation that the dispersion of judgments for reading Round 1 cut score was relatively narrow compared to Round 1 judgments for other skills.

A final, optional, question in the evaluation survey asked panelists to provide comments on the standard setting process. Thematic analysis of these comments revealed three main topics: (a) the meeting process and procedure, (b) personal

Table 17 Panelists' Evaluation of the Meeting Process

Question	Meeting	1	2	3	4	5
Inefficient (1)—Efficient (5)	L & R				2	15
	S & W			1	4	12
Uncoordinated (1)—Coordinated (5)	L & R					17
	S & W			1	1	15
Confusing (1)—Understandable (5)	L & R				1	16
	S & W				4	13
Dissatisfying (1)—Satisfying (5)	L & R				1	16
	S & W				6	11

Note. L = listening; R = reading; S = speaking; W = writing.

Table 18 Panelists' Confidence in the Standard Setting Results

Test section	Very uncomfortable	Somewhat uncomfortable	Somewhat comfortable	Very comfortable
Listening			6	11
Reading			2	15
Speaking		1	3	13
Writing			8	9

reflections about the experience, and (c) challenges and suggestions. (See Appendix G for subthemes and example quotes for each theme.) In terms of the meeting process and procedure, the great majority of panelists' comments were positive and related to the guidance given by the project team, the organization and efficiency of the meeting, and the group discussions. Many panelists mentioned that the meetings were well organized and efficient and that they did not experience any interruption or delay. The panelists also found the group discussion to be a helpful activity to make cut score judgments, for the most part; however, a few of them mentioned that the discussion deviated from the topic at hand from time to time.

Almost all panelists expressed overwhelmingly positive feelings about their experience in the standard setting meetings. Their comments often included positive key words and phrases such as "learned a lot," "enjoyed," "interesting," "rewarding," and "great experience." However, some panelists also mentioned challenges they encountered during the standard setting meetings and made a few suggestions to improve the experience. Challenges mentioned were mostly related to the development of the JQCs and judgment tasks. For example, some panelists commented that they felt the JQC descriptors developed as a group were sometimes not directly applicable to the test tasks in the TOEFL Essentials test and that some scaffolding might have helped them in making independent cut score judgments. A few logistical suggestions were also made by panelists such as extending the length of breaks and using a different online conference platform.

Final Score Mapping

Performance on each of the four sections of the TOEFL Essentials test is reported in the form of band scores from 1 to 12. The overall band score is computed as the average of the four section band scores and reported in half-band increments, with section scores reported in full-band increments (Papageorgiou et al., 2021). To facilitate the standard setting judgment task, the panelists made recommendations based on raw scores only, without consideration of the conversion of raw scores to the reported band scores. This approach meant that they could easily connect their judgments to the number of correct items needed to be classified at a specific CLB level for the selected-response sections and the descriptors listed in the scoring rubrics for the constructed-response sections. Therefore, it was necessary for the project team to render the panel's recommended raw cut scores into reported band scores to arrive at the final score mapping. In setting the final cut scores, the project team also considered other sources of information relevant to alignment of TOEFL Essentials results with CLB levels. Such information included the MST design used for the listening, reading, and writing sections; the results of the construct congruence study; the mapping of the TOEFL Essentials band scores to the CEFR levels (ETS, 2021, Table 19), and the available information regarding the proposed correspondence between the CEFR levels and the CLB (North & Piccardo, 2019, Table 20). In finalizing the score mapping, the project team also considered implications for false positive and false negative classifications (ETS, 2020), with priority given to minimizing false positive classifications. That is, it was felt important to avoid classifying test takers into a CLB level higher than appropriate,

Table 19 Mapping of TOEFL Essentials Section and Overall Band Scores to the Common European Framework of Reference (CEFR) Levels (ETS, 2021)

CEFR level	Section band score (1–12)	Overall band score (1–12)
C2	12	12
C1	10–11	10–11.5
B2	8–9	8–9.5
B1	5–7	5–7.5
A2	3–4	3–4.5
A1	2	2–2.5
Below A1	1	1–1.5

Table 20 Proposed Correspondence Between Common European Framework of Reference (CEFR) Levels and Canadian Language Benchmarks (CLB; North & Piccardo, 2019)

CEFR	CLB
C2	12
C1	11
	10
B2	9
	8
B1	7
	6
	5
A2	4
	3
A1	2
Below A1	1

given the likely use of test scores for gatekeeping purposes such as admission to academic programs. Accordingly, in cases of ambiguity regarding the appropriate cut score, a relatively higher value was chosen.

The final score mapping of TOEFL Essentials test scores to the CLB levels is presented in Table 21. Based on the construct congruence study results and also the relatively small number of panelists recommending cut scores for the top and bottom levels (see Results of the Standard Setting Judgments Section), no cut scores are recommended for CLB 1 and CLB 12 for any test section. No reading or listening cut scores are recommended for CLB 2 for the reading and listening sections for the same reason. In addition, the recommended raw cut scores for CLB 2 were 1.4 and 1.0 for listening and reading, respectively, which could have been achieved by chance. The cut scores for the overall band score were produced by taking the average of the cut scores for each of the four sections, rounded up to the nearest whole or half band, in accordance with the priority given to minimizing false positive classifications.

Cut Score Validation

As mentioned earlier in this report, three primary sources of validity evidence are relevant to standard setting: procedural, internal, and external evidence (Council of Europe, 2009; Hambleton et al., 2012; Tannenbaum & Cho, 2014). In this section, we discussed these three aspects in relation to our study.

Table 21 Mapping of TOEFL Essentials Test Scores to the Canadian Language Benchmarks (CLB) Levels

CLB	Speaking band score	Writing band score	Listening band score	Reading band score	Overall band score
11	12	12	12	12	12
10	11	11	11	11	11–11.5
9	10	9–10	10	10	10–10.5
8	7–9	7–8	9	8–9	8–9.5
7	6	6	7–8	7	6.5–7.5
6	5	5	6	6	5.5–6.0
5	4	4	5	4–5	4.5–5
4	3	3	3–4	3	3–4
3	2	2	2	2	2–2.5
2	1	1	N/A	N/A	N/A

Table 22 Correlations Between Average Canadian Language Benchmarks (CLB) Level Judgments and Item Difficulty

Test section	Round	Correlation
Listening	1	−.70
	2	−.67
Reading	1	−.85
	2	−.85

Procedural Validity Evidence

Procedural evidence supports the recommended cut scores by establishing that the panel was appropriately selected and qualified, that training procedures were effective, and that the judgment process was conducted appropriately. In the current study, methods were based on best practices established in the standard setting literature, and information regarding the study procedures is given in prior sections of this report. We also note that the selection of the panelists was carefully coordinated by the project team and ETS staff in Canada, paying attention to regional representation and variety of educational institutions (Appendix A). This effort was made to help ensure that the panel's recommendations would incorporate a diversity of views, experiences, and contexts. A final piece of procedural evidence to support claims regarding the validity of the mapping results is the feedback provided by the panelists. As discussed in the Results of the Meeting Evaluation Survey Section, the panelists rated the various procedural aspects of the standard setting study positively, and most of them expressed high levels of confidence in the recommended cut scores.

Internal Validity Evidence

Internal validity evidence addresses issues of accuracy and consistency of the standard setting results. As discussed in the Results of the Standard Setting Judgments Section, variability of judgments generally decreased across judgment rounds for all four test sections. In addition, there was a clear relationship between the panelists' judgments of the proficiency level needed to answer each reading and listening item and the empirical difficulty for these items. In a separate analysis, we estimated the correlation between the mean CLB level assigned to each item by the panel and the empirical item difficulty value (percent of test takers answering the item correctly). This analysis was conducted for the first and second round for listening and reading, because the panelists provided judgments for each item, as opposed to the holistic judgment provided in Round 3. High correlations were observed in all cases (see Table 22), suggesting that, in general, more difficult items were also estimated by the panelists to require higher levels of language proficiency (hence correlations were negative). It should also be pointed out that high correlations were observed even in Round 1, during which panelists completed their judgments tasks individually without any information about item difficulty.

External Validity Evidence

External validity evidence is composed of independent sources of information that support claims regarding the appropriateness of the results produced in a standard setting study. For the current effort, an initial piece of external validity

Table 23 Correspondence Between Canadian Language Benchmarks (CLB) and Common European Framework of Reference (CEFR) Levels

	Equivalent CEFR level	
CLB level	North and Piccardo (2019)	TOEFL Essentials (ETS, 2021; current study)
12	C2	N/A
11	C1	C2
10	C1	C1
9	B2	C1
8	B2	B2
7	B1	B1/B2
6	B1	B1
5	B1	A2/B1
4	A2	A2
3	A1/A2	A1
2	A1	N/A
1	Below A1	N/A

evidence is the comparison of the mapping between CLB and CEFR levels as reported in North and Piccardo's study (2019) versus the mapping between the two frameworks implied by TOEFL Essentials band scores. Comparing the levels of different language frameworks based on separate score mapping studies can be challenging (see research in the volume edited by Tschirner, 2012), but nevertheless offers an additional perspective regarding the reasonableness of the score mapping presented in this report. This triangulation was produced by combining the data previously reported in Tables 19 and 20 and the overall TOEFL Essentials band scores reported in Table 21; the triangulation results are shown in Table 23.

Table 23 indicates that the two sources of information produce similar but not identical CLB–CEFR alignments. This result is to be expected given that the CLB and the CEFR are separate frameworks with overlapping but different perspectives, contexts of use, and development procedures. Accordingly, considerable interpretation is required in the alignment of levels. Also, some difference in the interpretation of the CLB levels is not surprising due to the lack of TOEFL Essentials total band scores for the bottom and top CLB levels.

Limitations

In this report, we provided a detailed rationale behind the mapping of TOEFL Essentials test scores onto the levels of the CLB, building on several sources of data, which included a construct congruence study and a standard setting study with expert panelists. We also presented evidence of procedural, internal, and external validity of the recommended cut scores.

The different sources of data provide support for the recommended score mapping. However, policymakers in the educational context where the CLB levels are used might want to further investigate the relationship between TOEFL Essentials test scores and the CLB to facilitate decision-making based on the test scores.

Based on our experience from this research project and the literature on mapping test scores to the proficiency levels of different language frameworks, we would also advise caution regarding the following potential issues with score interpretation.

- Mapping test scores to proficiency levels can be a useful tool for guiding score interpretations. However, the content of the test and how this content relates to the language abilities of interest is a critical issue to consider when deciding whether a given test result provides useful evidence for decision-making based on a set of proficiency levels such as the CLB.
- Similarly, two language tests mapped to the same proficiency levels are not necessarily equivalent in terms of content, and so their scores should not be considered interchangeable based solely on score mapping. Again, comparability of test results should be based on consideration of the abilities measured in each test, along with elements of technical quality such as the reliability of scores.
- Mapping test scores to the levels of a language proficiency framework is not necessarily simple, direct, or established as a one-time event. In the case of the TOEFL Essentials test, the appropriateness of score mapping will be evaluated on an ongoing basis as new information becomes available from contexts where the CLB are used.

Additionally, we found that it was possible to conduct standard setting activities in an entirely online format, although the potential for this arrangement to influence the deliberations of the panel remains an open question. The online format was dictated by limitations on travel associated with the COVID-19 pandemic, but this approach also had the advantages of convenience, reduced cost, and greater access for participants who might not have been able to travel. We also found that technical problems were few during group meetings and that panelists actively contributed to the discussions through both video and text chat. Nonetheless, it seems possible that the attention and participation of panelists may have been influenced in unknown ways by the very different circumstances associated with participating from home compared to a traditional face-to-face meeting. We saw no obvious indication that the online format impacted panelists' discussions or their judgments, but the interaction between standard setting processes and the method of work (i.e., online vs. face-to-face) is an intriguing question to be explored in future research.

Conclusion

In the current study, we employed information from a variety of sources to map scores from the TOEFL Essentials test to the CLB. An initial source of information for this mapping was a detailed analysis of test content, which established alignment between the language constructs measured by the TOEFL Essentials test and the language abilities specified at each level of the CLB. A second key source of information was a standard setting study, where specific cut scores for individual CLB levels were determined by a panel of Canadian experts from diverse educational contexts. These experts reviewed test content and test-taker responses in light of the descriptions provided in the CLB, as well as their own professional experiences, and provided recommended cut scores for each test section. The panel's recommendations were finally compared to other relevant information by the project team, who set the final cut scores.

Mapping TOEFL Essentials test scores to the CLB represented a challenge in terms of both the range of proficiency covered as well as the number levels for which cut scores needed to be identified. Nonetheless, using multiple lines of evidence, it was possible to establish what we believe is a reasonable and useful score mapping. The appropriateness of this score mapping will also continue to be evaluated as more information becomes available, to help ensure that the best possible data underlie score interpretations and the decisions based on them.

References

Centre for Canadian Language Benchmarks. (2012). *Canadian Language Benchmarks: English as a second language for adults*. www.canada.ca/content/dam/ircc/migration/ircc/english/pdf/pub/language-benchmarks.pdf

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

Council of Europe. (2001). *The common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

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

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment. Companion volume with new descriptors*. https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989

Davis, L., & Norris, J. (2021). Developing an innovative elicited imitation task for efficient English proficiency assessment. *ETS* https://doi.org/10.1002/ets2.12338, *2021*, 1, 30

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

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

ETS. (2020). *Guidelines for setting useful score requirements for the TOEFL iBT test* (TOEFL Research Insight Series Vol. 9). https://www.ets.org/s/toefl/pdf/toefl_ibt_insight_s1v9.pdf

ETS. (2021). *Setting score requirements*. https://www.ets.org/s/toefl-essentials/score-users/scores-admissions/set/

Ferrara, S., Perie, M., & Johnson, E. (2008). Matching the judgmental task with standard setting panelist expertise: The Item-Descriptor (ID) Matching method. *Journal of Applied Testing Technology*, *9*(1), 1–22.

Fleckenstein, J., Keller, S., Kruger, M., Tannenbaum, R. J., & Koller, O. (2020). Linking *TOEFL iBT®* writing rubrics to CEFR levels: Cut scores and validity evidence from a standard setting study. *Assessing Writing*, *43*, 1–15. https://doi.org/10.1016/j.asw.2019.100420

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

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.

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

Kostromitina, M., & Plonsky, L. (2022). Elicited imitation tasks as a measure of L2 proficiency: A meta-analysis. *Studies in Second Language Acquisition*, *44*(3), 886–911. https://doi.org/10.1017/S0272263121000395

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. MIT Press.

North, B., & Piccardo, E. (2019). *Aligning the Canadian Language Benchmarks (CLB) to the Common European Framework of Reference (CEFR)*. Centre for Canadian Language Benchmarks. https://www.language.ca/aligning-clb-and-cefr/

Papageorgiou, S., Davis, L., Norris, J. M., Garcia Gomez, P., Manna, V. F., & Monfils, L. (2021). *Design framework for the* TOEFL® Essentials™ *test 2021* (Research Memorandum No. RM-21-03). ETS.

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

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

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

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

Appendix A

Schedule for the Standard Setting Workshop

Friday, January 14, 2022, 11:00 AM–7:00 PM (EST)

- Welcome
- Developing just qualified definitions for speaking
- Training on standard setting method for productive skills and practice for the judgment task
- Round 1 judgments for speaking
- Round 1 discussion and Round 2 judgments for speaking
- Round 2 discussion, Round 3 judgments, and finalization of cut scores for speaking

Saturday, January 15, 2022, 11:00 AM–7:00 PM (EST)

- Developing just qualified definitions for writing
- Practice for the judgment task
- Round 1 judgments for writing
- Round 1 discussion and Round 2 judgments for writing
- Round 2 discussion, Round 3 judgments, and finalization of cut scores for writing

Friday, January 28, 2022, 11:00 AM–7:00 PM (EST)

- Welcome
- Developing just qualified definitions for listening

- Training on standard setting method for receptive skills and practice for the judgment task
- Round 1 judgments for listening
- Round 1 discussion and Round 2 judgments for listening
- Round 2 discussion, Round 3 judgments, and finalization of cut scores for listening

Saturday, January 29, 2022, 11:00 AM–7:00 PM (EST)

- Developing just qualified definitions for reading
- Practice for the judgment task
- Round 1 judgments for reading
- Round 1 discussion and Round 2 judgments for reading
- Round 2 discussion, Round 3 judgments, and finalization of cut scores for reading

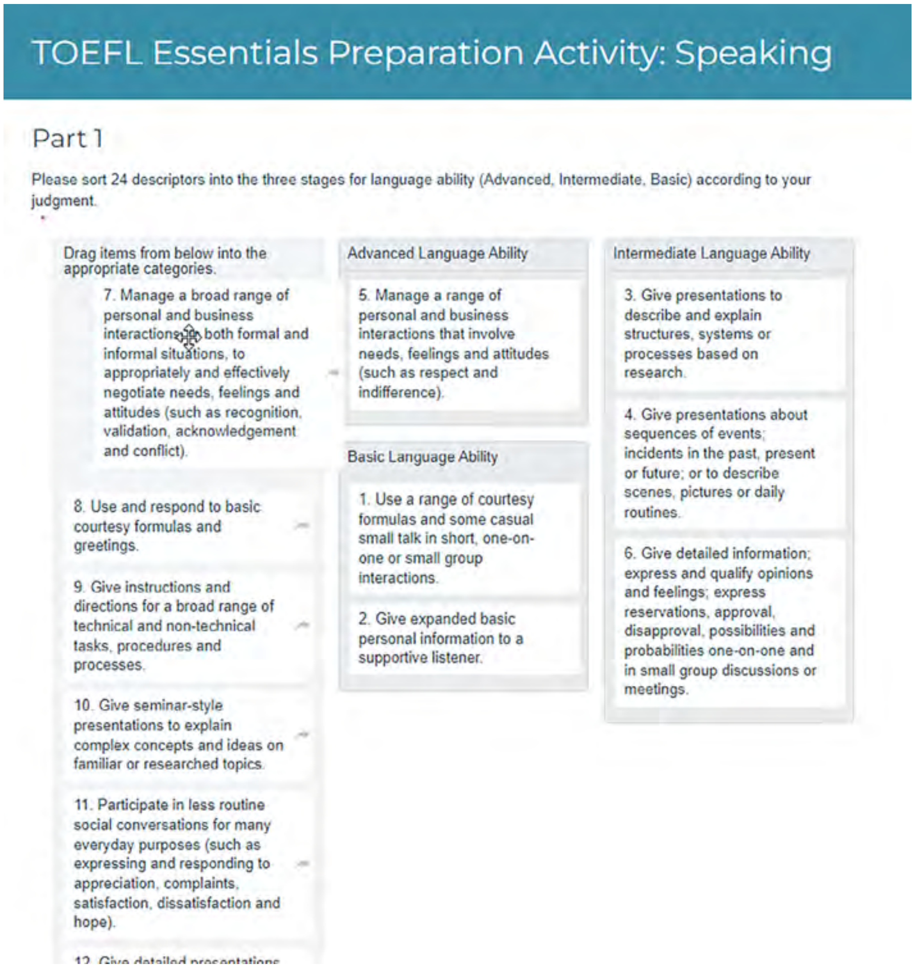
Appendix B

The Standard Setting Panelists, Affiliation, and Province

Panelist name	Institution	Province
Adam Saleh	York University	Ontario
Amy Yani	Sheridan College	Ontario
Andrea Szilagyi	University of British Columbia, Okanagan	British Columbia
Angel Arias	Carleton University	Ontario
Dani Stoyanova Zheleva	City University of Seattle, Vancouver campus	British Columbia
Emel Ortac	Saskatchewan Polytechnic	Saskatchewan
Janice GT Penner	Douglas College	British Columbia
Kevin Matthew Sison	Toronto District School Board	Ontario
Majid Nikouee	University of Alberta	Alberta
Mariya Petkova	Sheridan College	Ontario
Michael William Lynn	York University	Ontario
Nicola Sattler	Atlantic Canada Language Academy	Nova Scotia
Ray Rahimi	Acsenda School of Management	British Columbia
Ronan John Scott	Okanagan College	British Columbia
Setsu Anne Crawford-Kawahara	YMCA of Greater Halifax and Dartmouth	Nova Scotia
Sunny Man Chu Lau	Bishop’s University	Quebec
Zhi Li	University of Saskatchewan	Saskatchewan

Appendix C

Screenshots of Sample Task for Panelist Familiarization With the CLB Levels



TOEFL Essentials Preparation Activity: Speaking

Part 1

Please sort 24 descriptors into the three stages for language ability (Advanced, Intermediate, Basic) according to your judgment.

Drag items from below into the appropriate categories.

7. Manage a broad range of personal and business interactions in both formal and informal situations, to appropriately and effectively negotiate needs, feelings and attitudes (such as recognition, validation, acknowledgement and conflict).

8. Use and respond to basic courtesy formulas and greetings.

9. Give instructions and directions for a broad range of technical and non-technical tasks, procedures and processes.

10. Give seminar-style presentations to explain complex concepts and ideas on familiar or researched topics.

11. Participate in less routine social conversations for many everyday purposes (such as expressing and responding to appreciation, complaints, satisfaction, dissatisfaction and hope).

12. Give detailed presentations

Advanced Language Ability

5. Manage a range of personal and business interactions that involve needs, feelings and attitudes (such as respect and indifference).

Basic Language Ability

1. Use a range of courtesy formulas and some casual small talk in short, one-on-one or small group interactions.

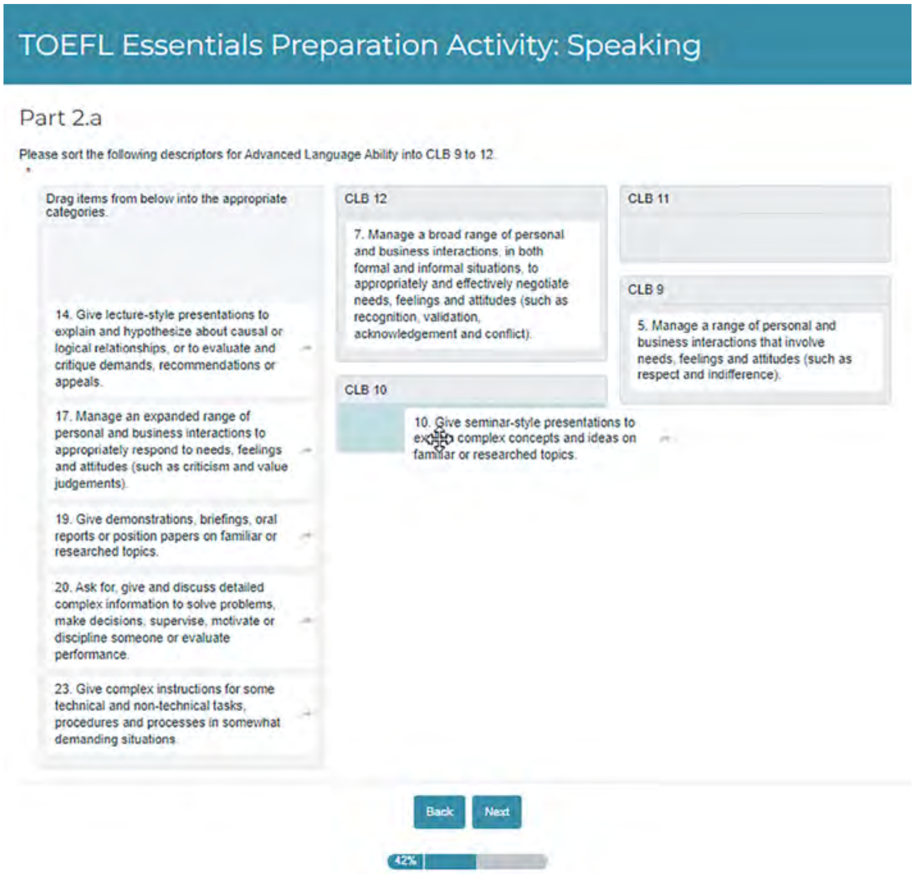
2. Give expanded basic personal information to a supportive listener.

Intermediate Language Ability

3. Give presentations to describe and explain structures, systems or processes based on research.

4. Give presentations about sequences of events; incidents in the past, present or future; or to describe scenes, pictures or daily routines.

6. Give detailed information; express and qualify opinions and feelings; express reservations, approval, disapproval, possibilities and probabilities one-on-one and in small group discussions or meetings.



TOEFL Essentials Preparation Activity: Speaking

Part 2.a

Please sort the following descriptors for Advanced Language Ability into CLB 9 to 12.

*

Drag items from below into the appropriate categories.

14. Give lecture-style presentations to explain and hypothesize about causal or logical relationships, or to evaluate and critique demands, recommendations or appeals.

17. Manage an expanded range of personal and business interactions to appropriately respond to needs, feelings and attitudes (such as criticism and value judgements).

19. Give demonstrations, briefings, oral reports or position papers on familiar or researched topics.

20. Ask for, give and discuss detailed complex information to solve problems, make decisions, supervise, motivate or discipline someone or evaluate performance.

23. Give complex instructions for some technical and non-technical tasks, procedures and processes in somewhat demanding situations.

CLB 12

7. Manage a broad range of personal and business interactions, in both formal and informal situations, to appropriately and effectively negotiate needs, feelings and attitudes (such as recognition, validation, acknowledgement and conflict).

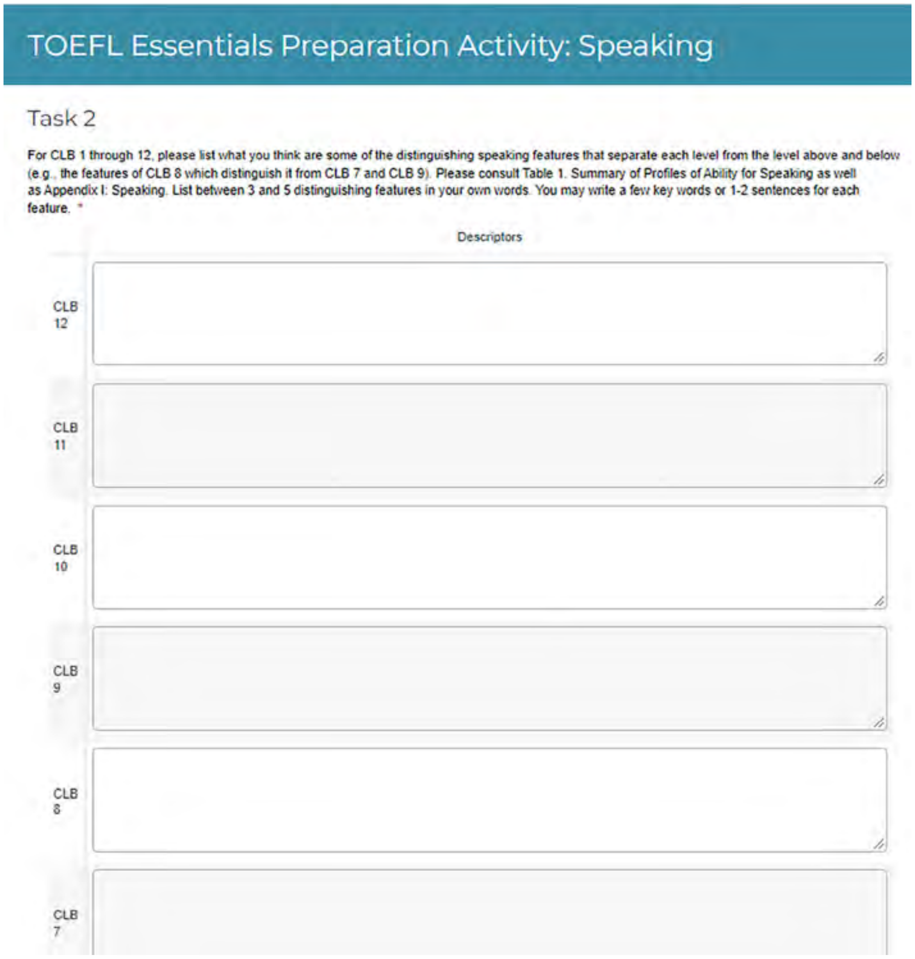
CLB 10

10. Give seminar-style presentations to explain complex concepts and ideas on familiar or researched topics.

CLB 11

CLB 9

5. Manage a range of personal and business interactions that involve needs, feelings and attitudes (such as respect and indifference).



TOEFL Essentials Preparation Activity: Speaking

Task 2

For CLB 1 through 12, 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 CLB 8 which distinguish it from CLB 7 and CLB 9). Please consult Table 1. Summary of Profiles of Ability for Speaking as well as Appendix I: Speaking. 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. *

	Descriptors
CLB 12	
CLB 11	
CLB 10	
CLB 9	
CLB 8	
CLB 7	

Appendix D

Borderline Student Definitions by the Panelists

Table D1 Speaking

CLB 1	CLB 2	CLB 3	CLB 4
● Can respond to most basic input ("hello") ● Relies heavily on gestures, reacts to gestures ● Rarely uses English words ● Relies on L1 ● Single words, basic vocabulary ● Listener must be highly supportive	● Give *very* basic info about self (+basic vocabulary for this purpose) ● Some phrases about self, very simple sentences ● One-word answers ● Rote/formulaic phrases ● Highly predictable situations ● Single familiar interlocutor ● Simple, nongrammatical questions, or basic formulaic questions ● Combination of gestures and words ● Requires high degree of support ● Name concrete and familiar objects, here and now, limited use of adjectives ● Give basic info about family ● Errors cause confusion	● Describe personal needs and experiences ● Can produce some full sentences ● Extensive errors ● Give basic info about self and family in a coherent, intelligible way ● Basic/formulaic structures may be produced accurately ● Errors frequently impede communication ● Speech rate slow and with hesitations ● Emerging basic structures ● Emerging ability to verbally indicate communication and comprehension problems ● Guided and encouraged by a supportive listener	● Somewhat limited vocabulary, somewhat basic control over structures. ● Errors may impede communication ● Start showing ability to produce connected discourse ● Can communicate about daily routines, personal information ● Starting to self-correct ● Start attempting to link ideas/sentences, greater willingness to attempt more complex structures ● Use of simple questions ● Comfort with basic structures ● Use/awareness of basic tense (simple present, present continuous), but with errors ● Basic social language (introduce self, etc.) basic turn taking, start/ending conversation ● Emerging comfort with communicating in small groups ● Emerging exchanges on digital media; ability to communicate in phone conversation ● Can verbally indicate communication and comprehension problems ● Moderate supportive listener

Table D1 (Continued)

CLB 5	CLB 6	CLB 7	CLB 8
• Express opinions and feelings • Social circumstances as context • Produce cohesive texts • Not requiring sympathetic listeners • Manage short routine conversations • Give a basic presentation about concrete topics (up to 5 min) • Use of past and future tenses • Describe pitches and scenes • Formality is present • Good control of simple structure • Basic interpersonal communication • Show awareness of nonverbal cues (gesture, body language) and respond accordingly • Stay on familiar topics • Extended conversations (phone) • Some varieties in sentence structure • Limited vocabulary (some idiomatic expressions) • Some connected speech and start talking about nonfamiliar topics • More natural speech, less hesitation, some fluency • Interact in small groups and respond to various interlocutors	• Less hesitation, more fluent delivery of speech • Use of comparative language on familiar topics • Express apologies, excuses, opinions, make suggestions or arrangements • Discuss topics that are relevant or interesting to the speaker • Talk about familiar topics • Attempt to use some complex structure • Communicate with strangers (e.g., phone calls) with challenges on familiar topics (different accents and contexts) • Give presentations with visuals about familiar topics for 5 – 6 min • Initial ability to use formulaic idiomatic language, adapt speech to certain degrees of formality • Willing to take risks • Communicate facts and ideas in some detail • Use reporting language • Provide simple narration • Attempt to use some idiomatic expressions on the fly • Cultural references • Some vocabulary to cover unfamiliar topics • Make indirect request appropriately • Ask for confirmation and clarification	• Emerging/initial ability to express abstract concepts • Start to paraphrase • Give presentation for more than 7 min • Beginning ability to keep the conversation going and hold the floor • Expanding ability to talk about complex topics and about others with complexity (school, friends, teachers, etc.) • Expanding/frequent use of idiomatic expressions • Engage in less routine conversations related to professional or study-related topics • Errors sometimes interfere with communication • Pronunciation issues that do not hamper communication • May be able to handle unpredictable situations with new words or expressions • Use compare and contract language under complex and abstract scenarios/situations • Give instructions • Limited/initial use of technical and academic language in their area of expertise	• Emerging ability to present arguments • Express counterarguments with limited coherence • Respond to others' comments • Emerging use of technical language • Provide solutions to problems • Give academic presentations based on research for up to 20 min • Natural speech rate • Increasing confidence and fluency • Justify opinions • Show register awareness and style (intonation, etc.) • Show disagreements • Expanded ability to talk about range of familiar topics • Expanded use of idiomatic expressions • Talk within unfamiliar groups • Start to handle professional phone calls • Pronunciation errors are present but do not interfere with communication • Emerging ability to handle conflicts and debates • Use more body language cues

Table D1 (Continued)

CLB 9	CLB 10	CLB 11	CLB 12
● Effectively engage in conversations in demanding and moderately complex contexts with some challenges ● Initial use of figures of speech ● Start to exchange key information on academic topics ● Deal with complaints and reassure others ● Coordinate tasks with others ● Participate in a 30-min communication/interaction ● Give a presentation for 15 – 30 min ● Manage a range of personal and professional (business and academic) interactions ● Participate in work meetings and discuss professional topics ● Appropriately use a variety of conversational strategies (rephrasing, clarifying, and interactions) ● Co-facilitate interactions ● Aware of formal situations and boundaries of degree and distance ● Good control of complex grammatical structure ● Grammar, pronunciation, and lexical errors are present but rarely impede communication ● Communication is more challenging and non-routine ● Provide solutions based on research ● Developing confidence with unfamiliar audience (peers and authoritative figures) ● Attempt to persuade persons in low-stakes situations and potentially in high-stakes situations	● Begin to train others, delegate responsibilities, and make decisions in meetings and discussions ● Attempt to resolve conflicts with nonaggressive language ● Give seminar-style workshops ● Give presentations longer than 30 min on researched topics ● Increased ability to persuade persons in authority and high-stakes situations ● Syntactic, grammar, vocabulary, and pronunciation errors are nonsemantic and nonsystematic (and do not impede communication) ● Co-facilitate large formal discussions (go beyond small group discussions) ● Recognize the needs and expectation of audience or conference	● Start chairing or facilitating a large-scale discussion ● Attempt to negotiate solutions in demanding situations ● Manage perceived hostility, blame, and sarcasm ● Use different strategies to persuade persons ● Meet the needs and expectation of diverse audiences ● Ability to present both pros and cons of challenging argument ● Demonstrate pragmatic and strategic competence	● Present on specialized topics ● Manage and mediate conflicts ● Able to speak for any length of time as needed ● Respond to critique ● Use appropriate tone, intonation, and pitch ● Start recognizing confrontation

Note. CLB = Canadian Language Benchmarks.

Table D2 Writing

CLB 1	CLB 2	CLB 3	CLB 4
● Write letters, numbers, single familiar words (e.g., names, age)	● Short familiar phrases ● Very basic vocabulary ● Starting awareness of spelling, punctuation ● Can complete simple, short forms (basic info—names, dates, single words)—often memorized info ● Write answers to simple questions about immediate needs with assistance ● Requires significant support/scaffolding ● Irrelevant answers if not assisted ● Message is unclear	● Attempt simple short sentences on familiar topics ● Mostly uses everyday words and phrases ● Highly supportive and familiar reader (infer message meaning) ● Personal details in response to short questions ● Awareness of some spelling and punctuation conventions ● Produce a short message ● Can complete simple forms without help ● Write simple one-line descriptions of personal photographs ● Difficulty communicating simple facts and ideas ● Significant difficulty with word order and word forms, greatly impacts comprehensibility of message	● Developing control of spelling, punctuation, capitalization ● Varying control in producing simple correct sentences ● Write a few connected sentences ● Some difficulty with word order and word forms, impacts comprehensibility of message ● Ability to write a short description (familiar person, object, place, situation, or event) ● In some cases can create a simple clear message

Table D2 (Comtinued)

CLB 5	CLB 6	CLB 7	CLB 8
● Emerging use of connectives, cohesive devices ● Include main ideas ● Good control of simple structures, difficulty with complex structures ● Awkward use of words and phrases ● Attempt to write in more detail with challenges ● Follow standard writing conventions (punctuations, tenses) ● Errors impede communication ● Write about concrete/familiar topics and go beyond personal topics ● Specialized/routine topics ● Write concrete descriptions and write a story ● Problem with unity ● Attempt to write a basic, short paragraph with issues of coherence (up to one paragraph) ● Complete forms that require personal information (up to 20 pieces of info) ● Reduce information from familiar sources (recipe books), not paraphrase ● Some awareness of formality	● Use connectives accurately ● Use supporting information adequately ● Errors sometimes impede communication ● Attempt to use some complex structures ● Write to familiar or defined audience ● Use points from notes ● Start to have a clear main idea sentence within a paragraph ● Listen and write down as they heard ● Identify difference between informal and formal writing ● Read a poster and write notes ● Emerging ability to reduce factual and familiar information based on written texts to notes or messages ● Identify and summarize information that is personally related ● Attempt to connect paragraphs	● Write 2 or 3 paragraphs about a wide range of topics in different contexts (professional, academic) ● Topics are usually concrete but can be somewhat abstract ● Write to familiar and unfamiliar audience ● Able to connect paragraphs ● Errors rarely impede communication ● Use sufficient range of vocabulary to express precise meaning ● Include both unified main and supporting ideas (paragraph includes a topic sentence, concluding sentence) ● Cohesion is present ● Limited use of idiomatic language and cultural references ● Produce formal writing ● Reduce, jot down, and organize information to create an outline or a summary ● Start to write claims and support claims	● Write 3–4 paragraphs ● Expanding use of idiomatic language ● Write connected paragraphs with detailed description ● Write about more abstract topics within predictable contexts that are familiar ● Write a report ● Errors very rarely impede communication ● Good control of spelling, punctuation, and format ● Mostly appropriate level of formality and register

Table D2 (Comtinued)

CLB 9	CLB 10	CLB 11	CLB 12
● Communicate abstract information that may require research ● Good control of grammatical structures and vocabulary, but errors may still be present ● May exhibit errors in word order, collocation, or word choice ● Developing ability to synthesize texts of immediate relevance ● Produce well-developed paragraphs that are coherent (up to four paragraphs) ● Write a short four-paragraph essay (approximately 1,000 words) ● Express and analyze opinions (reflection) ● Start to be proofreading own texts but still rely on others to improve ● Complex and diverse structures ● Begin to use genre-specific expressions/vocabulary ● Limited flexibility with tone/style and register	● May use citations ● Write over 1,000 words (1,500–3,000 words) ● Write a five-paragraph essay ● Summarize 5–10 pages of texts ● Synthesize information from magazines and newspapers ● Proofread and revise own work with help from others ● Write a literature review ● A few errors in grammar and diction ● Exhibit accurate use of vocabulary and grammar	● Occasional errors in grammar, diction, word choice, collocations ● Write more than 3,000 words ● Write to a wide range of audience (defined and undefined) ● Attempt to include charts and graphs ● Write in unlimited contexts ● Proofread and revise own work as well as others' work ● Use a broad range of complex and diverse structures ● Write in appropriate tone/style/register ● Synthesize complex information from multiple sources in a logical way	● Effectively proofread and revise own work as well as others' work ● Write potentially publishable texts in specific disciplines (social sciences and humanities) ● Rare minor errors in sentence structure, vocabulary, collocations, and grammar ● Use a wide range of genre-specific expressions and jargons

Note. CLB = Canadian Language Benchmarks.

Table D3 Listening

CLB 1	CLB 2	CLB 3	CLB 4
Context ● Nondemanding ● Immediate personal needs ● Relies on strong visual support ● No communication over the phone Character ● Communication is extremely short ● Incredibly slow rate ● Needs extensive assistance … (repetition, modification, demonstration, translation) Understands … ● Very little understanding ● Can parrot simple input ● Numbers, letters ● Uses gestures/pointing to indicate understanding	Context ● Nondemanding ● Immediate personal needs ● Relies on strong visual support (gestures) ● No communication over the phone Character ● Communication is very short ● Slow rate ● Short dialogs, up to two to four turns, immediate personal needs ● Relies on very active listener ● Needs extensive assistance … (repetition, modification, demonstration, translation) Understands … ● Minimal understanding ● Short simple sentences ● Few factual details	Context ● Nondemanding ● Introduction ability to follow small groups Character ● Communication is short ● Slow rate ● Short dialogs, up to five turns, immediate personal needs ● Considerable assistance Understands … ● Instructions of a few simple sentences ● Gist of a dialog (starts to understand)	Context ● Immediate personal relevance ● Emerging ability to understand over the phone, if speech is slow and clear ● Some understanding in small groups interaction Character ● Communication is relatively short ● Some assistance ● Short dialogs, up to seven turns, personal relevance ● Short, informal monologues, dialogs, simple instructions ● Slow to normal rate Understands … ● Initial understanding of some implied meaning ● Initial awareness of some common registers/idioms ● Some compound sentences

Table D3 (Continued)

CLB 5	CLB 6	CLB 7	CLB 8
• Rely on contextual cues • Comprehend small group interactions/phone conversations • Understand some implied meanings • Understand narrative monologues/communication related to everyday/familiar/extremely relevant topics • Distinguish main ideas from supporting ideas • Begin to understand complex sentence structures • Get the gist and identify factual details • Interpret descriptions, reports, instructions • Clear speech at a slow to normal rate • Moderately demanding context • Requires repetition and recasts • Concrete language/most common vocabulary • Recognize some registers • Getting familiar with idiomatic expressions that are very common • Understand information/presentation that is up to 5 min • Requires visual support and setting to support meaning when the topic is less familiar • The number of turn taking is up to 10 • Initial ability to identify explicitly stated main ideas, supporting details, and implied meanings	• Begin to understand directions and instructions for technical or nontechnical tasks • Slow to normal speech rate • Initial ability to understand instructions that are clear and sometimes not presented step by step • Understand information, presentations, monologues up to 10 min • Understand common and predictable social exchanges • Begin to understand abstract concepts related to personal experiences only • Comprehend on the phone when context and topic are highly familiar • Understand complex structures (maybe not fully and with some effort) • Start to be able to make some inferences • Recognize some registers and style • Occasionally rely on contextual cues	• Understand common idiomatic language • The number of turn taking is up to 12 • Understand group discussions and interactions • Begin to see understanding of work-related topics, information possibly not presented in a sequence • Normal speech rate • Understand abstract concepts related to personal experience and general knowledge • Understand some specialized language and increasing range of complex sentences • Follow moderately complex phone interactions • Recognize expanding range of register and style • Identify emotional states • May use contextual clues to enhance comprehension • Emerging understanding of longer/semiformal monologues up to 15 min	• The number of turn taking is up to 15 • Understand abstract concepts related to specialized or work situations • Begin to understand extended multi-step instructions, technical or specialized matters • Initial ability to identify main ideas even when they are not explicitly stated • Follow colloquial or idiomatic conversations with some effort • Expanded understanding of registers and style • Follow communications with multiple speakers or in small groups about abstract and complex ideas on familiar topics • Understand monologues up to 20 min

Table D3 (Continued)

CLB 9	CLB 10	CLB 11	CLB 12
● Begin to understand academic discourse ● Debate in small groups ● Begin to understand paraphrasing and summarizing ● Begin to understand complex instructions that are presented in any order but clear and coherent ● Demanding context ● Understand some technical language ● Recognize bias and attitude ● Understand larger group conversation, live or recorded speech ● Limited only by the purpose of listening, not time ● Comprehend multistep directions or familiar procedures ● May require knowledge of academic or business codes in formal or social contexts ● Use context clue to support comprehension in high-stakes situations ● Initial ability to understand cultural references, low-frequency idioms, and humor with difficulty ● May sometimes miss details	● Expanding range of complex and detailed, formal and informal information ● Understand extensive lectures but not too complex ● Understand expository/argumentative exchange, discussion, or presentations between several speakers ● Comprehend multistep, but clear directions for less familiar procedures ● Infer unstated meanings ● Specialized topics in their own field ● Identify, analyze, and evaluate the critical aspect of communication ● Only occasionally miss details ● Begin to recognize/identify nuances in register, diverse style of speech, informal/formal, language variety ● Able to recognize transitional signals ● Separate facts from opinions ● Sometimes has difficulty interpreting humor ● Low-frequency idioms	● Fast speech rate ● Interpret nuances and meaning ● Occasional difficulty in understanding sarcasm, irony, figurative language ● Understand specialized topics outside the field of study ● Presents occasional difficulty in interpreting verbal humor ● Sufficiently grasp a detailed discussion of paraphrase and summary ● Able to understand instructions for unfamiliar procedures ● Identify most instances of hedging and face-threatening talk ● Instructions are complex in any order ● Recognize/identify most nuances in register, diverse style of speech, informal/formal, language variety ● Infers meaning from most unstated information	● Identify all instances of hedging and face-threatening talk ● Able to understand, analyze, and evaluate aural input from a variety of abstract topics delivered in various ways, including conference presentations, academic lectures, and aural discourse tainted with white noise ● Highly complex exchanges on most general and specialized topics ● Critically evaluate complex, detailed, specialized discussion/lectures and make suggestions to improve content ● Understand nuances and subtleties of complex speech from diverse speakers ● Understand unfigurative language ● Infers meaning from almost all unstated information ● Recognize/identify almost all nuances in register, diverse style of speech, informal/formal, language variety

Table D4 Reading

CLB 1	CLB 2	CLB 3	CLB 4
● Letters, numbers, short simple phrases ● Nondemanding contexts ● Very little ability to decode unknown words ● Relies heavily on graphics/visuals to interpret meaning ● Able to understand only everyday words, extremely limited vocabulary ● Clear font, lots of white space ● Little ability to apply sound-symbol relationships and spelling conventions ● Able to answer questions that do not require much writing (multiple choice, check, responding with an action) ● Able to understand information in very simple and short visuals	● Limited to everyday words and phrases ● Individual words and simple learned phrase ● Some short simple sentences related to immediate need ● Nondemanding contexts ● Little ability to decode unknown words ● Relies heavily on graphics/visuals to interpret meaning ● Limited vocabulary related to familiar everyday situations ● Clear font, lots of white space ● Very limited knowledge of basic grammar ● Instructions are one step ● Responses to reading do not require much writing (e.g., matching, MC [multiple choice], etc.)	● Nondemanding contexts ● Limited ability to decode unknown words ● Able to identify the main idea of very short sentences ● Start to understand some connected discourse/sentences (one or two paragraphs) ● Able to understand some information from very simple texts, on familiar topics ● Limited knowledge of basic grammar ● Maps, diagrams, basic/common tables ● Starting to get the gist of short social messages on daily topics ● Can get information from short simple notices (business/service) ● Can understand instructions of one to four steps	● Nondemanding contexts ● Some ability to decode unknown words ● Can identify main ideas and purpose ● Can identify some details and connectivity between short paragraphs ● Topics are familiar, personally relevant, and predictable ● Beginning to understand some common idiomatic language ● Can understand instructions of one to five steps ● Some knowledge of basic grammar ● Able to identify types and purposes of simple, short texts on familiar topics

Table D4 (Continued)

CLB 5	CLB 6	CLB 7	CLB 8
● Short reading texts ● Identify purpose, main ideas, important details ● Connect ideas within a text of up to one page or two paragraphs ● Understand texts on topics that are not personally relevant ● Often re-read and need clarification ● Moderately complex text in a predictable situation ● Ready to work with a bilingual dictionary ● More concrete words ● Occasionally guess the meaning of unknown/low-frequency words based on context or without using a dictionary ● Identify some styles and registers ● Access and locate basic information from reference sources ● Initial understanding of common and some abstract or specialized vocabulary and occasional high frequency idioms ● Clear instructional text of 7 – 10 steps related to everyday situations ● Able to compare simple information ● Identify moods and attitude in a specific social context	● Understand moderately complex social messages related to a personal experience or a familiar context ● Adequate range of moderately complex text ● Developing understanding of complex sentence structures ● Identify some implied meaning ● Able to locate some information to make comparison ● Sometimes guess the meaning ● Instructions up to 10 steps, not always presented step by step ● Uses a concise unilingual ESL dictionary ● Mostly concrete, factual, and descriptive texts with some vocabulary being abstract and specialized ● Understand literacy texts ● Read a text of up to three paragraphs ● Occasionally supported by visuals ● Expanding range of styles and registers ● Able to find up to two pieces of information	● Locate detailed information for comparison and contrast ● Read a text of up to five paragraphs ● Texts are factual, descriptive, argumentative, opinions that are both explicit and implied ● A range of different styles and registers ● Able to find up to three pieces of moderately complex text/information ● More frequently guessing meaning of unknown terms/words ● Starting to identify various styles/registers ● Topics are less predictable ● Uses a dictionary independently ● Able to interpret information in moderately complex tables and figures ● Instructions up to 10 – 13 steps ● Distinguish fact from opinion ● Context is less predictable but relevant social, educational, or work-related ● Can understand some concrete, abstract, and specialized vocabulary	● Locate and integrate information for comparison and contrast ● Read a text of up to 10 paragraphs ● Begins to have a wider range of different styles and registers ● Able to find up to four pieces of information ● Initial ability to identify mood, attitude ● Usually understands the meaning of new words from context but may have difficulties with cultural references and figurative language ● Understands an adequate range of complex sentences and structures ● Identify moods and attitude in a more complex context ● Beginning to understand familiar topics that are partially predictable ● Able to understand some unfamiliar topics

Table D4 (Continued)

CLB 9	CLB 10	CLB 11	CLB 12
• Texts are visibly complex and lengthy • Context is demanding and unpredictable • Identify a point of view • Specialized vocabulary • Begin to use inferences to integrate stated information across paragraphs • Initial ability to interpret nuances • Starting to deal with complex topics (reports, essays, novels, poems) • Texts start to include more complex features (tables, graphs, process flow charts, blueprints, etc.) • Topics and texts can be familiar or less familiar • Paraphrasing key points • Starting to identify implied information • Start to use figurative and not low-frequency idiomatic language • Separates relevant from irrelevant details • Start to conduct a complex search of online reference sources to research a defined topic that is limited in scope • Separates relevant from irrelevant details • Complex grammar and vocab used to interpret nuances • Knows all the AWL (570 words) • Understand texts conveying disagreement or conflict	• Understand multipurpose texts in many unfamiliar situations • Understands values and assumptions from stated and implied information • Use inferences to integrate stated information throughout the text • Sometimes have difficulty with low-frequency words/idioms • Can use dictionary to figure out low-frequency words/idioms • Expanding range of complex texts • Instructions are not presented in a step-by-step form • Interpret and summarize information and ideas contained in complex, formatted texts	• Instructions are complex and related to unknown procedures • Texts may be very long • Topics are partially familiar and starting to be unfamiliar • Interpret, summarize, and critically evaluate information • Language may be figurative and cultural references • Begins to use reference materials as required • Begin to understand, summarize, and outline the message, position, assumptions, bias, values and motives from fragments of different texts • Identifies line of reasoning and structure • Understand abstract and specialized texts • Occasional difficulty in interpreting low-frequency idioms and figurative language	• Interpret most low-frequency idioms and figurative language • Texts may require high-level inference • Understands most complex and unfamiliar academic or multipurpose written discourse • Interpret, summarize, and critically evaluate information public/semipublic business texts • Begin to evaluate content for validity, appropriateness, and relevance

Note. CLB = Canadian Language Benchmarks.

Appendix E

Sample Panelist Rating Form for Speaking (All Rounds)

	CLB 1	CLB 2	CLB 3	CLB 4	CLB 5	CLB 6	CLB 7	CLB 8	CLB 9	CLB 10	CLB 11	CLB 12
Round 1	—	☒	24	☒	36	☒	44	☒	64	68	☒	—
Round 2	10	12	22	32	36	40	44	52	62	68	72	75
Round 3	10	12	24	32	36	42	47	52	62	66	70	75

Note. CLB = Canadian Language Benchmarks.

Appendix F

Sample Panelist Forms for Listening

Question	CLB 1	CLB 2	CLB 3	CLB 4	CLB 5	CLB 6	CLB 7	CLB 8	CLB 9	CLB 10	CLB 11	CLB 12
1				x								
2					x							
3					x							
4					x							
5					x							
6						x						
7						x						
8					x							
9						x						
10						x						
11								x				
12									x			
13										x		
14											x	
(Items 15–54)	—	—	—	—	—	—	—	—	—	—	—	—
Total score (0–54)	0	1	3	6	21	32	33	39	41	50	53	54

Note. CLB = Canadian Language Benchmarks.

Round 3	CLB 1	CLB 2	CLB 3	CLB 4	CLB 5	CLB 6	CLB 7	CLB 8	CLB 9	CLB 10	CLB 11	CLB 12
Minimum expected listening section total score (0–54)	0	1	3	7	21	31	35	40	44	50	53	54

Note. CLB = Canadian Language Benchmarks.

Appendix G

Themes and Quotes From the Evaluation Survey

Theme	Example quote
Meeting process and procedure	
Guidance given by facilitators	● Even though some of our questions were not worded well, you seemed to figure out what we wanted to know and answered very well. (ID#07, L & R) ● Thank you for explaining the rationale and background to the various steps. As I am new to this, it was very helpful. (ID#13, L & R) ● Expectations for each task were all well explained! (ID#17, S & W)
Organization and efficiency of the meeting	● I am amazed at how well everything was organized and how smoothly those 4 days went! (ID#11, L & R) ● I thought the method used for the mapping process was helpful and efficient. (ID#05, S & W) ● The standard setting sessions were so well organized and efficient. (ID#18, S & W)
Group discussion	● [I thought] this kind of study would "just be a bunch of people talking … " Thus, you can see that I had no idea about the depth of thought required for this "just talking" experience. (ID#07, L & R) ● I really enjoyed the group activities about reviewing the JQC for the different levels. (ID#02, S & W) ● I found it very useful to hear from the other panelists and their input did sway my opinion several times. (ID#15, S & W) ● I would have preferred to dedicate more time to discussing the scores of the different writing levels after round 1. (ID#02, S & W) ● There were some less productive discussions about the test tasks. (ID#05, S & W)
Personal reflection	● I learned a lot and enjoyed every aspect of the process very much. (ID#03, L & R) ● Definitely will inform my future teaching as it relates to the CLBs. (ID#09, L & R) ● I learned a lot from everyone, including the highly professional presenters/coordinators and felt my contributions were appreciated. (ID#14, L & R) ● Thank you very much for this extremely rewarding and educational experience. (ID#16, L & R) ● I felt like everyone's voices were heard and maximum participation was encouraged. (ID#6, S & W)
Challenges and suggestions	
JQCs and judgment tasks	● Perhaps a Graphic Organizer for the JQC for each CLB would have been helpful. … that would help make the JQC document even more efficient and a quicker reference. (ID#13, L & R) ● There were challenges because the CLB and TOEFL are not exactly on the same page. … (ID#01, S & W) ● I felt that some discussions went off the rails a bit as some participants strayed away from the JQC requirements and more toward the descriptors of a particular CLB level. (ID#16, S & W) ● Some of the descriptors of JQCs are not very applicable. So, when we discuss the descriptors for the JQCs at each level, it may be good to direct the panelists' attention to the test tasks at some point so that some of the descriptors can be more relevant in the judgment task stage. (ID#18, S & W)
Other (logistics)	● … switch from Teams to Zoom, if at all possible! (ID#14, L & R) ● It would definitely be advantageous to have more set time to have breaks. (ID#16, S & W)

Note. L = listening; R = reading; S = speaking; W = writing.

Suggested citation:

Papageorgiou, S., Davis, L., Ohta, R., & Garcia Gomez, P. (2022). *Mapping* TOEFL® Essentials™ *test scores to the Canadian Language Benchmarks* (TOEFL Research Report No. RR-100). ETS. https://doi.org/10.1002/ets2.12357

Action Editor: John Norris

Reviewers: Liane Patsula and Ching-Ni Hsieh

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