

**Mapping Test Scores Onto the Canadian Language Benchmarks:
Setting Standards of English Language Proficiency on
The Test of English for International Communication (TOEIC),
The Test of Spoken English (TSE), and
The Test of Written English (TWE)**

Richard J. Tannenbaum

E. Caroline Wylie

Educational Testing Service, Princeton, NJ

April 2004

Abstract

The Canadian Language Benchmarks (CLB) describe language proficiency in reading, writing, speaking and listening on a 12-level scale that runs from Level 1 (Initial Basic) to Level 12 (Fluent Advanced). These levels provide guidance to language educators and instructors to identify existing levels of language competency of non-native communicators and to develop curriculum and courses to advance communicative competence. This paper describes a study conducted with a panel of English Language experts, from various regions in Canada, to map scores from three tests that collectively assess Reading, Writing, Speaking, and Listening on to three levels of the CLB. The panel recommended Level 4 (Fluent Basic Proficiency), Level 6 (Developing Intermediate Proficiency) and Level 8 (Fluent Intermediate Proficiency) cut scores for The Test of English for International Communication (TOEIC), The Test Of Spoken English (TSE), and The Test Of Written English (TWE). A modification of the Angoff (1971) standard-setting approach was used for multiple-choice questions, and a Benchmark Method (Faggen, 1994)—also referred to as an Examinee Paper Selection Method (Hambleton, Jaeger, Plake, & Mills, 2000)—was used for constructed-response questions.

Table of Contents

Introduction.....	1
Purpose of study.....	1
Canadian Language Benchmarks.....	1
Standard Setting.....	2
Section 1: Methods.....	3
Panelist Orientation.....	3
Panelist Training.....	4
Standard-Setting Process for Selected-response (Multiple-choice) Tests.....	5
Standard-Setting Process for Constructed Response Tests.....	9
Participants.....	11
Section 2: TOEIC Results.....	13
Linkage with the CLB.....	13
Cut Score Judgments.....	16
Section 3: TSE Results.....	17
Linkage with the CLB.....	17
Cut Score Judgments.....	18
Section 4: TWE Results.....	18
Linkage with the CLB.....	19
Cut Score Judgments.....	19
Summary and Conclusion.....	20
Statistical Distinctiveness Between Cut Scores.....	21
References.....	23

List of Tables

Table 1: Panel Demographics.....	12
Table 2: Listening Section Linkage Agreements.....	14
Table 3: Number Of Items Judged To Be At Each CLB Level For The Listening Section.....	14
Table 4: Section Linkage Agreements.....	15
Table 5: Number Of Items Judged To Be At Each CLB Level For The Reading Section.....	15
Table 6: First- And Second-Round TOEIC Judgments.....	16
Table 7: Number Of Items Judged To Be At Each Level For The TSE.....	17
Table 8: Cut Scores For The TSE.....	18
Table 9: Cut Scores For The TWE.....	19
Table 10: Summary Of Recommended Cut Scores.....	21
Table 11: Conditional Standard Error Of Measurement At Each Cut Score.....	22
Table 12: Distance Between Cut Scores In CSEMs.....	22

List of Figures

Figure 1: Hypothetical Angoff Ratings for three items.....	6
--	---

Introduction

Purpose of study

Currently, Citizenship and Immigration Canada (CIC) considers six criteria when reviewing applications for immigration. Each criterion has a point-value associated with it, for a grand total of 100 points. The criteria and point values are: Language (24 points), Education (25 points), Work Experience (21 points), Age (10 points), Arranged Job (10 points), and Adaptability (10 points). An applicant must earn 67 points to qualify. Educational Testing Service (ETS) is seeking designation by the CIC to be an authorized language-testing organization. If ETS is designated, applicants for immigration into Canada would be able to attempt to satisfy the Language criterion (a maximum of 16 points for the first Official Language) by taking The Test of English for International Communication (Listening and Reading), The Test of Written English (Writing) and The Test of Spoken English (Speaking) to demonstrate their English language ability.

In order to facilitate the use of these three tests for immigration purposes, ETS conducted a standard-setting study to map scores from these tests on to the Canadian Language Benchmarks (CLB). The study goals were (a) to document the alignment between the skills-content of each test and the skills-content of the corresponding Benchmarks (e.g., The Test of Spoken English was judged in relation to CLB Speaking) and (b) to identify minimum test scores (cut scores) that delineated three different proficiency levels on the CLB: Level 4 (Fluent Basic Proficiency), Level 6 (Developing Intermediate Proficiency), and Level 8 (Fluent Intermediate Proficiency). These three particular levels were identified by the CIC. In essence, the Level 4 cut score delineates the boundary or borderline between Levels 3 (Adequate Basic Proficiency) and 4 of the CLB; the Level 6 cut score delineates the borderline between Levels 5 (Initial Intermediate Proficiency) and 6 of the CLB; and the Level 8 cut score delineates the borderline between Levels 7 (Adequate Intermediate Proficiency) and 8 of the CLB. An expert-judgment approach was used to address each part of the study.

Canadian Language Benchmarks

The Canadian Language Benchmarks were first released in 1996 and revised in 2000. The Benchmarks describe language proficiency in four areas—Reading, Writing, Speaking, and

Listening—on a 12-level scale. The role of the Benchmarks is to provide a common framework on which to place adult immigrants according to their language proficiency in English or French, thus helping both learners and teachers better monitor progress. The Benchmarks are structured into three major groupings (basic, intermediate, and advanced) with four levels within each band (initial, developing, adequate and fluent). Thus CLB Level 1 is known as Initial Basic and CLB Level 12 as Fluent Advanced.

Skilled worker applicants for immigration are awarded increasingly more points for language proficiency according to whether they are deemed to have basic, moderate or high language proficiency. The thresholds (entrance points) of these three levels are considered to be the CLB Levels 4, 6, and 8, respectively. These three bands of proficiency (basic, moderate, high) apply independently to reading, writing, speaking, and listening.

Standard Setting

The process followed to map test scores onto the CLB is known as standard setting. Standard setting is a general label for a number of approaches used to identify test scores that support decisions about test takers' (candidates') level of knowledge, skill, proficiency, mastery, or readiness. For example, an international employer might require a non-native English speaker to achieve a certain score on The Test of English for International Communication (TOEIC) for placement in a job position in a predominantly English-speaking country. This cut score, set by each employer, reflects the minimum level of English language competence the particular employer believes necessary in order for an employee to function successfully in a particular role and setting. The score reflects a standard of “readiness to perform job tasks in English” for that position. People with TOEIC test scores at or above the cut score have demonstrated a sufficient level of English proficiency; those with test scores below the cut score have not yet demonstrated a sufficient level of English language proficiency to function in that role. In this example, one cut score classifies test-takers into two levels of proficiency; more than one cut score may be established on the same test to classify candidates into multiple levels of proficiency.

It is important to recognize that a cut score, a threshold test score, is a function of informed expert judgment. There is no absolute, unequivocal cut score. There is no single “correct” or “true” score. A cut score reflects the values, beliefs, and expectations of those experts who participate in its definition and adoption, and different experts may hold different

sets of values, beliefs, and expectations. Its determination may be informed by empirical information or data, but ultimately, a cut score is a judgment-based decision.

As noted by the *Standards for Educational and Psychological Testing* (1999), the rationale and procedures for a standard-setting study should be clearly documented. This includes the method implemented, the selection and qualifications of the panelists, and the training provided. With respect to training, panelists should understand the purpose and goal of the standard-setting process (e.g., what decision or classification is being made on the basis of the test score), be familiar with the test, have a clear understanding of the judgments they are being asked to make, and have an opportunity to practice making those judgments. The standard-setting procedures in this study were designed to comply with these guidelines; the methods and results of the study are described below.

This report is presented in five major sections. The first section describes the standard-setting methods (for the selected-response and constructed-response tests) that were implemented to establish the threshold scores corresponding to Levels 4, 6 and 8 on the CLB for each of the English language tests. This section also includes a description of the study participants. The next three sections, in turn, present the results for three tests. The fifth section presents an overall summary and conclusion.

Section 1: Methods

Panelist Orientation

Panelists were provided with an overview of the purpose of the study and a definition of threshold scores (or cut scores), as applied to the current purpose. Appendix A provides the agenda for the study. Cut scores were defined as the level of performance on each of the tests that reflected the English language proficiency of a candidate who was just at Level 8, just at Level 6, and just at Level 4 on the CLB. Each cut score was defined as the minimum score believed necessary to qualify a candidate at each of the three levels. The panelists were also provided with brief overviews of each of the tests for which they would be mapping scores onto the CLB (setting cut scores).

- Test of English for International Communication (TOEIC). The TOEIC measures the ability of non-native English communicators to communicate in English in the global workplace. The TOEIC addresses listening comprehension skills and reading

comprehension skills. The test items are developed from samples of spoken and written English from countries around the world. The TOEIC is a selected-response test that is reported on a scale that ranges from a low of 10 to a high of 990. Score reports provide both candidate's section-level and total scores.

- Test of Spoken English (TSE). The TSE measures the ability of non-native speakers of English to communicate orally in English. It consists of nine items for which a candidate must generate a verbal response involving, for example, narration, persuading, recommending, and giving and supporting an opinion. Responses to each item are scored using a rubric ranging from a low of 20 to a high of 60 in 10-point intervals. As many as 12 independent assessors contribute to a candidate's overall TSE score. Item scores are averaged to arrive at the overall score, which is reported in intervals of five: 20, 25, 30, 35, 40, 45, 50, 55, 60.
- Test of Written English (TWE). The TWE measures the ability of non-native writers of English to produce an essay in response to a given topic, demonstrating their ability to generate and organize ideas, to support those ideas with examples, and to use conventions of standard written English. The response is scored using a rubric ranging from a low of 1 to a high of 6 in 1-point intervals. Two independent assessors score the response and an average score is computed; the overall TWE score, therefore, is reported in half-point intervals: 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0, 5.5, 6.0.

Different reporting scales are used across the tests (TOEIC, TSE, TWE) to avoid confusion and to help ensure that one score is not substituted for a score on another test that has a different meaning.

Panelist Training

The first major event of the training process had panelists summarizing the key descriptors of the Canadian Language Benchmarks. This was done in two small groups. Panelists had been sent a homework task to review the CLB Levels 4, 6, and 8 for each language area and to select critical descriptors that defined each level (See Appendix B). In their small groups, panelists were asked to consider their homework responses, focusing on the English-language skill(s) being measured by the particular test that was the immediate focus. The first test section

to be addressed was the TOEIC Listening section; therefore, Levels 8, 6, and 4 of the CLB Listening section were summarized. One group focused on what distinguished a Level 6 candidate from a Level 4 candidate in listening skills while the other group focused on what distinguished a Level 8 candidate from a Level 6 candidate. Each group's charted summary was posted and discussed so that the whole panel had an opportunity to comment and, as appropriate, suggest modifications. The whole-panel agreed-upon summaries remained posted to guide the standard-setting judgment process for the TOEIC Listening section. Collectively, the whole group then spent time considering what the listening skills would be of a candidate who was above a CLB Level 8. The charts generated by the two small groups are presented in Appendix C. (The charts differ in format, which reflects how the groups approached the exercise.)

The exercise of summarizing the CLB Levels 8, 6, and 4 was repeated in turn for each language skill addressed by the test of focus. Once the standard-setting judgments were completed for the TOEIC Listening section, the TOEIC Reading section was presented, so the summary process was repeated for Reading. After standard-setting judgments were completed for the TOEIC, the TSE became the focus, followed by the TWE.

Standard-Setting Process for Selected-response (Multiple-choice) Tests

The Listening and Reading sections of the Test of English for International Communication (TOEIC) consists of 100 selected-response items each, in which candidates chose or select a response to an item from a given set of options. The same approaches that were used to determine cut scores and to judge content alignment for the Listening section were applied to the Reading section. For the Listening section, however, panelists listened to taped-recorded speaking stimulus for each item, whereas for the Reading section they read printed text.

The general standard-setting process applied to the TOEIC is known as the Angoff Method (Angoff, 1971). The general approach remains the most widely used standard-setting method for selected-response tests (Mehrens, 1995; Cizek, 1993; Hurtz & Auerbach, 2003). The first section of the TOEIC test addressed was Listening. This section measures the ability of a non-native communicator to comprehend spoken English. As applied to the Listening section, panelists were asked to read an item, listen to the stimulus for that item, consider the difficulty of the English-language skill addressed by the item, and to judge three separate probabilities: that a Level 8, Level 6, and Level 4 candidate would know the correct response. Level 8 was the first

judgment to avoid a potential ceiling effect. A ceiling effect could occur if panelists began the judgment process for a Level 4 cut score and set too high an English-language proficiency expectation, restricting the range of the score scale available for Level 6 and Level 8 cut scores.

Panelists recorded their item-level judgments on a form (see the Appendix D for a copy of the judgment form used for the Listening section of the TOEIC), with the following probability scale: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9. A judgment of 0.1, for example, corresponds to a 10 percent probability of knowing the correct answer. As a rule-of-thumb, panelists were informed that a difficult item—that is, one that requires a relatively high-level of English proficiency—might fall into the range of 0.1 to 0.3: a 10- to 30-percent probability of knowing the correct answer. A relative easy item might fall into the 0.7 to 0.9 range: 70- to 90-percent probability of knowing the correct answer; and a moderately difficult item might fall into the range of 0.4 to 0.6: 40- to 60-percent probability of knowing the correct answer. For each panelist, the sum of the Level 8 probabilities represents that panelist’s Level 8 recommended cut score. Similarly the sum across the Level 6 and Level 4 item probabilities represents the Level 6 and Level 4 cut scores recommended by each panelist. Cut scores were then averaged across all panelists to determine the first-round average recommended cut scores. Figure 1 illustrates item-level judgments for three items made by one panelist. The sum of the hypothetical item probabilities in each column represents this panelist’s three cut score recommendations for Levels 8, 6, and 4—2.4, 1.8, and 1.1, respectively.

	Circle the probability that a <u>Level 8</u> candidate would get the item correct	Circle the probability that a <u>Level 6</u> candidate would get the item correct	Circle the probability that a <u>Level 4</u> candidate would get the item correct
1	0.10.2 0.3 0.4 0.5 0.6 0.7 <u>0.8</u> 0.9	0.10.2 0.3 0.4 0.5 <u>0.6</u> 0.7 0.8 0.9	0.10.2 <u>0.3</u> 0.4 0.5 0.6 0.7 0.8 0.9
2	0.10.2 0.3 0.4 0.5 0.6 <u>0.7</u> 0.8 0.9	0.10.2 0.3 0.4 <u>0.5</u> 0.6 0.7 0.8 0.9	0. <u>0.2</u> 0.3 0.4 0.5 0.6 0.7 0.8 0.9
3	0.10.2 0.3 0.4 0.5 0.6 0.7 0.8 <u>0.9</u>	0.10.2 0.3 0.4 0.5 0.6 <u>0.7</u> 0.8 0.9	0.10.2 0.3 0.4 0.5 <u>0.6</u> 0.7 0.8 0.9

Figure 1: Hypothetical Angoff Ratings for three items.

Panelists were also asked to make a dichotomous (“yes” or “no”) judgment concerning the content alignment or linkage between each test item and the CLB. Panelists were asked to consider the question “Does the English language skill measured by the item address an English language skill covered by the corresponding CLB modality (language area)?” It was clarified for panelists that a “yes” response to items would reflect their judgment that the skills measured by

the items were included in the corresponding CLB description. It would not mean, however, that all of the CLB skills for a language area necessarily were reflected by the test items. The alignment question, in essence, asked about the presence or absence of a skill-based connection, not about the extent of skill domain coverage. For the linkage part of the data collection, panelists were not asked to attribute an item to a particular CLB level, only to determine whether the skill represented by the item was addressed in the corresponding Benchmark description.

In addition to skills alignment, the CIC was interested in information about the classification of test items relative to the three CLB levels targeted in the study. The fundamental question posed was, “how are the test items distributed across the three levels?” This information was derived from the Angoff judgments used to arrive at the cut scores for each level. In order to attribute a particular level on the CLB to each item, first panelists’ responses to the dichotomous question regarding the connection between each item and the CLB were examined. A threshold of at least 11 of the 15 panelists (73%) affirming the connection was established, a priori, as the needed level of agreement to judge an item as aligned with the CLB. (This criterion reflects a clear majority of panelists.) The Angoff ratings were then used to infer a particular level on the CLB for each item. An item was associated with the first of the three levels for which probability judgments met or exceeded 0.6 (60% probability of a candidate at that level answering the item correctly). Referring again to Figure 1 and using the criterion of 0.6, Item 1 would be considered to be at Level 6. For Item 2 the probability judgments do not meet or exceed 0.6 until Level 8, whereas Item 3 would be classified as a Level 4 item. For each panelist the CLB level for each test item under review was similarly derived. By item, across the panelists the modal classification was calculated since the mode indicates the level with the maximum panelist agreement.

Prior to making their “live” first-round standard-setting judgments for the Listening items, panelists were given an opportunity to practice making judgments on five sample Listening items from a previously administered (1997) edition of the TOEIC. For each sample item, each panelist was asked to answer yes/no to the alignment question and to record the probability that a Level 8, Level 6, and Level 4 candidate would know the correct answer (practice recording forms were provided.) Once each panelist noted his or her response, a whole-group discussion occurred whereby panelists were asked to share their item-level decision rationales. After the discussion of each item, the correct answer was revealed, as was the

proportion of approximately 10,000 randomly sampled examinees that chose the correct answer, and whether the item would be classified as being easy, of medium difficulty, or difficult, based on our rule-of-thumb guidelines. (It was clarified that these percent correct values were based on the general population of TOEIC examinees and that the panel's task was to consider how an examinee at Level 8, Level 6, and Level 4 would perform.) The practice session helped to calibrate the panelists and to make explicit the diversity of relevant professional perspectives reflected by the panel. The practice session also helped to clarify any misunderstanding of the judgment process. At this point, panelists were formally asked to acknowledge if they understood what they were being asked to do and the overall judgment process. They did this by signing a training evaluation form confirming their understanding and readiness to proceed. In the event that a panelist was not yet prepared to proceed, he or she would have been given additional training by one of the ETS facilitators. All panelists signed off on their understanding and readiness to proceed. Panelists were then asked to complete their "live" judgments for the first three items of the Listening section and then to stop. This provided an opportunity to answer panelists' questions. The panelists confirmed that they understood the process and were then asked to complete their round-one judgments for the Listening section.

The ETS facilitators computed each panelist's Level 8, 6, and 4 standard-setting judgments for the TOEIC Listening section, summing the probabilities across the 100 items, first for the Level 8 judgments then for the Level 6 judgments, and finally for the Level 4 judgments. For example, if a panelist had recorded 0.8 for each of the 100 items for a Level 8 candidate, that panelist's Level 8 cut score would be 80; so according to that panelist 80 items would need to be answered correctly for a candidate to be considered at the Level 8 on the CLB. If a panelist had recorded a 0.5 for each of the 100 items for a Level 6 candidate, that panelist's Level 6 cut score would be 50; so according to that panelist, 50 items would need to be answered correctly for a candidate to be considered at the Level 6 on the CLB. The average Level 8, Level 6 and Level 4 cut scores across all panelists were computed, as was the median, standard deviation, minimum score, and maximum score at each level. The cross-panelist summary information was posted and used to facilitate a discussion. Each panelist also had his or her own Level 8, Level 6, and Level 4 TOEIC Listening cut scores. In general, the panelists with the minimum score and maximum score were asked to begin the discussion, with other panelists encouraged to share their judgments. At the conclusion of the group discussion, the panelists were given an

opportunity to change their overall Level 8, 6, and 4 TOEIC Listening scores. Panelists were reminded that they could keep their first-round section-level scores; they were not obligated or expected to change their scores. Panelists then recorded their second-round (final) judgments.

This same process of practice and discussion followed by “live” round-one judgments, discussion, and a final (round-two) judgment, was followed for the 100 Reading items of the TOEIC.

Standard-Setting Process for Constructed Response Tests

The TSE and the TWE are considered constructed-response tests in that candidates are required to produce original responses, not to select from a set of given options, as in the case of selected-response tests. The standard-setting process as applied to the TSE will be described in some detail. An abbreviated presentation of the process will follow for the TWE because the same process was used in both cases.

The standard-setting process applied to the TSE is variously known as the Benchmark Method (Faggen, 1994) or the Examinee Paper Selection Method (Hambleton, Jaeger, Plake, & Mills, 2000). As applied to the TSE, the process included the panelists first reviewing the nine items of the TSE and the scoring rubric. Operationally, the panelists were asked to read a TSE item and to listen to sample spoken responses to the item that served to illustrate each whole-number score point on the rubric (20, 30, 40, 50, 60). The panelists were asked to consider the difficulty of the English language skill addressed by the item, the language features valued by the rubric, and the skill set of a Level 8 candidate (as previously defined). Panelists, independently, were asked to pick the lowest scoring sample response that, in their expert judgment, most appropriately reflected the response of a candidate who was just at Level 8 proficiency on the CLB. Because, as noted previously, TSE responses are averaged, panelists were able to pick from among the range of reported scores (20, 25, 30, 35, 40, 45, 50, 55, 60). So for example, if a panelist believed that a Level 8 candidate would score higher than a 50 on an item, but not quite as high as a 60, the panelist would be able to pick a score of 55. They were then asked to repeat the judgment process for a candidate at CLB Level 6 and Level 4. This basic process was followed for each of the nine TSE items. Panelists independently completed their Levels 8, 6, and 4 judgments for the first TSE item and were asked to stop. Panelists were then asked to share their judgments for the first item—what scores did they give for the Level 8, 6, and 4 candidates?

The purpose of the facilitated discussion was for panelists to hear the judgment rationales of their peers. The goal was to make more explicit the diversity of relevant perspectives reflected by the panel and to give panelists an opportunity to consider a viewpoint that they had not previously considered; the goal was not to have panelists conform to a single expectation of performance levels for CLB Levels 8, 6, and 4 on TSE items. This practice opportunity was also used to clarify any misunderstandings of the judgment process. Panelists were given the chance to change their Level 8, 6, and 4 judgments for the first item before proceeding, independently, on to the remaining eight items of the TSE. The completion of the Level 8, 6, and 4 judgments for all nine of the TSE items was considered to be first-round judgments.

The ETS facilitators computed each panelist's Level 8, 6, and 4 standard-setting judgments for the TSE, taking the average score across the nine items for each panelist. The average Level 8, Level 6 and Level 4 cut scores across all panelists were computed, as was the median, standard deviation, minimum cut score, and maximum cut score for each level. The cross-panelist summary information was posted and used to facilitate a discussion. Each panelist also had his or her own Level 8, 6, and 4 TSE cut scores. In general, the panelists with the minimum score and maximum score were asked to begin the discussion, with other panelists encouraged to share their judgments and decision rationales. At the conclusion of the group discussion, the panelists were given an opportunity to change their overall Level 8, 6, and 4 TSE cut scores if they felt that they wished to reflect some aspect of the discussion in their final judgment. Panelists were reminded that they could keep their first-round scores; they were not obligated or expected to change their scores. Panelists then recorded their second-round (final) judgments. (See the Appendix D for a copy of the judgment recording form completed by each panelist.)

Similar to the alignment question asked for each TOEIC item, panelists had also been asked to indicate whether the English language skill measured by each TSE item addressed an English language skill covered by the CLB Speaking description. The same criterion of 11 of 15 panelists responding "yes" was used to confirm the linkage to the CLB Speaking description. The nine TSE items were classified in relation to the three CLB levels using a different approach from that applied to the TOEIC items. In this instance—where no item-level probabilities were collected—classifications were derived by locating the item at the first level for which panelists' selected 40 (or greater) as the benchmark score. This mark was chosen as it reflects the transition

point on the TSE rubric between weaker and stronger performances. (A 40 reflects “somewhat effective communication,” while a 30 reflects “generally not effective communication.”) The modal classification was calculated.

This overall standard-setting process was also applied to the Test of Written English (TWE). The TWE is also a constructed-response assessment for which candidates produce an essay in response to a given topic. There is only one topic, so, in essence, the TWE is a single-item test. As with the TSE, panelists reviewed the essay topic and scoring rubric. They then reviewed sample essays illustrative of each of the rubric score points. Panelists, independently, were asked to pick the sample response that, in their expert judgment, reflected most appropriately the response of a candidate just at CLB Level 8 proficiency, just at Level 6 proficiency, and then just at Level 4 proficiency. As with the TSE, panelists were able to use the full reporting scale. So, for example, if a panelist believed that a Level 8 candidate would score higher than a 5, but not quite as high as a 6, the panelist would be able to pick a cut score of 5.5. The first-round of independent judgments was followed by a whole-group discussion. Panelists were then given the opportunity to change their Level 8, 6, and 4 judgments. Panelists also made an alignment judgment for the TWE; the criterion of 11 of 15 panelists responding “yes” was applied. The item classification (TWE is a one-item test) was defined by the lowest level for which panelists selected a score of 4 as the benchmark score. Similar to the TSE, this mark was chosen as it reflects the transition point on the TWE rubric between weaker and stronger performances.

Participants

The panel consisted of 15 Canadian language experts who were familiar with the Canadian Language Benchmarks and with the test-taking population. A senior executive from TOEIC Services Canada organized the recruitment of the experts. Initially, all of the CLB experts listed with the Centre for Canadian Language Benchmarks were invited. Several were able to participate but a number of openings remained. As a result, the TESL provincial offices were contacted and asked to suggest CLB experts in the province. TOEIC Services Canada selected 15 panelists, with consideration for diversity of geographic location and gender. The panel consisted of ESL teachers and assessment experts, including those in professional workplaces, who are involved in assessment decisions. Table 1 describes the demographic

characteristics of the panel. Panelists were also asked to identify the levels on the CLB with which they were most familiar. The majority of the panelists reported working with persons who span CLB Level 1 through 10. Appendix E provides the panelists’ affiliations and brief bios of the panelists and two ETS researchers who conducted the study.

Table 1: Panel Demographics

	Number	Percent
Gender		
Female	10	67%
Male	5	33%
Area of Expertise ¹		
Assessment	6	
Curriculum	5	
Language instruction for immigrants	2	
Workplace programs	4	
Research	3	
Province		
Alberta	5	33%
British Columbia	3	20%
New Brunswick	1	7%
Newfoundland	1	7%
Ontario	4	27%
Saskatchewan	1	7%
CLB Experience		
1 to 6	2	13%
1 to 8	4	27%
1 to 10	8	53%
6 to 10	1	7%

¹ Some members met more than one criterion so percentages are not reported.

Section 2: TOEIC Results

The Test of English for International Communication (TOEIC) is a two-hour, selected-response test designed to assess English language skills (listening and reading) in examinees for whom English is not their native language. Within each section, there are multiple item-types to assess different aspects of listening and reading proficiency. The Listening section consists of four sub-sections:

- Selection of most appropriate description of a photograph (20 items)
- Selection of most appropriate response to a question or statement (30 items)
- Selection of most appropriate response to a question based on a short conversation (30 items)
- Selection of most appropriate response to a question based on a longer conversation or talk (20 items)

The Reading section consisted of three sub-sections:

- Selection of most appropriate response to complete a sentence (40 items)
- Selection of the grammatical/syntactical error in a given sentence (20 items)
- Selection of most appropriate response to questions based on reading passages (40 items)

Linkage with the CLB

In response to the question “Does the English language skill measured by the item address an English language skill covered by the corresponding Listening CLB description?” Eleven or more of the panelists indicated the affirmative for every TOEIC Listening item, satisfying the alignment criterion. For each subsection in the Listening section, Table 2 provides the average percent of panelists who agreed that there was a linkage between the skill(s) addressed by the items and the descriptors in the Listening section of the CLB. An overall Listening percentage is also provided, which reflects the question-weighted average of the section linkages. On average, 95% of the panelists agreed that each item was aligned with the corresponding CLB Listening description.

Table 2: Listening Section Linkage Agreements

Subsection	Average percentage
Photographs	97%
Response to comment	94%
Short conversations	94%
Longer conversations or talk	96%
Average Listening	95%

Table 3 reports the results of the item-classification analysis, by subsection, and for the overall Listening section. Overall most of the Listening items were classified as being at Level 6, with more than one-quarter being classified at Level 8. Comparatively few items were classified as being at Level 4. The proportions of items at each level vary considerably by subsection with most of the Level 4 items occurring in the first subsection. The results seem to match the anecdotal comments made by some of the panelists who indicated that few items were directly accessible to Level 4 candidates.

Table 3: Number Of Items Judged To Be At Each CLB Level For The Listening Section

Subsection	Level 8		Level 6		Level 4	
	N	%	N	%	N	%
Photographs (20 items)	2	10%	11	55%	7	35%
Response to comment (30 items)	2	7%	27	90%	1	3%
Short conversations (30 items)	13	43%	17	57%	0	0%
Longer conversations or talk (20 items)	10	50%	9	45%	1	5%
Total (100 items)	27	27%	64	64%	9	9%

In response to the question “Does the English language skill measured by the item address an English language skill covered by the Reading CLB?” Eleven or more of the panelists indicated the affirmative for 40 of the 100 reading items, clearly differentiating their judgments by subsection. All 40 of the Reading Comprehension items were judged to be linked to the CLB Reading description, but none of the Sentence Completion items or any of the Error Recognition items met the alignment criterion. Panelists did not see the skills measured by these items as

primarily reading skills. When asked, some panelists agreed that those items did address skills found within other language areas within the CLB (e.g., Writing), but they did not agree with assessing them out of their appropriate context, and the alignment task was only focused on connections to the CLB reading description.

For each subsection in the Reading section, Table 4 provides the average percent of panelists who agreed that there was a linkage between the skills addressed by the items and the descriptors in the Reading section of the CLB. An overall Reading percentage is also provided, which reflects the question-weighted average of the section linkages. On average, 61% of the panelists agreed that each item was aligned with the corresponding CLB Reading description.

Table 4: Section Linkage Agreements

Subsection	Average percentage
Sentence Completion	40%
Error Recognition	27%
Reading Comprehension	99%
Average Reading	61%

Table 5 reports the results of the item-classification analysis, by subsection and for the overall Reading section. Note that the 40 Reading Comprehension items were the ones judged to be linked to the CLB. Similar to the Listening results, most items were classified at Level 6, with more than one-quarter being classified as Level 8, and very few items classified at Level 4.

Table 5: Number Of Items Judged To Be At Each CLB Level For The Reading Section

Subsection	Level 8		Level 6		Level 4	
	N	%	N	%	N	%
Sentence Completion (40 items)	10	25%	29	73%	1	3%
Error Recognition (20 items)	11	55%	9	45%	0	0%
Reading Comprehension (40 items)	7	18%	32	80%	1	3%
Total (100 items)	28	28%	70	70%	2	2%

Cut Score Judgments

The first-round and second-round cut score judgments for the TOEIC Listening and Reading Section are presented in Appendix F Tables A (Listening) and B (Reading). Each panelist’s individual Level 8, 6, and 4 cut scores are presented for each round, as are the cross-panel summary statistics (mean, median, standard deviation, minimum and maximum.) Table 6 presents the Level 8, 6, and 4 cross-panel statistics for both sections. The TOEIC section level scaled score means and medians were obtained from a raw-to-scaled score conversion table for the TOEIC. The total TOEIC cut scores are the sum of the two section cut scores.

Table 6: First- And Second-Round TOEIC Judgments

	Round 1 Judgments			Round 2 (final) Judgments		
	Mean	Median	SD	Mean	Median	SD
<u>Level 8</u>						
Listening (raw scores)	77	79	7	76	77	6
Listening (scaled scores)	435	450	46	430	435	37
<u>Level 6</u>						
Listening (raw scores)	60	58	11	59	60	7
Listening (scaled scores)	325	310	71	320	325	48
<u>Level 4</u>						
Listening (raw scores)	31	30	11	29	28	7
Listening (scaled scores)	130	125	72	115	110	41
<u>Level 8</u>						
Reading (raw scores)	77	76	8	76	75	7
Reading (scaled scores)	365	360	46	360	355	38
<u>Level 6</u>						
Reading (raw scores)	60	59	13	59	60	10
Reading (scaled scores)	270	265	73	265	270	57
<u>Level 4</u>						
Reading (raw scores)	28	28	10	26	28	6
Reading (scaled scores)	90	90	57	75	90	32
<u>Level 8</u>						
TOEIC (raw scores)	155	154	14	153	154	12
TOEIC (scaled scores)	800	820	85	790	805	72
<u>Level 6</u>						
TOEIC (raw scores)	120	124	21	118	120	15
TOEIC (scaled scores)	595	620	130	585	595	92
<u>Level 4</u>						
TOEIC (raw scores)	59	58	20	55	55	11
TOEIC (scaled scores)	220	205	121	190	190	64

The Reading and Listening Level 8, 6, and 4 raw cut score means (and medians) decreased slightly for all levels from round one to round two as can be seen in Table 6. The variability (standard deviation) of the panelists’ judgments also decreased from round one to round two for all levels, indicating a greater degree of panelist consensus.

The second-round mean scaled scores may be accepted as the panel-recommended cut scores, that is the minimum scores necessary to qualify for Levels 8, 6, and 4 on the CLB. Thus the TOEIC Level 8, 6, and 4 scaled cut scores for Listening are 430, 320 and 115, respectively, and for Reading they are 360, 265, and 75, respectively. The total TOEIC cut scores are 790, 585, and 190, for Levels 8, 6, and 4, respectively.

Section 3: TSE Results

The Test of Spoken English (TSE) assesses speaking language skills in a nine-item constructed-response format. Each of the nine responses is scored according to a five-point rubric (20 to 60, in 10-point increments). The overall score is the average across item scores, and is reported in intervals of five: 20, 25, 30, 35, 40, 45, 50, 55, 60.

Linkage with the CLB

In response to the question “Does the English language skill measured by the item address an English language skill covered by Speaking CLB?” Eleven or more of the panelists indicated the affirmative for all nine items, satisfying the alignment criterion. The average percentage of panelists who agreed that there was a linkage between the skill(s) addressed by the items and the descriptors in the Speaking section of the CLB was 96%. Table 7 reports the results of the item-classification analysis. All but one item was classified at Level 6.

Table 7: Number Of Items Judged To Be At Each Level For The TSE

Level	Number of items	%
8	1	11%
6	8	89%
4	0	0%

Cut Score Judgments

The first-round and second-round cut score judgments for the TSE are presented in Appendix G. Each panelist’s individual Level 8, 6, and 4 cut scores are presented for each round, as are the cross-panel summary statistics (mean, median, standard deviation, minimum and maximum.) Table 8 summarizes the results for the Round 1 and Round 2 cut score judgments that the panelists made on the TSE. The presented TSE scores reflect the reporting scale for this test. The means decreased slightly and the standard deviations increased slightly from Round 1 to Round 2.

Table 8: Cut Scores For The TSE

	Round 1 Judgments			Round 2 (final) Judgments		
	Mean	Median	SD	Mean	Median	SD
Level 8	52	51	3.1	48	50	3.1
Level 6	41	41	2.0	38	40	3.6
Level 4	31	30	4.4	27	25	4.6

The second-round mean scores may be accepted as the panel-recommended cut scores, that is the minimum scores necessary to qualify for Levels 8, 6, and 4 on the CLB. Thus the TSE Level 8, 6, and 4 cut scores are 50², 40³ and 25⁴ respectively.

Section 4: TWE Results

The Test of Written English (TWE) assesses written language skills in a constructed-response format. TWE is a 30-minute examination of a candidate’s ability to respond in writing to a single prompt, thus providing information about candidates’ ability to generate and organize ideas on paper, to support those ideas with evidence or examples, and to use the conventions of standard written English. Essays are scored according to a seven-point rubric (0 to 6). Two raters score each essay independently, and the reported score is average of these two scores. Thus the

² The TSE Round 2 mean level 8 judgment was 48, but the reporting scale is in increments of 5. Thus, the level 8 cut score is 50.

³ The TSE Round 2 mean level 6 judgment was 38, but the reporting scale is in increments of 5. Thus, the level 6 cut score is 40.

⁴ The TSE Round 2 mean level 4 judgment was 27, but the reporting scale is in increments of 5. Thus, the level 4 cut score is 25.

score scale ranges from 0 to 6 in half point increments (0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0, 5.5, 6.0).

Linkage with the CLB

In response to the question “Does the English language skill measured by the item address English language skills covered by Writing CLB?” All (100%) of the panelists indicated the affirmative, thus satisfying the alignment criterion. The item-classification analysis for the essay prompt classified it at Level 8.

Cut Score Judgments

The first-round and second-round cut score judgments for the TWE are presented in Appendix H. Each panelist’s individual Level 8, 6, and 4 cut scores are presented for each round, as are the cross-panel summary statistics (mean, median, standard deviation, minimum and maximum.) Table 9 summarizes the results for the Round 1 and Round 2 cut score judgments that the panelists made on the TWE. The presented TWE scores reflect the reporting scale for this test.

Table 9: Cut Scores For The TWE

	Round 1 Judgments			Round 2 (final) Judgments		
	Mean	Median	SD	Mean	Median	SD
Level 8	4.7	5.0	0.7	4.5	4.5	0.5
Level 6	3.3	3.5	0.6	3.2	3.5	0.4
Level 4	1.3	1.0	0.7	1.3	1.5	0.6

The mean judgments were slightly lower for two of the three cut scores and the standard deviations decreased slightly from Round 1 to Round 2. The second-round mean scores may be accepted as the panel-recommended cut scores, that is, the minimum scores necessary to qualify for Levels 8, 6, and 4 on the CLB. Thus the TWE Level 8, 6, and 4 cut scores are 4.5, 3.0⁵ and 1.5⁶ respectively. As a side note, one panelist initially felt quite strongly that a Level 4 candidate

⁵ The TWE Round 2 mean level 6 judgment was 3.2. The reporting scale is in increments of 0.5. Thus, the level 6 cut score is 3.0.

⁶ The TWE Round 2 mean level 4 judgment was 1.3. The reporting scale is in increments of 0.5. Thus, the level 4 cut score is 1.5.

would struggle so much with the reading demand of the essay prompt itself that he or she would be unable to produce an essay that would get any score above a zero, although as a result of the discussion between Rounds 1 and 2, the panelist increased the Level 4 cut score.

Summary and Conclusion

The purpose of this study was to arrive at Canadian Language Benchmark (CLB) Level 8, Level 6, and Level 4 recommended cut scores on a series of language proficiency tests, thus creating an operational bridge between the descriptive levels of the CLB and standardized tests of English language proficiencies. A panel of 15 experts was invited to participate in the standard-setting study. The Benchmark Method (Faggen, 1994)—also referred to as the Examinee Paper Selection Method (Hambleton, Jaeger, Plake, & Mills, 2000)—and a modification of the Angoff Method (1971) were applied to the constructed-response questions and selected-response questions respectively.

In the process of going through the linkage and standard-setting process, panelists, on several occasions expressed some reservations about the experiences of a Level 4 candidate since they felt that the majority of all three tests would be difficult for these low proficiency candidates, and thus the test-taker's experience would be very discouraging for him or her. In addition, the nature of these assessments differs from the CLB-Assessment (CLBA), used in Canada. The CLBA is a progressive assessment, meaning that candidates stop once they reach a level beyond which they cannot perform adequately. Since this assessment requires highly trained assessors to administer it, it is not feasible for world-wide implementation, as is required by Citizenship and Immigration Canada. The panelists struggled somewhat with the differences between the assessment with which they were familiar and the three tests examined during the study. Several panelists noted that while the tasks in the TSE addressed skills found within the CLB, there were aspects of the Speaking CLB not addressed by the assessment, and the lack of non-verbal cues given the form of the assessment was seen as potentially presenting an additional challenge to lower level candidates.

Together the two sections (Listening and Reading) of TOEIC, with the Test of Spoken English and the Test of Written English address the four modalities of the Canadian Language Benchmarks (CLB), although the match between the TOEIC Reading section and the CLB is the

weakest. The panelists were concerned, for example, that aspects of the TOEIC Reading section addressed skills that they did not consider part of the CLB Reading domain and this concern was reflected in their linkage ratings for two of the three subsections of the reading. The item-classification analysis indicated that the majority of items on each assessment tended to be at Level 6, with approximately one quarter of the items at Level 8. In general few items were indicated at Level 4. Table 10 below summarizes the Level 8, 6, and 4 cut scores for each of the tests.

Table 10: Summary Of Recommended Cut Scores

Test	Level 8	Level 6	Level 4
TOEIC	790	585	190
Listening	430	320	115
Reading	360	265	75
Test of Spoken English	50	40	25
Test of Written English	4.5	3.0	1.5

Statistical Distinctiveness Between Cut Scores

The standard-setting process established three expert-judgment-based cut scores on each test, providing the boundaries between basic, moderate, and high proficiency levels. One question that could be asked, however, is whether these cut scores are statistically distinct from one another? This question was addressed by examining the conditional standard error of measurement (CSEM) around each cut score.

Rather than assume that a test is equally reliable at all points along the scale, the conditional standard error of measurement is calculated for every point along the raw score scale (Lord, 1984) and then transformed to the scaled score scale. Table 11 provides the conditional standard error of measurement (CSEM) on the scaled score scale for each cut score on the different tests. If a candidate’s true score (that is, the score he or she would obtain on a perfectly reliable test) is at one of the cut scores, there is approximately a 0.95 probability that he or she will earn a score within with two CSEMs of his or her true score, and a 0.99 probability of earning a score within three CSEMs of his or her true score. Considering a candidate whose true

score is at each of the established cut scores, allows us to estimate the probability that the candidate would be misclassified at a different CLB level.

Table 11: Conditional Standard Error Of Measurement At Each Cut Score

Test	Level 4	Level 6	Level 8
TOEIC – Listening	27	30	26
TOEIC – Reading	24	26	23
Test of Spoken English	- ⁷	1.95	1.70
Test of Written English	0.31	0.50	0.47

The number of CSEMs each cut score was from one another was calculated. These values are presented in Table 12. For example, the number of CSEMs the Level 6 TOEIC Listening cut score of 320 is from the Level 4 and Level 8 cut scores is 6.8 and 3.7, respectively. This means that a candidate with a true score of 320 would need to earn a score that is 6.8 CSEMs below that score to be misclassified as a Level 4 candidate, and would need to earn a score that is 3.7 CSEMS above that score to be misclassified as a Level 8 candidate. The likelihood of either instance occurring is negligible.

Table 12: Distance Between Cut Scores In CSEMs

	Listening	Reading	TSE	TWE
Level 4 cut score (1 CSEM)	115 (27)	75 (24)	25 (-)	1.5 (0.31)
# of CSEMs to Level 6 cut score	7.6	7.9		4.8
# of CSEMs to Level 8 cut score	11.7	11.9		9.8
Level 6 cut score (1 CSEM)	320 (30)	265 (26)	40 (1.95)	3.0 (0.50)
# of CSEMs to Level 4 cut score	6.8	7.3	7.7	3.0
# of CSEMs to Level 8 cut score	3.7	3.7	5.1	3.0
Level 8 cut score (1 CSEM)	430 (26)	360 (23)	50 (1.70)	4.5 (0.47)
# of CSEMs to Level 6 cut score	4.2	4.1	5.9	3.2
# of CSEMs to Level 4 cut score	12.1	12.4	14.7	6.4

⁷ No candidates scored as low as the Level 4 cut-score on the TSE. Thus the CSEM could not be calculated.

Since there are three or more CSEMs between a candidate whose true score was located exactly at one of the cut scores, and the next cut score above or below the one under consideration, the probability of that candidate being misclassified approximates zero. The expert-judgment, standard-setting process gave meaning to the distinctiveness of each cut score. This analysis illustrates that the cut scores are also statistically distinct.

References

- Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.
- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Cizek, G.J. (1993). *Reactions to National Academy of Education report: Setting performance standards for student achievement*. Washington, DC: National Assessment Governing Board.
- Faggen, J. (1994). *Setting standards for constructed response tests: An overview* (ETS RM 94-19). Princeton, NJ: Educational Testing Service.
- Hambleton, R.K., Jaeger, R.M., Plake, B.S., & Mills, C. (2000). Setting performance standards on complex educational assessments. *Applied Psychological Measurement*, 24, 355-366.
- Hurtz, G.M., & Auerbach, M.A. (2003). A meta-analysis of the effects of modifications to the Angoff method on cutoff scores and judgment consensus. *Educational and Psychological Measurement*, 63, 584-601.
- Mehrens, W.A. (1995). Methodological issues in standard setting for educational exams. In *Proceedings of Joint Conference on Standard Setting for Large-Scale Assessments* (pp.221-263). Washington, DC: National Assessment Governing Board and National Center for Education Statistics.
- http://www.language.ca/pdfs/clb_adults.pdf Canadian Language Benchmarks 2000: English as a second language – for adults

AGENDA: Mapping TOEIC, TWE and TSE onto the Canadian Language Benchmarks***February 17th to 19th, 2004 - Toronto, Canada*****Day 1**

8:00 – 8:30	Breakfast
8:30 – 9:00	Introductions
9:00 – 9:30	Overview of ETS language tests, the CLB and the purpose of the study
9:30 – 10:00	Standard-setting training: selected-response items
10:00 – 10:45	Define candidate focal groups for Levels 4, 6 and 8 on the CLB (Listening)
10:45 – 11:00	Break
11:00 – 12:00	Standard-setting training: Practice judgments (Listening items)
12:00 – 1:00	Lunch
1:00 – 3:00	Standard-setting judgments on Listening items
3:00 – 3:15	Break
3:15 – 4:00	Define candidate focal groups for 4, 6 and 8 on the CLB (Reading)
4:00 – 4:45	Standard-setting training: Practice judgments (Reading items)
4:45 – 5:00	Wrap up for the day and adjourn

Day 2

8:00 – 8:30	Breakfast
8:30 – 9:30	Discussion of first round of Listening judgments, final judgments
9:30 – 9:45	Review of Level 4, 6 and 8 Reading focal group descriptions
9:45 – 12:00	Standard-setting judgments on Reading items (take break as needed)
12:00 – 1:00	Lunch
1:00 – 1:30	Standard-setting training: constructed response items
1:30 – 2:15	Define candidate focal groups for Levels 4, 6 and 8 on the CLB (Speaking)
2:15 – 3:45	Training and standard-setting judgments on TSE
3:45 – 4:00	Break
4:00 – 4:45	Review and discussion of TSE cut scores, final judgments
4:45 – 5:00	Wrap up for the day and adjourn

Day 3

8:00 – 8:30	Breakfast
8:30 – 9:30	Discussion of first round of Listening judgments, final judgments
9:30 – 10:30	Discussion of first round of Speaking judgments, final judgments
10:30 – 10:45	Break
10:45 – 11:45	Define candidate focal groups for Levels 4, 6 and 8 on the CLB (Writing)
11:45 – 12:45	Lunch
12:45 – 2:00	Training and standard-setting judgments on Writing component
2:00 – 2:15	Break
2:15 – 3:15	Discussion and final judgments
3:15 – 3:30	Wrap up and adjourn

**STUDY TO MAP THE TEST OF ENGLISH FOR INTERNATIONAL COMMUNICATION, THE TEST OF
SPOKEN ENGLISH, AND THE TEST OF WRITTEN ENGLISH ONTO THE
CANADIAN LANGUAGE BENCHMARKS**

In order to facilitate the use of three English Language tests (TOEIC, TSE and TWE) for immigration purposes, ETS is conducting a study to map the three tests onto the Canadian Language Benchmarks (CLBs). The study goals are (a) to document the alignment between the content of each test and the content of the Benchmarks and (b) to identify test scores that correspond to Benchmark proficiency levels, using an expert judgment standard-setting approach. At the study you will be familiarized with the tests, receive training in the standard-setting process, and have an opportunity to practice making judgments.

During the study itself, the discussions will focus around proficiency levels 4, 6, and 8 of the Canadian Language Benchmarks. In order to facilitate discussions, it is very important that you become familiar with the CLBs in general and these three levels in particular. A PDF version of the CLB can be found at the following address: http://www.language.ca/pdfs/clb_adults.pdf

The ETS tests that we will be benchmarking at the study address the four modalities of Speaking, Listening, Reading and Writing, and we will be discussing the characteristics of a Level 4, 6, and 8 candidate by language modality. In the section below, relevant tables from the CLB have been identified by page number and title. Please review these CLB tables, paying close attention to levels 4, 6, and 8. Highlight key words or phrases that you believe best represent (most clearly define) each of the levels.

Speaking: Global Performance Descriptors – Page 4 and 54

Writing: Global Performance Descriptors – Page 40 and 98

Listening: Global Performance Descriptors – Page 16 and 74

Reading: Global Performance Descriptors – Page 28 and 86

Having reviewed the relevant CLB tables, complete the attached sheet by briefly noting in your own words, in the space provided, the key characteristics or indicators from the CLB tables that describe an English Language learner who:

1. you believe is at the 4-level of proficiency
2. you believe is at the 6-level of proficiency
3. you believe is at the 8-level of proficiency

For example, considering first the tables that defining proficiencies related to Speaking, review each of the CLB tables listed above, and identify critical descriptors in the tables that help you distinguish between levels 4, 6 and 8 of proficiency. For example, you might note among other things that a level 4 learner has “control of basic grammar; uses correct past tense with common verbs” while a level 6 learner can use a “variety of structures with some omission/reduction of elements (e.g., past tense)” and finally a level 8 learner can use “a variety of sentence structures and an expanded inventory of concrete, idiomatic and conceptual language.”

Your notes, along with those of your colleagues, will form the starting point for discussion during the study itself.

Key Characteristics by Language Modality of a Level 4, 6, and 8 English Language Learner

	<i>Level 4</i>	<i>Level 6</i>	<i>Level 8</i>
<i>Speaking</i>			
<i>Writing</i>			
<i>Reading</i>			
<i>Listening</i>			

Please bring this completed sheet with you to the meeting February 17th. Thank you

Canadian Language Benchmarks – Listening

Compare CLB Level 6 with 4

- A six can probably handle listening activities that are purely audio (no visual component) but a four would “crash and burn” in that instance.
- A six doesn’t require simplified instructions but a four does.
- A six can listen, summarize and recall details, but a four can only recall details that are familiar and relevant.
- A six can understand a wider range of content, but a four recognizes content based on familiar keywords/vocabulary.
- A six can recognize a typical script, but a four would recognize the topics of the script (relies on “expected” responses, etc.)
- A six can comprehend idiomatic language/common expressions but a four is “out to lunch” with them.
- A four requires simplified/repetitive/clarification to be successful on a task, but a six generally doesn’t, however, if he/she does he/she will ask for it.

Compare CLB Level 8 with 6

Benchmark 6

- Content – familiar, relevant topics mostly concrete and factual language.
- Speed - very familiar and relevant topics – moderate to normal rate
 - Less familiar, but still relevant – slower rate, may need repetitions
- Comprehension – main ideas, gist, important details.

Benchmark 8

- Content – abstract concepts, understands some idioms.
- Speed - topics of general popular interest, normal rate of speech.
- Comprehension – main ideas and subtleties
 - Interpretive listening
 - Can make intended inferences

Canadian Language Benchmarks – Reading

Compare CLB Level 6 with 4

Benchmark 4

- Short, simplified texts on familiar predictable topics
- Low level of inference (slow to moderate reading rate)
- More comfortable with bilingual dictionary
- Locate specific details to compare and contrast facts
- Gets main idea, key points

Benchmark 6

- Short, plain language, authentic texts (3-5 paragraphs)
- Moderate level, most concrete and factual
- Strategies for determining new words in context comfortable with unilingual learners dictionary
- Comfortable in a familiar context – beginning to make inferences.

Compare CLB Level 8 with 6

- 8 can glean sub-textual items, but a 6 is more surface
- 8 understands most details, but 6 needs more concretely
- 8 can scan and infer from context unfamiliar vocabulary
- 6 needs more dictionary support
- 8 reads more “authentic” text, but 6 reads more “plain English” text
- 8 can handle longer texts
- 8 can read academic/business as well as text
- 8 is at ease with English text, but 6 relies more on strategies to deal with text
- Eight’s are more critical
- Six’s are more literal than interpretive
- Eight can more likely handle extended metaphor

Canadian Language Benchmarks – Speaking

Compare CLB Level 6 with 4

Benchmark 4

- Common vocabulary, basic grammar (rely on generic vocabulary), everyday topics, routine questions, face-to-face communication
- Control of basic structures and tenses
- Pronunciation errors may impede communication-processing time
- Short, predictable telephone exchange-processing time
- Relate a short story
- Visual clues
- False cognates
- Connected discourse but not fluently
- Limited ability to repair, but with frequent self-correction (rephrasing)

Benchmark 6

- Has more confidence, more phrases, structures, idioms
- Phone exchange stressful with strangers
- Confusing pronouns, dropping of articles
- Can have small group discussions, give detailed facts and ideas
- Conversation management, active listening
 - Asking for clarification, repetitive explanation, speech reasonably fluent, however; pauses occur at times

Compare CLB Level 8 with 6

- 6 can speak on familiar topics, using a variety of structures, however; errors are frequent in complex structures and discourse is reasonably fluent, but 8 can speak on familiar and unfamiliar topics with familiar discourse, is fluent but unfamiliar and more abstract fluency breaks down.
- 6 grammar and pronunciation can impede communication, 8 don't or rarely do.
- 6 express obligation/ability/certainly/advise/ warning
- 8 can synthesize
- 8's are highly cohesive, but 6's are more fragmented
- 6's require more processing time, where 8 doesn't require as much.
- 6 is sometimes incomprehensible but 8 rarely is.
- 8 has longer utterances, more connected discourse, more natural (fewer pauses/hesitations)
- 6 hesitates more often
- 8 understands/uses colloquialisms/vernacular, where as 6 struggles but uses learned idiomatic language.

Canadian Language Benchmarks – Writing

Compare CLB Level 6 with 4

Benchmark 4

- Simple structures
- Simple connectors
- Control of basic conventions (capitalization/punctuation)
- Avoids complex structures
- Write simple descriptions on familiar topics of relevance
- Narrate simple events on familiar topics of relevance
- Can complete simple form on familiar topics of relevance
- Write simple direction on familiar topics of relevance
- Writes in simple paragraph form on familiar topics of relevance
- Uses common vocabulary
- Spelling can be problematic

Benchmark 6

- Variety of structure
- Variety of connectors
- More advanced conjunctions
- Can handle moderately complex tasks (“business” letter/resume)
- Attempts complex structures, but errors are frequent
- More detailed descriptors
- Direct translation from L1 when trying to explain more complex ideas
- Initial sense of audience
- Some creativity is evident
- Some sense of formal/independent register
- Use of expanded vocabulary (more technical, etc.)
- 1-2 paragraphs, relatively good control of simple structure
 - attempts complex structure
 - uses some connectors
 - concrete topics
- Misspells some words, word forms
- Awkward phrasing
- Have trouble with more formal language

Compare CLB Level 8 with 6

Benchmark 8

- 3-4 paragraphs, complex sentence structure, generally correct
- Better organized, lexical phrases
- Better sense of audience, purpose
- Abstract topics, use of idiomatic language
- Occasional spelling errors
- Writes clear instructions, can complete a complex form
- Can do summaries and outlines
- Able to write reports

Round 1 Judgments

Item	Addresses part of the CLB	Circle the score that a <u>Level 8</u> candidate would achieve on each item	Circle the score that a <u>Level 6</u> candidate would achieve on each item	Circle the score that a <u>Level 4</u> candidate would achieve on each item
1	Y N	20 25 30 35 40 45 50 55 60	20 25 30 35 40 45 50 55 60	20 25 30 35 40 45 50 55 60
2	Y N	20 25 30 35 40 45 50 55 60	20 25 30 35 40 45 50 55 60	20 25 30 35 40 45 50 55 60
3	Y N	20 25 30 35 40 45 50 55 60	20 25 30 35 40 45 50 55 60	20 25 30 35 40 45 50 55 60
4	Y N	20 25 30 35 40 45 50 55 60	20 25 30 35 40 45 50 55 60	20 25 30 35 40 45 50 55 60
5	Y N	20 25 30 35 40 45 50 55 60	20 25 30 35 40 45 50 55 60	20 25 30 35 40 45 50 55 60
6	Y N	20 25 30 35 40 45 50 55 60	20 25 30 35 40 45 50 55 60	20 25 30 35 40 45 50 55 60
7	Y N	20 25 30 35 40 45 50 55 60	20 25 30 35 40 45 50 55 60	20 25 30 35 40 45 50 55 60
8	Y N	20 25 30 35 40 45 50 55 60	20 25 30 35 40 45 50 55 60	20 25 30 35 40 45 50 55 60
9	Y N	20 25 30 35 40 45 50 55 60	20 25 30 35 40 45 50 55 60	20 25 30 35 40 45 50 55 60

Do Not Write in this Space.

	End of round 1, <u>Level 8</u> cut score	End of round 1, <u>Level 6</u> cut score	End of round 1, <u>Level 4</u> cut score
My initial recommended cut score (range 20 – 60)			
Group average			

Round 2 Judgments

	Write the overall score that a <u>Level 8</u> Candidate would achieve on this test	Write the overall score that a <u>Level 6</u> Candidate would achieve on this test	Write the overall score that a <u>Level 4</u> Candidate would achieve on this test
My final recommended cut score (range 20 – 60)			

(Signature)

INSERT SCAN FORM

Scan form page 2

Scan form page 3

Standard-Setting Participants⁸

Name	Affiliation
Christa Bruns	Southern Alberta Institute of Technology - CLB assessor and instructor of ESL in the workplace
Marry Ann Harrington	University of New Brunswick - Testing manager and CLB assessor
Ruth Hungerland	Memorial University - ESL program coordinator
Annette Kreider	Language Assessment, Referral and Counseling Centre - Team leader and language assessor
Deidre Lake	Canadian Language Research and Consulting - Researcher and former LINC program manager
Daniel L. Love	Calgary Immigrant Aid Society: ILVARC - CLB assessor and teacher
Ken Maneker	Forward Consulting - Former language school director of education
Ron Mang	Saskatchewan Institute of Applied Science and Technology - CLB assessor
Alison P. Norman	Vancouver Community College - Test developer and program administrator for ELSA
Bruce Russel	LCRT Consulting - ESL consultant to Toronto District School Board
Christine Scardicchio	Dufferin-Peel Catholic District School Board - CLB assessor for adult ESL and co-op workplace programs
Andrea Strachan	LCRT Consulting - Curriculum developer and certified CLB assessor
Reena Taviss	Education and Training Consultant - Consultant with 25 years of experience in evaluation and program development
David L. E. Watt	University of Calgary - Researcher on language assessments

⁸ One panelist asked not to have their name included in the final report.

ETS Researchers' Bios

Dr. Richard Tannenbaum is the Executive Director of Research and Technical Analysis for the Higher Education Division at ETS. He is responsible for the development, management, and coordination of research initiatives in support of the lines of business within the Division (English Language Learning; Graduate and Professional Programs, College and University Services; and Teacher Quality). Prior to his current role, Dr. Tannenbaum was the Director of standard setting studies for the Praxis Series™ of teacher licensure assessments. Richard's areas of expertise include licensure and certification, standard setting, job analysis, validation, assessment and criterion development, and survey-based research.

Dr Caroline Wylie is a researcher at ETS. Her work has been mainly focused in the area of complex performance assessments. She also provides technical measurement expertise and research support for assessments and professional products developed at ETS; included among these responsibilities are standard setting studies, validation studies, and job analysis studies.

Table A: Judgments for the TOEIC Listening Section

	Round 1 Judgments			Round 2 (final) Judgments		
	Level 8	Level 6	Level 4	Level 8	Level 6	Level 4
P1	73	57	41	75	58	33
P2	73	56	34	73	56	30
P3	69	58	23	75	60	30
P4	70	39	19	70	45	25
P5	79	58	27	79	58	28
P6	84	74	17	80	65	20
P7	78	61	27	77	60	27
P8	67	50	31	70	50	25
P9	79	62	27	79	60	27
P10	82	52	14	80	52	14
P11	66	47	30	66	55	25
P12	82	66	40	82	66	35
P13	87	72	50	70	60	35
P14	86	72	49	85	70	40
P15	87	77	39	85	72	35
Mean (truncated)	77	60	31	76	59	29
Median (truncated)	79	58	30	77	60	28
Standard Deviation	7.3	10.8	11.2	5.8	7.2	6.6
Minimum	66	39	14	66	45	14
Maximum	87	77	50	85	72	40

Table B: Judgments for the TOEIC Reading Section

	Round 1 Judgments			Round 2 (final) Judgments		
	Level 8	Level 6	Level 4	Level 8	Level 6	Level 4
P1	75	59	40	75	60	35
P2	74	50	24	74	50	24
P3	81	68	23	80	64	25
P4	65	32	14	65	40	19
P5	75	56	28	75	56	28
P6	76	58	10	75	55	10
P7	68	49	20	70	50	20
P8	72	58	31	75	58	30
P9	79	62	24	79	60	25
P10	88	79	31	85	75	30
P11	63	43	27	65	45	25
P12	84	69	38	84	69	30
P13	88	76	52	70	60	30
P14	85	70	32	85	70	32
P15	85	71	28	85	71	28
Mean (truncated)	77	60	28	76	59	26
Median (truncated)	76	59	28	75	60	28
Standard Deviation	7.9	12.8	10.2	6.8	10.0	6.2
Minimum	63	32	10	65	40	10
Maximum	88	79	52	85	75	35

Judgments for Test of Spoken English (TSE)

	Round 1 Judgments			Round 2 (final) Judgments		
	Level 8	Level 6	Level 4	Level 8	Level 6	Level 4
P1	54	44	34	50	40	30
P2	48	37	28	45	35	25
P3	52	41	36	45	35	30
P4	51	41	33	45	35	25
P5	51	43	37	50	45	35
P6	51	41	23	50	40	20
P7	51	41	31	45	35	25
P8	51	41	30	50	35	25
P9	47	38	27	45	35	25
P10	51	40	28	50	40	30
P11	52	42	34	45	35	25
P12	52	41	29	50	40	30
P13	60	42	29	50	40	30
P14	55	44	36	55	45	35
P15	54	43	23	50	40	20
Mean (truncated)	52	41	31	48	38	27
Median (truncated)	51	41	30	50	40	25
Standard Deviation	3.1	2.0	4.4	3.1	3.6	4.6
Minimum	47	37	23	45	35	20
Maximum	60	44	37	55	45	35

Judgments for Test of Written English (TWE)

	Round 1 Judgments			Round 2 (final) Judgments		
	Level 8	Level 6	Level 4	Level 8	Level 6	Level 4
P1	4.5	3.5	2.0	4.5	3.0	2.0
P2	5.0	3.5	2.0	5.0	3.5	1.5
P3	5.0	4.0	1.5	4.0	3.5	1.5
P4	3.5	2.0	1.0	3.5	2.5	1.0
P5	5.0	3.5	2.0	5.0	3.5	2.0
P6	5.0	3.5	0.5	4.5	3.0	0.5
P7	4.0	2.5	0.5	4.0	2.5	0.5
P8	4.0	3.0	1.0	4.0	3.0	1.0
P9	4.0	3.0	1.5	4.5	3.0	1.5
P10	5.5	4.0	1.0	5.0	3.5	0.5
P11	6.0	4.0	1.0	5.0	4.0	1.0
P12	4.5	2.5	0.0	4.5	3.0	1.5
P13	5.5	4.0	2.5	5.0	3.5	2.0
P14	5.0	3.5	2.0	5.0	3.5	2.0
P15	4.5	3.5	1.0	4.5	3.5	1.0
Mean (truncated)	4.7	3.3	1.3	4.5	3.2	1.3
Median (truncated)	5.0	3.5	1.0	4.5	3.5	1.5
Standard Deviation	0.7	0.6	0.7	0.5	0.4	0.6
Minimum	3.5	2.0	0.0	3.5	2.5	0.5
Maximum	6.0	4.0	2.5	5.0	4.0	2.0