Pearson | PTE

Full Research Report

Establishing alignment between PTE Core and the Canadian Language Benchmarks (CLB)

Contents

1. Purpose

The purpose of this study is to establish an alignment between PTE Core test scores and the Canadian Language Benchmarks (CLB) to support the use of PTE Core scores as evidence of English language proficiency for applications with Immigration, Refugees, and Citizenship Canada (IRCC). The outcomes of this research identify threshold PTE Core scores for CLB 4 to 10 for each of the four skills: listening, reading, speaking, and writing.

2. Background

PTE Core is a computer-based English language test that assesses listening, reading, speaking, and writing skills relevant to general migration contexts, such as working visas and citizenship. The main use of PTE Core test scores is Canadian migration.

Immigration, Refugees, and Citizenship Canada (IRCC) requires applicants for specific visa classes to demonstrate a specified level English language proficiency as described by the Canadian Language Benchmarks (CLB). The CLB is the official standard of reference for English language proficiency in Canada. The CLB describes 12 distinct levels of performance along the spectrum of basic, intermediate, and advanced ability (CIC Canada, 2012a).

PTE Core was developed with the Canadian migration context in mind to address CLB levels 4 through 10. The development process included iterative phases of research and development engaging with Canadian English language teachers, academics, stakeholders, and test takers. An initial Test Review Group, which included an independent expert panel of English language academics and teachers, reviewed the design intentions and assessed construct for each item type and mapped these demands onto the competencies and levels of proficiency articulated in the CLB to define the range of skills and abilities covered by PTE Core item types.

This was followed by an intensive mapping activity, in which English language experts mapped each item in the entirety of the item bank to specific CLB skills and levels. These activities informed the development of the test blueprint for PTE Core, ensuring that the appropriate balance of item types and difficulties were represented in the test form structure. Sample test forms were piloted with test takers in focus groups and interviews to gather feedback on the accessibility, difficulty, and relevance of PTE Core to their Canadian migration experience.

A large-scale field test was conducted with a representative population of Canada-bound test takers to calibrate the item bank, evaluate the measurement properties of the test, and gather test taker performances to be used in standard setting activities. The standards setting and verification activities described in this paper build on these earlier phases of research to establish a strong relationship between PTE Core and the CLB.

3. Overview of study

This paper describes the standard setting activities that underpin the alignment of PTE Core to the CLB. This alignment was established through standard setting activities in 2021, and a verification activity was conducted in 2022 to validate the findings of the initial standard setting. PTE Core launched in 2024, and the alignment with the CLB will continue to be monitored and validated throughout its lifecycle.

PTE Core test scores were related to the CLB using established standard setting methodologies that have previously been used to successfully align high stakes test scores to external performance standards of language proficiency (Fox & Fraser, 2014; Jones et al., 2017; Tannenbaum & Wylie, 2004). In particular, we followed procedures, analysis, and reporting structure similar to those used to establish alignment between PTE Academic and the CLB in order to maintain consistency within the PTE suite of assessments (Jones et al., 2017).

Four separate standard setting activities were conducted, one for each skill. For the receptive skills of listening and reading, a modified Angoff method was used to judge the difficulty of test items in relation to the CLB (Angoff, 1971; Hambleton & Pitoniak, 2006). The judged item difficulties were related to the empirical item difficulties established during field testing to produce the alignments for listening and reading. For the productive skills of speaking and writing, a modified Analytic Judgement method was used to judge the language proficiency demonstrated by test takers over multiple responses in relation to the CLB (Hambleton & Pitoniak, 2006; Plake & Hambleton, 2001). The judgements of test taker proficiency were related to the empirical test taker abilities established during field testing to produce alignments for speaking and writing.

This paper describes the first standard setting activities for the newly developed PTE Core test and outlines plan for the continued validation and monitoring of the relationship between PTE Core test scores and the CLB.

4. Panel of judges

A panel of 14 judges was recruited through existing contacts between Pearson and relevant institutions and researchers in Canada. Judges were required to have extensive experience in teaching and assessing English language, and to have familiarity with the CLB.

Table 1- Characteristics of panel of judges

Panel Characteristic	Category	N	%
Gender	Female	12	86
	Male	1	7
	Non-binary	-	-
	Prefer not to say	1	7
Job Role	Professor	4	29
	Graduate student	4	29
	Instructor	3	21
	Independent specialist	1	7
	Researcher	1	7
	Director / Head of department	1	7
Years of Experience	Under 5	2	14
	5 to 9	2	14
	10 to 14	2	14
	More than 15	8	57
Canadian Province	Alberta	3	21
	British Columbia	1	7
	Manitoba	-	-
	New Brunswick	-	-
	Newfoundland and Labrador	-	-
	Nova Scotia	1	7
	Ontario	9	64
	Prince Edward Island	-	-
	Quebec	-	-
	Saskatchewan	-	-
Affiliation	University	10	71
	Language Instruction for Newcomers to Canada (LINC)	2	14
	College	1	7
Familiarity with CLB	Very familiar	6	43
	Moderately familiar	6	43
	Slightly familiar	2	14
	Not familiar	-	-

The majority of the panel was female and had more than 15 years of experience in language learning and assessment in Canada. Most participants were affiliated with a university or Language Instruction for Newcomers to Canada (LINC) programme and were based in Ontario, Alberta, British Columbia, or Nova Scotia. All participants were familiar with the CLB, with most participants being moderately to very familiar.

5. Method

5.1 Standard setting methods

A modified Angoff method was selected for receptive skill standard setting activities. The Angoff method is an item-centered approach to standard setting. Judges typically review individual test items and determine the probability that a borderline test taker could answer the item successfully *(Angoff, 1971)*. Because we require cut scores for multiple levels of the CLB, the Angoff procedure was modified to avoid the repetition of considering several different borderline test takers. Rather than asking judges to determine the probability that a test taker of each CLB level would succeed on an item, we asked judges to identify which CLB level a test taker would need to be in order for the probability of success to be 50%. Prior research has indicated that this question can be difficult for judges to conceptualize, so a large portion of the training was dedicated to discussing this concept until judges developed an intuitive understanding of the meaning and application *(Jones et al., 2017)*.

A modified Analytic Judgement method was selected for productive skills. The Analytic Judgement method is a test-taker centered, holistic approach to standard setting. Judges typically review a sample of full exam papers and classify each test taker's holistic performance into categories of basic, proficient, and advanced, often with subcategories introduced to refine classification *(Plake & Hambleton, 2001)*. This method aligns with the goals of this standard setting study for productive skills, as we aimed to classify test taker speaking and writing performance on the CLB, which provides three stages of basic, intermediate, and advanced proficiency, each with four subcategories of benchmarks. Some modifications were made to the Analytic Judgement method for this study. Rather than reviewing entire test performances, judges reviewed speaking and writing performances separately. Judges focused on item types where test takers were required to produce longer responses that enabled judges to form a holistic impression of each test taker's speaking and writing ability.

5.2 Materials

The standard setting activities used a mixture of test items and test taker responses. The booklets for Listening and Reading activities each contained 20 items from 4 different item types. The quantity of each item type reflected the proportion of each item type in the full test, and aligned with the sampling strategy used in prior PTE alignment research related to the CLB *(Jones et al., 2017)*. The samples for the Speaking and Writing activities each contained 4 responses from 20 test takers.

The samples were selected to represent a cross section of item difficulty and test taker ability within each skill area, as determined through field testing. The items and responses were presented in random order to encourage judges to make independent judgements about each item or response. Tables 2.1 and 2.2 show the number and type of items used for each standard setting activity.

Table 2.1 – Composition of standard setting materials for receptive skills

Activity	Item type	Description	Number of items
Listening	209-LL-SAMC	Single Answer Multiple Choice Listening	5
	210-LL-MAMC	Multiple Answer Multiple Choice Listening	5
	211-LL-GAPS	Gap Fill Listening	5
	212-LR-HOTS	Highlighting Incorrect Words	5
Reading	201-RR-SAMC	Single Answer Multiple Choice Reading	4
	202-RR-MAMC	Multiple Answer Multiple Choice Reading	4
	205-RR-GAPS	Gap Fill Reading	6
	218-RW-GAPS	Gap Fill Reading and Writing	6

Table 2.2 – Composition of standard setting materials for productive skills

Activity	Item type	Description	Number of Items per Test Taker
Speaking	219-SS-DESC	Describe Image	2
	223-SS-SITU	Speaking Situations	2
Writing	215-LW-SUMM	Listening Summary Writing	1
	222-WW-EMAI	Email Writing	2
	208-RW-SUMM	Reading Summary Writing	1

5.3 Training and standardization

Due to the Covid-19 pandemic, it was not possible for training and standard setting activities to take place with judges in the same physical location. All standard setting activities were adapted to take place virtually. Activities were restructured into a combination of independent work and virtual meetings in Microsoft Teams.

Training took place virtually on 18 June 2021. The training session included activities organized around four aims:

- To familiarize judges with the standard setting procedures
- To familiarize judges with the PTE Core test purpose and structure, with particular emphasis on the item types used in the standard setting activities
- To standardize judges' understanding of CLB proficiency levels
- To provide judges practice in applying CLB descriptors to exemplar items and performances.

Following the training meeting, judges were instructed to independently review the CLB profiles of ability for each skill and to identify points of differentiation between the levels (CIC Canada, 2012a). They were instructed to consider the CLB profiles of ability alongside the exemplar tasks and performances provided in the CLB Support Kit to prepare for the standard setting activities *(CIC Canada, 2012b)*.

5.4 Standard setting activity structure

The standard setting activities took place virtually between 21 and 27 June 2021. Four separate activities were run, one for each skill.

Each activity began with either a discussion to re-familiarise judges with the standard setting task, the relevant CLB descriptors, and important points of differentiation across the CLB levels. For each activity, judges completed two rounds of independent ratings. Judges used randomly assigned ID numbers when submitting their ratings to ensure anonymity throughout the activities.

In Round 1, judges worked independently to rate either test items or test taker abilities. For receptive skills, ratings were based on the question: "At which CLB level would a learner have a 50% chance of answering the item correctly?" For productive skills, ratings were based on the question: "What level of CLB proficiency best describes this speaker/writer?"

Judges submitted their Round 1 rating forms so that agreement statistics could be calculated. The agreement statistics were presented back to the group, along with empirical data from field testing, to support the group's discussion of their Round 1 ratings.

Following the discussion, judges completed their Round 2 ratings independently. Judges were instructed to review their ratings from Round 1 in light of the discussion and data presented. They were instructed to not amend their ratings in any way if their opinion had not changed. Judges submitted their updated ratings in their Round 2 forms.

The Round 2 ratings were then aggregated, judge agreement statistics were recalculated, and the results were presented back to the judges for summary discussion.

6. Analysis

6.1 Data cleaning and judge agreement

Following Round 2, the data were prepared for analysis. Drawing on prior PTE and CLB alignment research, three criteria were applied to clean the data *(Jones et al., 2017).*

- Remove any **ratings** more than 1.5 levels from the average rating for an item. A total of 5.54% of ratings were removed.
- Remove any **judges** whose correlation with the average of all judges is less than 0.5. No judges had correlations below 0.50, so no judges were removed.
- Remove any **items** with low certainty. Items are considered to have low certainty when the maximum proportion of judges rating in any two adjacent categories is less than 70%. A total of 3.75% of items were removed.

Table 3 describes the level of agreement between judges at the end of each round, as well as the final level of agreement in the sample after data cleaning procedures. The mean rating, standard deviation (SD) of ratings, and standard error of judgement (SEJ) were calculated for each item. Table 3 shows the mean rating, mean standard deviation, and mean standard error of judgement across all items judged in each round of the four activities. Table 3 also includes the number of items considered to be low certainty at each stage.

Table 3 – Judge agreement for each round and after cleaning

Skill	Round	Mean Rating	Mean SD	Mean SEJ	Low certainty (n)
Listening	Round 1	5.83	1.32	0.35	11
	Round 2	5.66	0.88	0.23	6
	Final	5.60	0.72	0.20	0
Reading	Round 1	5.80	1.18	0.32	11
	Round 2	5.70	0.78	0.21	1
	Final	5.63	0.61	0.17	0
Speaking	Round 1	5.45	1.10	0.30	7
	Round 2	5.35	0.78	0.21	2
	Final	5.34	0.59	0.16	0
Writing	Round 1	4.96	1.11	0.30	9
	Round 2	4.90	0.69	0.18	0
	Clean	4.84	0.58	0.16	0

After both rounds of rating and data cleaning, a good level of consensus is shown in the data, with the mean standard errors of judgement all less than a quarter of a CLB level.

6.2 Calculating alignment

PTE Core reports scores on a scale of 10 to 90 for the overall score, as well as the skill scores for listening, reading, speaking, and writing. The scores are linear transformations of an underlying ability scale derived from IRT calibration of field test data.

To produce the alignment between PTE Core test scores and the CLB, linear regression was used to relate the standard setting ratings to field test data. For receptive skills, PTE Core item difficulty values obtained in field testing were regressed against the CLB ratings obtained through standard setting. For productive skills, test takers' PTE Core skill scores from field testing were regressed against the holistic judgement of their CLB ability determined through standard setting.

All four skills show strong correlations between the field test data and the standard setting data. The proportion of explained variance is consistent across all four skills, from 0.537 to 0.652. This relationship will be monitored over time, as a diverse range of test takers is exposed to the PTE Core test in a high stakes testing environment.

.73 to .81

All four skills show strong correlations between field test data and CLB ratings

Table 4 – Regression functions by skill

Skill	Regression function	Correlation (r)	Explained variance (r^2)
Listening	PTE Core = 10.6 x CLB – 14.2 **	0.733	0.537
Reading	PTE Core = 9.1 x CLB – 3.3 ***	0.807	0.652
Speaking	PTE Core = 8.4 x CLB + 9.0 ***	0.784	0.615
Writing	PTE Core = 9.3 x CLB + 3.9 ***	0.750	0.562

*** significant at p=0.001; ** significant at p=0.01

6.3 Alignment table

Table 5 shows the PTE Core scores calculated for each CLB level using the regression functions in Table 4.

Table 5 – Alignment of PTE Core scores to CLB levels

CLB	PTE Core Listening	PTE Core Reading	PTE Core Speaking	PTE Core Writing
10	89–90	88–90	24–32	90
9	82–88	78–87	84–88	88–89
8	71–81	69–77	76–83	79–87
7	60–70	60–68	68–75	69–78
6	50–59	51–59	59–67	60–68
5	39–49	42–50	51–58	51–59
4	28–38	33–41	42–50	41–50
3	18–27	24–32	34–41	32–40

6.4 Quality of alignment

6.4.1 Decision consistency and decision accuracy

Decision consistency refers to the extent to which test takers are classified into the same categories over repeated administrations of the same test, where these classifications are used to make high stakes decisions about the test taker. Decision accuracy refers to the extent to which classifications based on observed test scores agree with true classifications *(Lee, 2010)*.

Different methods exist to estimate the decision consistency and accuracy from a single administration of a test, such as the field test for PTE Core. This analysis uses the method described by Rudner (2000, 2005) and implemented in the R package cacIRT to estimate decision consistency and accuracy for each PTE Core score associated with each CLB level in Table 5 *(Lathrop, 2015)*.

Using this method, decision consistency and accuracy metrics are estimated by creating a normal distribution of ability for each test taker using their IRT ability estimate as the mean and the associated standard error as the standard deviation. For each test taker, decision accuracy is the proportion of the ability distribution plotted between the cut scores for their classification. Decision consistency is the probability of being classified the same way (either inside the cut score range or outside of the cut score range) on two independent tests. Table 6 shows the accuracy and consistency statistics for each cut score identified in Table 5.

Table 6 – Decision Consistency and Decision Accuracy for CLB thresholds

CLB	PTE Core Listening	PTE Core Reading	PTE Core Speaking	PTE Core Writing
	DC/DA			
10	0.94 / 0.92	0.94 / 0.92	0.93 / 0.90	0.95 / 0.93
9	0.92 / 0.89	0.91 / 0.87	0.92 / 0.88	0.94 / 0.92
8	0.90 / 0.85	0.90 / 0.86	0.89 / 0.85	0.91 / 0.88
7	0.90 / 0.86	0.91 / 0.88	0.90 / 0.86	0.89 / 0.85
6	0.93 / 0.90	0.94 / 0.92	0.92 / 0.89	0.89 / 0.85
5	0.98 / 0.97	0.97 / 0.96	0.95 / 0.93	0.93 / 0.90
4	0.99 / 0.99	0.98 / 0.98	0.97 / 0.95	0.97 / 0.96
3	0.99 / 0.99	0.99 / 0.99	0.98 / 0.97	0.99 / 0.98

For PTE Core, decision accuracy ranges from 0.89 to 0.99 and consistency from 0.85 to 0.99. For both accuracy and consistency, the values are strongest in the range of CLB 3 to 6. As the PTE Core test is intended to target a general, non-academic testing population, this aligns with the purpose of the test. CLB 10 also shows strong values for consistency and accuracy. This reflects the fact that the CLB extends beyond the score reporting scale for this test, and there will be some test takers who demonstrate ability at or above CLB 10 who would be consistently and correctly classified as not below CLB 10.

6.4.2 Conditional Standard Error of Measurement (CSEM)

Conditional Standard Errors of Measurement (CSEM) express the amount of error associated with different scores across the reporting scale for a test. CSEM is calculated based on the score information functions at each ability value that corresponds to the scaled cut scores. Table 7 shows the CSEM for the PTE Core cut scores given in Table 5.

Table 7 – CSEM for PTE Cores scores at CLB levels

CLB	PTE Core Listening	PTE Core Reading	PTE Core Speaking	PTE Core Writing
10	11.3	9.3	13.5	9.2
9	9.0	7.1	9.8	8.2
8	7.5	5.9	8.5	7.6
7	6.2	5.0	6.8	6.6
6	5.0	4.4	5.3	5.8
5	3.9	4.0	4.3	5.0
4	3.6	4.0	3.8	3.9
3	4.5	4.4	3.7	3.5

The CSEM for cut scores at CLB 3 to CLB 7 are in the range of 3.5 to 6.8 points, which indicates a high degree of precision in the levels most important to decisions related to general, non-academic language proficiency. CSEM values are larger toward the top end of the scale, though still within an acceptable range, given the size and granularity of the PTE Core score reporting scale. These figures will be monitored and updated as further data is gathered on PTE Core in the live testing environment.

6.4.3 Judge evaluation questionnaire

Judges were asked to complete a survey to gather feedback on the clarity of instructions, their comfortability with the standard setting tasks, and their confidence in their final ratings. Survey responses were provided by 11 of the 14 judges.

3.5 to 6.8

points – High degree of precision in CSEM for cut scores in the range important for migration decisions.

Table 8 – Standard setting feedback survey responses

	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree
The advance information was clear.	8	2	–	1	–
The training session provided me with a clear understanding of the standard setting process.	8	3	–	–	–
I was able to relate Listening items to the CLB.	3	6	2	–	–
The discussion after Round 1 judgements for Listening helped me to refine my judgements for Round 2.	9	2	1	–	–
I feel confident in my final judgements for Listening.	5	6	–	–	–
I was able to relate Reading items to the CLB.	4	6	1	–	–
The discussion after Round 1 judgements for Reading helped me to refine my judgements for Round 2.	8	1	2	–	–
I feel confident in my final judgements for Reading.	8	2	1	–	–
I was able to relate Speaking responses to the CLB.	4	7	–	–	–
The discussion after Round 1 judgements for Speaking helped me to refine my judgements for Round 2.	11	–	–	–	–
I feel confident in my final judgements for Speaking.	8	3	–	–	–
I was able to relate Writing items to the CLB.	5	6	–	–	–
The discussion after Round 1 judgements for Writing helped me to refine my judgements for Round 2.	11	–	–	–	–
I feel confident in my final judgements for Writing.	9	2	–	–	

The feedback indicates that the majority of judges found the advance information and the training sessions clear. Generally, judges had more confidence in their ratings for productive skills than receptive skills. Judges commented that it was more straightforward to judge test taker responses than test item difficulty because the judging criteria aligned more naturally with their experience evaluating students.

The receptive skills pose a challenge, as judges are required to use more abstract reasoning to determine the difficulty of test items. That said, all judges either agreed or strongly agreed that they had confidence in their final ratings for both receptive and productive skill activities.

7. Standards verification

7.1 Initial standards verification

The aim of the verification activities was to review the appropriacy of CLB classifications of PTE Core test takers, particularly for productive skills. Standards verification took place approximately a year after the original standard setting, using additional field test data. The results are presented here in brief as initial evidence of the reasonability of the PTE Core to CLB alignment table. Further research is planned to verify the alignment in greater depth.

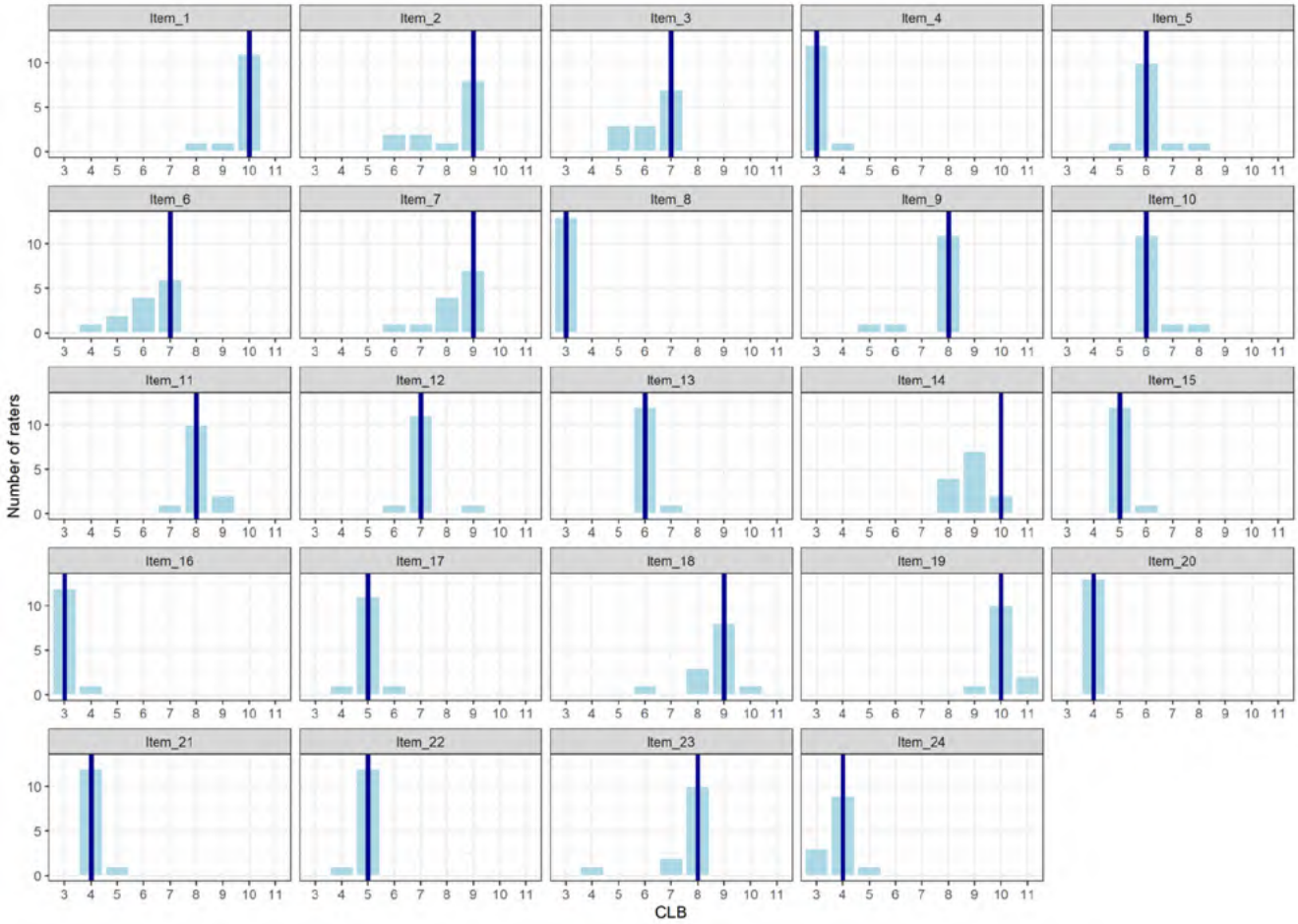
The standards verification activity took place virtually in July 2022. The same panel of 14 language experts participated. Judges carried out independent re-standardization activities prior to joining the virtual activity. Each activity began with an exercise to re-familiarize judges with the item types and scoring rubrics, the relevant CLB descriptors, and important points of differentiation across the CLB levels.

A sample of 24 test takers was selected from across the PTE Core score range for Speaking and for Writing. Test taker performances were classified in terms of CLB levels using the newly established alignment between PTE Core and the CLB. Judges were presented with the extended productive responses from each test taker, along with the CLB level assigned to each test taker based on their PTE Core score in the given skill.

Judges were asked to determine if each test taker's CLB classification was reasonable based on their holistic performance across multiple items for the given skill. If judges determined the classification was not reasonable, they were asked to provide the appropriate CLB classification for each test taker. This verification activity differs from the original standard setting in that judges were made aware of the test taker's CLB classification. This is intended to mimic the real-world scenario of test takers presenting their test scores as evidence of a CLB classification, and others forming opinions of whether that classification aligns with their impression of the test taker's proficiency.

For writing, the modal judge CLB rating agreed with the CLB rating based on PTE Core score for 23 out of the 24 test takers. Figure 1 shows Round 1 judge ratings for writing. Each panel represents an individual test taker, with the light blue bars showing the distribution of judge ratings for that test taker. The dark blue vertical reference line indicates the CLB classification for the test taker based on their PTE Core writing score.

Figure 1 – Writing Round 1 Ratings

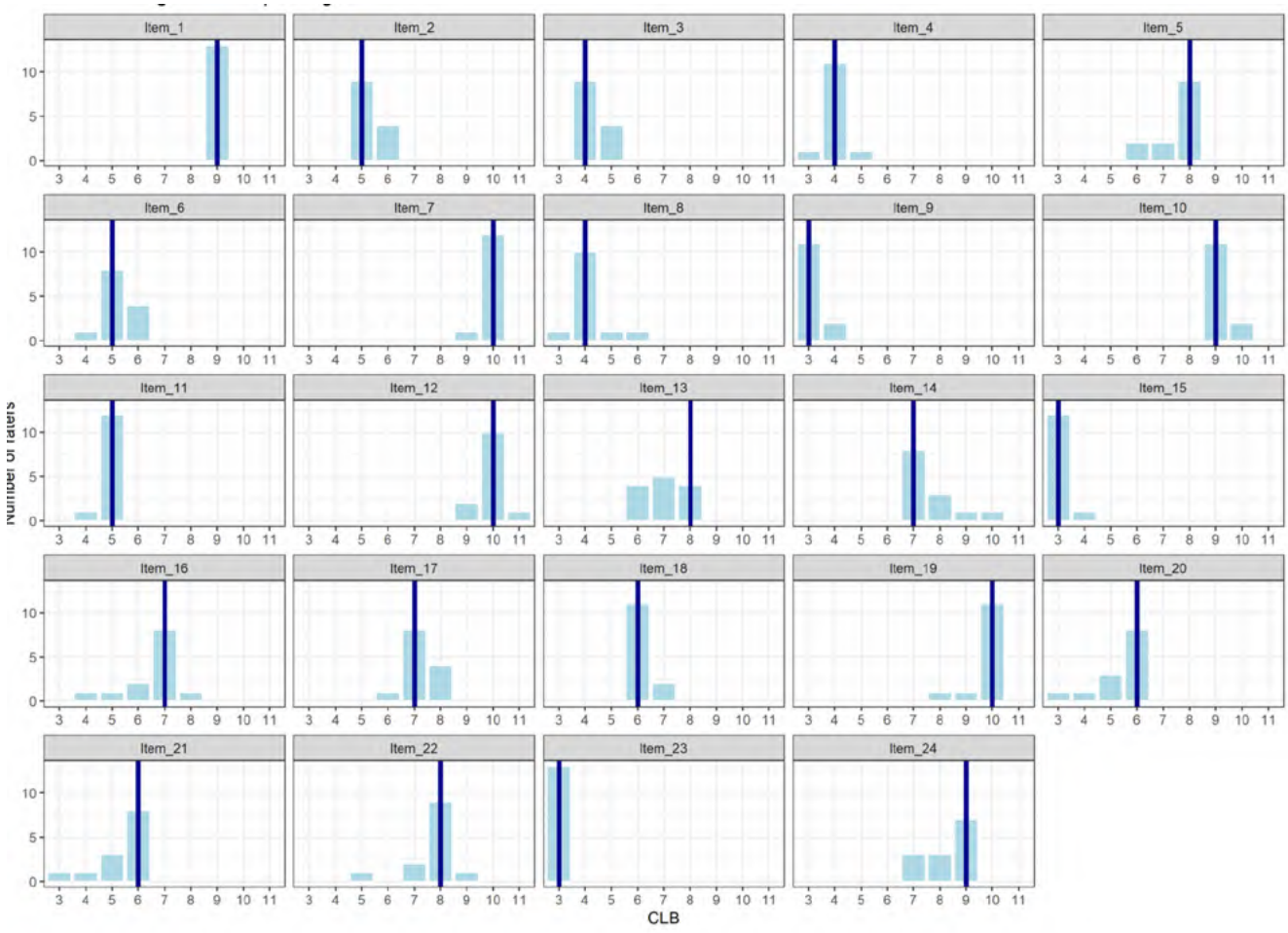


Only one test taker was classified differently by the majority of judges than by their PTE Core score *(Test Taker 14)*. This test taker was classified as CLB 10 by PTE Core, and received ratings ranging from CLB 8 to CLB 10 by expert judges. Following discussion and a second round of rating, this test taker's classification became less clear, with judge ratings between CLB 7 and CLB 10. The panel showed the least amount of agreement on their ratings for this test taker, indicating that something about this performance was particularly difficult to assess holistically.

For speaking, the modal judge CLB rating agreed with the CLB rating based on PTE Core score for 23 of the 24 test takers.

Figure 2 shows the Round 1 judge ratings for speaking.

Figure 2 – Speaking Round 1 Ratings



Similar to the writing activity, a single test taker in the speaking activity appeared difficult to classify *(Test Taker 13)*. The test taker was classified as CLB 8 by PTE Core, and received ratings of CLB 6 to CLB 8 from the judges. However, unlike the writing activity, judges were able to find a consensus rating after discussion.

>95%

the majority of judges agreed with the CLB classifications of 23 of 24 test takers (>95%) based on their PTE Core score.

The Round 2 ratings included ratings of CLB 7 and CLB 8, with the majority of judges settling on CLB 7.

In both the speaking and writing activities, the majority of judges agreed with the CLB classifications of 23 of 24 test takers (>95%) based on their PTE Core score. In the two instances of disagreement, the judges either could not arrive at a consensus rating or arrived at a consensus rating of the adjacent CLB level. While this verification activity is only an initial step, it provides a positive indication that the alignment between PTE Core scores and the CLB is reasonable and reflective of the standards expected by a panel of Canadian English language teachers and academics.

7.2 Future plans for standards verification

PTE Core is now available in the market, and standards will need to be verified and continually monitored as the test taker population grows. Additional triangulation of evidence is planned to validate the alignment between PTE Core test scores and the CLB, including:

- Standards verification activities to review the classification decisions of a large, representative sample of test takers in the operational test environment. These verification activities will include multiple panels of judges representing diverse stakeholder perspectives.
- Comparative analysis of score performance on multiple tests in controlled settings. Score concordance studies are planned to validate the theoretical alignment of performance standards with empirical evidence of score comparisons across distinct tests used for similar purposes.

8. Conclusions

Standards for PTE Core have been established throughout the development, field testing, and standard setting activities described in this paper. The alignment table between PTE Core and the Canadian Language Benchmarks (CLB) is based on robust field test data and the consensus judgements of a panel of expert practitioners who are highly experienced in both English language learning and assessment and the CLB. The results clearly define the existing relationship between PTE Core test scores and the Canadian Language Benchmarks, which will continue to be validated and monitored throughout implementation and live operation of the test.

9. References

Angoff, W. (1971). Scales, norms, and equivalent scores. In L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 508–597). American Council on Education.

CIC Canada. (2012a). *Canadian Language Benchmarks: English as a Second Language for Adults*. Ottawa: Centre for Canadian Language Benchmarks.

CIC Canada. (2012b). *CLB Support Kit*. Ottawa: Centre for Canadian Language Benchmarks.

Fox, J., & Fraser, W. (2014). *Setting university language proficiency entry requirements on the Pearson Test of English Academic (PTE Academic) in relation to performance categories on the Canadian Academic English Language (CAEL) Assessment. Pearson*.

Hambleton, R., & Pitoniak, M. (2006). Setting performance standards. In R. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 433–470). American Council on Education/Praeger.

Jones, G., De Jong, J., Zheng, Y., & Booth, D. (2017). *A standard setting study to establish concordance between the Pearson Test of English (PTE A) and the Canadian Language Benchmarks (CLB)*. Pearson. https://pearsonpte.com/wp-content/uploads/2018/07/Pearson-Standard-Setting-Study-between-PTEA-and-CLB.pdf

Lathrop, Q. (2015). *cacIRT: Classification Accuracy and Consistency under Item* Response Theory (R package version 1.4) [Computer software]. https://CRAN.R-project.org/package=cacIRT

Lee, W. (2010). Classification Consistency and Accuracy for Complex Assessments Using Item Response Theory. *Journal of Educational Measurement, 47*(1), 1–17. JSTOR.

Plake, B., & Hambleton, R. (2001). The analytic judgement method for setting standards on complex performance assessments. In G. Cizek (Ed.), *Standard setting: Concepts, methods, and perspectives* (pp. 283–312). Routledge.

Rudner, L. (2000). Computing the expected proportions of misclassified examinees. *Practical Assessment, Research, and Evaluation, 7*(1). https://doi.org/10.7275/an9m-2035

Rudner, L. (2005). Expected Classification Accuracy. *Practical Assessment, Research, and Evaluation, 10*(1). https://doi.org/10.7275/56a5-6b14

Tannenbaum, R., & Wylie, E. (2004). *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)*. Princeton, NJ: Educational Testing Service.

Pearson | PTE

About PTE

Learn about who we are and what we do.

PTE is a world-leading provider of secure English language tests

Pearson Test of English (PTE) provides secure English language testing for study applications worldwide, and for visa applications for work and migration in Australia, the UK, New Zealand, and Canada.

Every year, tens of thousands of people trust PTE to help them rapidly prove their English proficiency – and to open doors to their future lives and careers.

Download the full report
www.pearsonpte.com/research

pearsonpte.com