December 2017

A Standard Setting Study to Establish Concordance between the Pearson Test of English Academic (PTE A) and the Canadian Language Benchmarks (CLB)

Written by
Glyn Jones, Dr. John HAL de Jong,
Dr. Ying Zheng and David Booth

Reviewed by
Andrea Strachan

Pearson

Contents

1. Aims

The aim of this study was to establish a link between PTE Academic (PTE A) scores and the Canadian Language Benchmarks (CLB) four to ten. That is to say, the desired outcome of the research is a set of cut scores that correspond to the respective boundaries between the CLB levels four to ten for Listening, Reading, Speaking and Writing.

A standard setting methodology was selected as this type of study has been used successfully for similar studies establishing test cut scores for work or migration purposes against descriptions of language proficiency such as the Common European Framework for Languages (CEFR) and the CLB. These include: Tannenbaum, R. J. & Wylie, C. (2004), Tannenbaum, R. J. & Wylie, C. (2008), Woo, A., Dickison, P. & De Jong, J.H.A.L. (2010) and CaMLA (2017).

A separate standard setting session was conducted for each of the four skills. For Listening and Reading, a variant of the Angoff method (Hambleton & Pitoniak, 2006, p. 440f.; Cizek & Bunch, 2007, p. 81ff.) was used while for Speaking and Writing, a variant of the Contrasting Groups method (Hambleton & Pitoniak, 2006, p. 445.; Cizek & Bunch, 2007, p. 105ff.) was used. The sequence of steps was substantially the same for each session as described in detail in the methodology section of this report.

2. Panellists

A panel was recruited by advertising on appropriate regional, national and international discussion lists and through the contacts Pearson Canada already had with appropriate institutions.

Panellists were required to be experienced ESL practitioners with a thorough theoretical understanding of the CLB descriptors and practical experience in their application to English language instruction, curriculum development or assessment. On application, the panellists completed a form which collected key demographic and professional data. It was important that the panel represented Canada's geographical diversity as well as the range of educational establishments where the CLB are used. The key criterion, however, was familiarity with the Canadian Language Benchmarks as this was critical for the success of the study.

Table 1: Characteristics of the panel

Panel Characteristic	Category	%	(N)*
Gender	Female	91	21
	Male	9	2
Job Role	Instructor (Teacher, Trainer)	61	14
	Head of Department/Specialist	13	3
	Director	13	3
	(Assistant) Professor	13	3
Years of Experience with CLB	Under 5	9	2
	5 to 9	43	10
	10 to 15	26	6
	More than 15	13	3
	Not known	9	2
Canadian Provinces	Ontario	39	9
	Quebec	4	1
	British Columbia	9	2
	Alberta	17	4
	Nova Scotia	9	2
	Newfoundland and Labrador	-	-
	Saskatchewan	13	3
	Manitoba	9	2
	New Brunswick	-	-
	Prince Edward Island	-	-
Affiliation	University	35	8
	College	30	7
	Language Instruction for Newcomers to Canada (LINC)	26	6
	Other Institution	9	2

*Note one panellist attended the first day only

The panellists were primarily female, with most being practising instructors from programmes preparing students using CLB-based curricula. The panellists were experienced in using the benchmarks, with the vast majority having used them for five years or more. In addition, a number of panellists occupied more senior roles as heads of department, university professors, and directors of learning institutions. These panellists brought significant experience to the study including involvement in the elaboration and field testing of the CLB in the initial stages of their development.

The panellists were drawn from different regions of Canada, with the majority coming from Ontario, where the standard setting workshop was held.

3. Venue

The workshop took place on the 14th and 15th of September 2017 in a conference room at the Hyatt Regency Hotel in downtown Toronto. The room was equipped with a large screen data projector, and sound reproduction (for audio prompts and speaking samples) was provided via a high-quality audio system with large speaker units which was operated by a professional sound engineer.

The panellists sat at four large tables, sufficiently spaced so that they could not see each other's work when rendering individual judgments. Roving microphones, connected to the audio system, were provided, but the acoustics in the room were such that it was possible to hear the workshop facilitator as he presented the materials and conducted whole-panel discussions, without strain and without amplification.

4. Method

4.1 Advance preparation

In advance of the workshop panellists were sent, by email, a PowerPoint presentation giving an overview of PTE A, as well as links to online sources of information about the test.

For each of the receptive skills (Listening and Reading), a set of test items was prepared representing a cross section of the item types used to assess the respective skill in PTE A. Pearson has robust Rasch difficulty parameters for all of these items, derived from extensive field testing and live test administration, so the set of items represented the full range of difficulty. Rasch analysis, which provides the difficulty parameters for the test items, is recognised as an efficient way of establishing the difficulty of an item in relation to the ability of the test taker. It is a probability based model which predicts the likelihood of any particular test taker getting the item correct. This allows item difficulties to be estimated based on the probability threshold (normally 50%) rather than based on the performance of a particular test cohort i.e. norm-referenced.

For each of the productive skills (Speaking and Writing) a set of test taker responses was prepared. These were obtained in the course of field testing or from live test administration. Pearson has robust Rasch ability estimates (see above) and reported scores for each of the respective test takers, for the skill in question (Speaking or Writing), as well as for global language proficiency.

The items and/or responses were compiled into booklets for use in the workshop. The composition of the booklets and their use are described in greater detail below.

4.2 At the workshop

During the first session of the workshop, the facilitator provided an introduction to the test using a PowerPoint presentation and spoken explanations, concentrating on the item types that were to be used for standard setting. An example of each of these item types was shown in the same way as it would appear to a test taker, with audio playback where applicable. The scoring criteria for each of these item types was explained, as was the contribution that each item makes to the scores for the separate skills. The panellists were reminded that although PTE A awards separate scores for each of the four skills, several of the item types (integrated items) test more than one skill.

4.3 Collecting judgments

At the start of the workshop, each panellist was given a randomly assigned number which they were asked to keep confidential. This served to preserve anonymity not only in the course of feedback and discussions during the workshop, but also in subsequent reporting.

For each skill, a booklet was prepared containing the set of test items (and, in the case of Writing, test taker responses) to be used in the session. The booklet was used to present items (and responses) to the panellists and to record their judgments, which, in each case, consisted of a number between CLB 1 and 12. Within each booklet, the items or responses were presented in random order; they were neither ranked according to empirical difficulty or proficiency, nor grouped according to item type. This arrangement was chosen in order to encourage

the panellists to consider each item or response independently, and according to the descriptors, rather than by comparison with similar items or responses.

Two rounds of judgments were collected for each skill. In the first round, the panellists submitted their judgments individually, and without conferring, by entering them in the appropriate pages of their booklet. At the end of the first round, the booklets were collected and the judgments were transferred to a spreadsheet for analysis in preparation for the second round. A spreadsheet was created for each skill. Each spreadsheet was formatted with a row for each panellist (identified only by his or her secret ID number) and a column for each item. Two rows were added at the bottom of the spreadsheet to show the mode and the standard deviation of judgments for each item.

In the second round, the panellists were shown the first round judgments for the whole panel. Each panellist was identified by his or her confidential number thereby ensuring anonymity of actual scores. Measures of central tendency, means and modes were also computed and displayed. This enabled each panellist to see how their own judgments compared with those of the panel as a whole. The items which showed the greatest variance – generally those where the standard deviation was more than one benchmark, and/or where there was a range spanning five or more benchmarks – were selected for discussion.

While it was stressed that they should not feel obliged to revise their first round judgments, the panellists were given the opportunity to do so in the light both of the collective results and of the discussion. The panellists then submitted their second round judgments, again by entering them in their booklets. The booklets were collected once again and the revised judgments transferred to a second spreadsheet for analysis.

4.4 Listening

A selection of 23 Listening items was used, comprising five different item types. Table 2 shows the distribution of these item types.

Table 2: Distribution of Listening item types

Item type	Description	No. in booklet
Multiple choice, choose single answer	The test taker hears an audio clip (<60 seconds) and answers a single multiple choice comprehension question; only one option is correct.	4
Multiple choice, choose multiple answers	The test taker hears an audio clip (<90 seconds) and answers a single multiple choice comprehension question; two or more options are correct.	4
Highlight correct summary	The test taker hears an audio clip (<90 seconds) and identifies which of four written paragraphs is an accurate summary of the content of the clip.	4
Select missing word	The test taker hears an audio clip (<70 seconds) which is truncated, the last word or phrase being replaced by a beep. The test taker identifies which of four written options is the missing word or phrase.	4
Fill in the blanks	The test taker hears an audio clip (<60 seconds) while reading on screen a transcript of the clip from which certain words are missing. The test taker types the missing words in the gaps.	3
Highlight incorrect words	The test taker hears an audio clip (<50 seconds) while reading on screen a transcript of the clip in which certain words are altered from the original. The test taker identifies (by clicking) the words in the transcript that differ from the audio text.	4
	Total in booklet	23

It was explained to the panellists that in order to arrive at their judgment, they should, for each item, ask themselves: "What Canadian Language Benchmark best describes a learner who has a 50% chance of answering this item correctly?"

The booklet for Listening contained the prompts for the Listening items as they appear on screen in the test. For each item, the panellists were given a short time – approximating to the time allowed to test takers – in which to read the prompt in the booklet before the audio for the item was played once over the sound system. The panellists were then given a short time – again approximating to the time allowed in the test – in which to answer the item as if they themselves were taking the test. The answer key for the item was then displayed on the data screen in order to enable the panellists to check their answers. They were then given time to decide on their judgments and to record these on the respective page of the booklet. When all the panellists had entered a judgment, the facilitator moved on to the next item.

In the second round discussion, several panellists said that they had difficulty conceptualising a learner who has a 50% probability of answering correctly, and that they would have been happier with a more definite formulation such as "a learner who *can* answer correctly". This may be in part due to guidelines given to instructors in relation to the CLB:

"*As a general rule, the benchmarks assigned to a learner at the time of placement assessment, summative in-class assessment, or high-stakes language test, mean that the learner has achieved, and demonstrated, the level of communicative ability associated with most or all (traditionally, 70 to 100%) of the descriptors for the benchmarks assigned in each of the four skills.*" (CIC, 2013)

However, it was explained that the underlying measurement model of PTE A assumes that a test taker who is "at" a given level is one who has a 50% probability of successfully answering items at that level and that, therefore, it was appropriate to apply this principle when forming their judgments.

4.5 Reading

The procedure for Reading was the same as that for Listening except that in the first round, the panellists were allowed to work through the booklet at their own pace, as there was no audio to be played. The booklet contained 20 Reading items, comprising four different item types. Table 3 shows the distribution of these item types.

Table 3: Distribution of Reading item types

Item type	Description	No. in booklet
Multiple choice, choose single answer	The test taker reads a written text (<110 words) and answers a single multiple choice comprehension question; only one option is correct.	4
Multiple choice, choose multiple answers	The test taker reads a written text (<275 words) and answers a single multiple choice comprehension question; two or more options are correct.	4
Reading: Fill in the blanks	A text (<80 words) is presented with a number of gaps. The test taker fills the gaps by dragging the missing words from a pool of options.	8
Reading and Writing: Fill in the blanks	A text (<200 words) is presented with a number of gaps. The test taker fills the gaps by selecting from a drop-down list for each gap.	4
	Total in booklet	20

As with Listening, the panellists were asked to answer each item as a test taker, then to specify for each item the level of a hypothetical learner who has, in their judgment, a 50% chance of answering correctly. When all the panellists had worked through the booklet, the answer keys were displayed so as to enable them to check their answers. They were then given time to finalise their judgments and to enter these in the booklet.

4.6 Speaking

For Speaking, the panellists were asked to consider 20 responses recorded by test takers in the course of field testing or from live test administration. The booklet contained the prompts for the respective items as they would have appeared on screen for the test takers. The 20 responses were elicited by nine different items (so some items accounted for more than one of the responses to be rated) representing two item types: *Describe image* and *Retell lecture*. In all, nine different items were represented, so up to four responses were provided for each item. Table 4 shows the distribution of item types, items and responses.

Table 4: Distribution of Speaking items and responses

Item type	Description	No. different items	No. responses
Describe image	An informative graphic image is presented (a diagram, graph, chart, table or map). The test taker describes the image.	6	14
Retell lecture	The test taker hears an audio clip (<90 seconds). The test taker re-tells the content of the clip.	3	6
	Totals	9	20

It was explained that for each response the panellists should consider the question: "What Canadian Language Benchmark best describes the English language proficiency of the speaker?"

The responses were played over the sound system in booklet order. The *Retell lecture* items (the lecture extract which the test taker should re-tell) have an audio prompt as well as on-screen instructions. This was played the first time each of the three items of this type occurred, followed by the response. The panellists were given the option to hear the lecture again when a response to the same item recurred later in the booklet but in each case they declared unanimously that this was not necessary. For the *Describe image* responses, the panellists were given time to study the graphic prompt in the booklet before the response was played. After each response, they were given time to consider their judgments and record them in the booklet.

In the course of discussion in the second round, some panellists reported that they found it difficult to rate the responses to the *Describe image* items because these involve listening as well as speaking, so a failure to express something accurately might reflect a misinterpretation of the audio prompt rather than a deficit in speaking ability. Similar concerns were expressed in relation to other integrated items in the course of the workshop. As the aim

of the workshop was to link specifically to the CLBs for Speaking, the panellists were urged to consider, as far as possible, the responses as evidence of Speaking ability.

4.7 Writing

The booklet for Writing contained 20 responses derived from two item types: *Summarise spoken text* and *Write essay*. Six different items were used across the two types, with between one and seven responses presented for each item. As with Speaking, the responses were obtained in the course of field testing or from live test administration. Table 5 shows the distribution of item types, items and responses.

Table 5: Distribution of Writing items and responses

Item type	Description	No. different items	No. responses
Summarise spoken text	The test taker hears an audio clip (<90 seconds). The test taker types a summary of the content of the clip	4	12
Write essay	The test taker types an essay of up to 300 words on a given topic.	2	8
	Totals	6	20

For each response, the panellists were asked to consider the question: "What Canadian Language Benchmark best describes the English language proficiency of the writer?"

The audio prompts for the *Summarise spoken text* items were played over the sound system. As with the *Retell lecture* items (see above), the panellists agreed that they needed to hear each different prompts once only.

In the second round discussion some panellists expressed the view that the responses, particularly those to the *Summarise spoken text* items, were too short to enable them confidently to attribute a CLB level to the writer. At best, they felt that they could only place the writer within a range of levels, between a minimum (someone for whom this response represents the very best they can produce) and a maximum (someone who might also be able to produce writing of a higher order, but of which there is no evidence). They were advised to try to conceptualise the test taker who was *most likely* to produce the response.

5. Analysis

5.1 The Global Scale of English

The Global Scale of English (GSE) is a scale of 10 – 90 which is based on the PTE A reporting scale. The scale is truncated at 10 and 90 because at very low levels of ability, it is difficult to be confident of a scaled score and its relationship to other similar scores. Similarly, no person can attain a perfect level in the language so we truncate the highest scores at 90. This does not mean that people are unable to exhibit higher scores, but merely that these scores are not reported.

5.2 Methodology

A number of different data sources were used to relate the PTE A scale to the Canadian Language Benchmarks:

- Panel-rated difficulty of items on the CLB for receptive skills (Reading and Listening) based on the text of the items: this is termed an item-centred method.
- Panel-rated ability of test takers on the CLB for productive skills (Speaking and Writing) based on responses from test takers: this is termed a test taker-centred method.
- The difficulty values of the Reading and Listening items used in the workshop. In PTE A many items are polytomous and thus have multiple score points. For the analysis, all the score points were used, and not just the averages. The difficulties were computed from the very large data set which resulted from the field testing of PTE A. This field test had over 10,000 participants representing over 120 language groups.
- The ratings of the performances of Speaking and Writing. These were at the trait level for each task. Trait level refers to the specific concept which is measured. This varies by item type but in general for speaking the traits are: Content, Pronunciation and Fluency, for Writing the traits are Content, Grammar and Vocabulary with Discourse and Spelling additionally assessed in the Essay question. In addition to trait scores, the Speaking test score for test takers and the overall score on PTE A for all test takers' responses as presented in the workshop were used.
- The concordance between PTE A scores and IELTS scores as established in the PTE A/IELTS concordance study. Zheng, Y. & De Jong, J.H.A.L. (2011).

5.3 Data cleaning

Before analysis could begin, the data was cleaned. The two parameters used to do this are shown below:

- Remove individual panellist ratings if they differ by more than 1.5 benchmarks from the average rating given by panellists overall. In total 111 ratings (6%) were removed.

- Remove individual panellist ratings if their correlation with the average of all panellists is less than .5. In total 4 panellists were removed.

5.4 Certainty of ratings

Categorical scales like the CLB, are in fact simplified representations of an underlying continuous scale. One panellist might consider an item to represent an example of a high CLB 7, whereas another panellist may judge it to be an example of a low CLB 8. Therefore, although the panellists are observed to be a full level apart in their judgement, in fact they may be only a fraction of one level apart. To estimate the agreement of the panellists we can therefore best compute the maximum proportion of panellists rating within any two adjacent categories. For example, if we observe the following scoring pattern produced by the panellists for one of the items:

Table 6: Example of ratings from one item									
CLB Level	CLB 4	CLB 5	CLB 6	CLB 7	CLB 8	CLB 9	CLB 10	CLB 11	CLB 12
n raters	0	0	1	6	14	1	0	1	0
Proportion of raters	0.00	0.00	0.04	0.26	0.61	0.03	0.00	0.04	0.00

The average rating for this item is 7.83, so a high CLB level 7, almost a CLB level 8. We can also see that the maximum proportion of panellists within any two adjacent categories can be found in categories CLB 7 and CLB 8. They total a proportion of 0.87 of all ratings. There are only a few panellists who either rated lower than CLB 7 or higher than CLB 8. This means there is a high level of agreement among the panellists that the item is almost at CLB level 8.

If, however, for another item, we see a pattern like this:

Table 7: Example of ratings from a second item									
CLB Level	CLB 4	CLB 5	CLB 6	CLB 7	CLB 8	CLB 9	CLB 10	CLB 11	CLB 12
n raters	0	5	0	2	11	1	1	2	1
Proportion of raters	0.00	0.22	0.00	0.09	0.48	0.04	0.04	0.09	0.04

The average rating is also 7.83, but the maximum proportion of panellists within any two adjacent categories is just 0.57, which indicates quite some uncertainty about the level of this item. About one third of the panellists estimate that either this item is well below level 8 or well above it. Fortunately, such extreme cases did not occur in the data. There were only two items which were rated with less than 0.70: one item for reading with a certainty value of 0.60 and one item for speaking with a certainty value of 0.67.

Table 8: Average certainty of cleaning		
Average certainty after cleaning		
Skill	**Certainty**	**n items <0.7**
Listening	0.83	0
Reading	0.86	1
Speaking	0.83	1
Writing	0.80	0

5.5 Best fitting regression function from CLB to GSE (Method 1)

By plotting the original GSE values of the PTE A items against the average CLB ratings from the panellists a linear regression function can be computed. Four regression functions were computed, one for each of the skills, listening, reading, speaking and writing. In the following table the regression functions are given as also the squared correlation, which is an indication of the proportion of variance in the GSE explained.

Table 9: Regression functions by skill

Skill	Regression function	Correlation	Squared correlation
Listening	GSE = 9.41 × CLB - 14.64	r = 0.70*	r^2 = 0.60
Reading	GSE = 8.18 × CLB - 6.30	r = 0.70*	r^2 = 0.60
Speaking	GSE = 8.67 × CLB - 10.48	r = 0.90**	r^2 = 0.81
Writing	GSE = 7.72 × CLB - 5.88	r = 0.92**	r^2 = 0.85
*significant at p = 0.10; **significant at p = 0.01			

It can be noted that the high value of the explained variance for the productive items indicates a very accurate prediction of GSE values from the CLB ratings by panellists. The estimation of difficulty for the receptive items is generally less precise as there is no access to observable test-taker behaviour: the level estimates can only be based on what panellists think is the likely difficulty of the items for test takers.

Using the regression functions from the previous table, the corresponding GSE values on PTE A for each CLB can be computed for each of the skills scores and for the overall PTE A score.

Table 10: Computed PTE A scores for CLB

CLB	PTE A Reading	PTE A Writing	PTE A Listening	PTE A Speaking	Overall Score
12	90	87	90	90	89
11	84	79	89	85	84
10	76	71	79	76	76
9	67	64	70	68	67
8	59	56	61	59	59
7	51	48	51	50	50
6	43	40	42	42	42
5	35	33	32	33	33
4	26	25	23	24	25

5.6 Validation of CLB-GSE correspondence using IELTS-CLB (Method 2)

From previous research the relation between the GSE values of PTE A Overall score and the IELTS score bands are known. Zheng, Y. & De Jong, J.H.A.L. (2011). On the other hand, we also obtained information about the relation of the IELTS band scores for the four skills with the CLB. We can therefore compute the relation between the GSE values for the PTE A Overall score and the CLB via a common anchor. A first step is to compute the IELTS Overall band scores from the four skills scores. This is straightforward as the IELTS Overall band score is simply produced by averaging over the four skills band scores. Next we can look up the predicted PTE A scores from the concordance table in the PTE A Score Guide. The next table shows the results of these computations. The first column presents the CLB. The next four columns present the corresponding IELTS band scores for each of the four skills. Column 6 presents the computation of the Overall IELTS band scores by averaging the four skills. The last column presents the GSE values for PTE A that correspond with the Overall IELTS band score. If the CLB-IELTS relation is correct and if also the IELTS-PTE A concordance is correct, then the relation between CLB and PTE A can be read from the first and the last column.

Table 11: CLB, IELTS and PTE A

CLB	IELTS Reading	IELTS Writing	IELTS Listening	IELTS Speaking	IELTS Overall	PTE A
12	9.0	9.0	9.0	9.0	9.0	86
11	8.5	8.0	9.0	8.0	8.4	82
10	8.0	7.5	8.5	7.5	7.9	77
9	7.0	7.0	8.0	7.0	7.3	69
8	6.5	6.5	7.5	6.5	6.8	62
7	6.0	6.0	6.0	6.0	6.0	50
6	5.0	5.5	5.5	5.5	5.4	41
5	4.0	5.0	5.0	5.0	4.8	33
4	3.5	4.0	4.5	4.0	4.0	26

5.7 Comparison of Method 1 and Method 2

We now have two independent estimates of the concordance between CLB levels and GSE values for PTE A Overall scores. Method 1 is based on regressing GSE values of PTE A items on ratings provided by panel members on the item texts of receptive items (item-centred approach) and on test-taker responses to productive items (test taker-centred approach). Method 2 is based on known estimates of the relation between IELTS band scores and the CLB on the one hand and the relationship between IELTS band scores and PTE A GSE scores on the other. By comparing the estimated relationship between the CLB and PTE A GSE scores resulting from each of these two independent methods we provide a triangulation validation of the relationship between CLB and PTE A GSE scores. The following table compares the estimated relationship resulting from each of these methods.

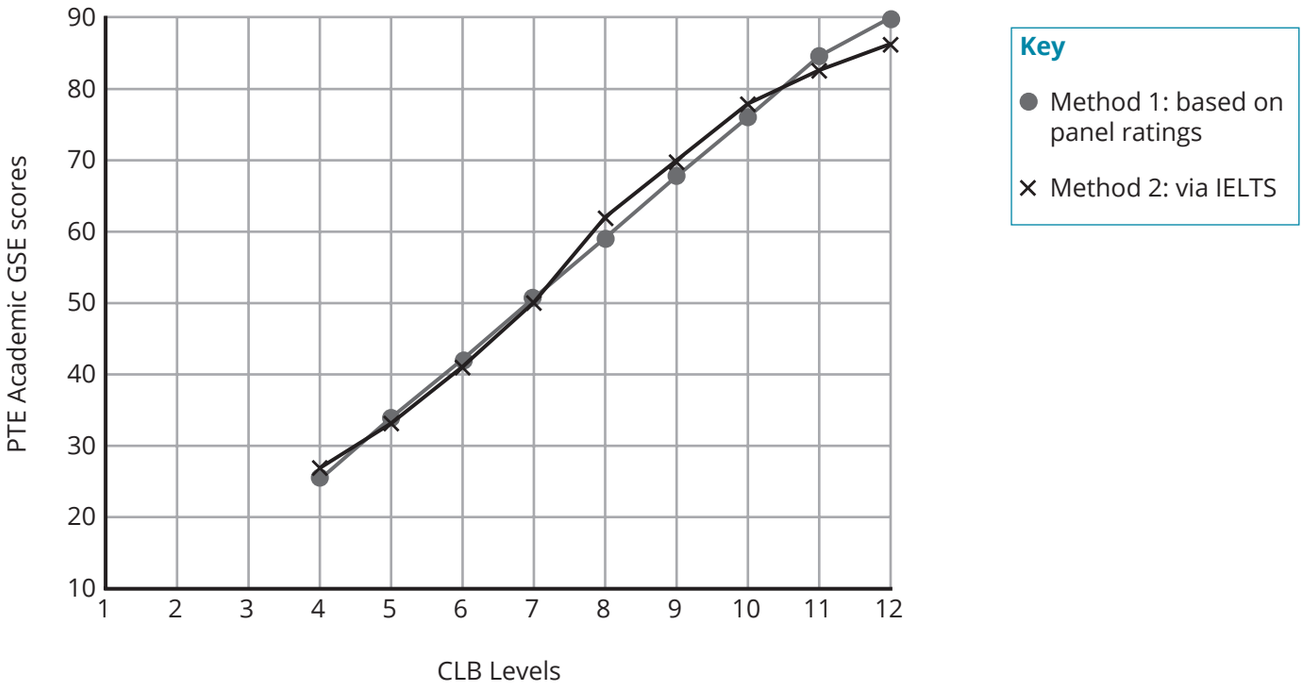
Table 12: Comparison of relationship from each method

	Method 1	Method 2	Difference
CLB	PTE A (GSE)	PTE A (GSE)	Method 1 - Method 2
12	89	86	-3
11	84	82	-2
10	76	77	2
9	67	69	2
8	59	62	3
7	50	50	0
6	42	41	-1
5	33	33	0
4	25	26	2

From the table we can conclude that the difference between the outcome of Method 1 and Method 2 is equal to or smaller than the 3-point error of measurement on the GSE scale. This provides strong support for the validity of the findings in this study.

A plot of both concordances reinforces the conclusion of close correspondence of the outcome of both methods.

Figure 1: Plot of concordance of two methods



6. Panel feedback summary

The panellists were invited to give feedback on their experience of the standard setting workshop by responding to an online survey. The survey was sent to the panellists in the week following the workshop, allowing them time to reflect before responding. All of the panellists responded.

The survey took the form of a series of statements to which panellists were asked to ascribe Likert scale categories: "strongly agree", "agree", "neither agree nor disagree", "disagree " and "strongly disagree". The survey also included an open-ended text item to enable panellists to submit additional comments. The summary of Likert responses in each category is shown in Table 13. (The option "strongly disagree", although available, was not selected in any instance so is not included in the table.)

Table 13: Summary of feedback survey responses

	strongly agree	agree	neither agree nor disagree	disagree
The advance information was clear.	7	13	0	2
The introductory presentation provided me with a clear understanding of the standard setting process.	6	10	1	5
I was able to relate the Listening items to the Canadian Language Benchmarks.	1	12	7	2
The discussion after Round 1 judgements for Listening helped me to refine my judgments in Round 2.	5	13	3	1
I feel confident in my final judgments for Listening.	5	15	1	1
I was able to relate the Reading items to the Canadian Language Benchmarks.	2	14	5	1
The discussion after Round 1 judgements for Reading helped me to refine my judgments in Round 2.	4	14	3	1
I feel confident in my final judgments for Reading.	5	15	0	2
I was able to relate the Speaking samples to the Canadian Language Benchmarks.	2	13	2	3
The discussion after Round 1 judgements for Speaking helped me to refine my judgments in Round 2.	4	11	4	2
I feel confident in my final judgments for Speaking.	4	13	1	2
I was able to relate the Writing samples to the Canadian Language Benchmarks.	4	13	1	3
The discussion after Round 1 judgments for Writing helped me to refine my judgments in Round 2.	4	14	2	1
I feel confident in my final judgments for Writing.	5	14	2	0

As can be seen, a clear majority of the panellists considered that the advance information they had received was clear. However, a significant minority (5 out of 22) felt that the introductory presentation on day one did not give them an adequate induction into the process. One of the panellists who responded "disagree" to this item remarked in additional comments that they would have appreciated more explanation of the aims of the workshop; another suggested that it would have been useful to work through some example items together before starting the standard setting process proper.

For each of the four skills, the survey addressed three areas: whether the panellists found it easy to relate the items or samples to the CLB; whether they found the discussions between Rounds 1 and 2 helpful; and the degree of confidence they had in their final ratings. Regarding the experience of the process itself, responses were overwhelmingly positive. Responses relating to Speaking were, however, marginally less positive than those for the other three skills. Four of the panellists remarked in additional comments that they found the speaking samples too short to enable them to make a confident judgment, and one commented on the poor recording quality of some of the samples. These factors may partially explain the less positive responses in relation to this skill. A particular issue identified both here and in the discussions which took place during the workshop was the difficulty in rating integrated tasks. This is most likely due to the way individual skills are dealt with as separate concepts in language frameworks and common assessment instruments as opposed to as a sub-set of language proficiency as in PTE A.

Interestingly, and for all four skills, the highest levels of agreement were reflected in the statements regarding confidence, suggesting that even those who found it difficult to relate the items to the CLB, or who felt they gained little from the discussion, were nevertheless able to arrive at decisions that they felt comfortable with. For Listening and Reading, 20 (out of 22) panellists responded "agree" or "strongly agree" to the statement about their confidence in their final judgments. For Speaking and Writing these figures were 17 and 19 respectively (out of 21; one panellist was absent on Day 2 and so missed the sessions for Speaking and Writing).

In general, the panellists' comments indicated that while they found the experience challenging, they also found it interesting and rewarding. One panellist commented: "Very informative session and very well run. The only issue I came across during the workshop is that I had some difficulties relating some of the items to a CLB benchmark, as CLB is generally task-based. The conversation and discussion which ensued during the rounds was enlightening and thought-provoking. Very interesting workshop which I enjoyed taking part in, thank you to all." This comment may highlight some of the differences between learning and testing where test tasks are designed to elicit specific pieces of language for assessment purposes rather than setting all activities in a task-based construct. That said all PTE A items require the test taker to perform a language task.

7. Conclusions

The standard setting panel are representative of a broad range of practitioners of the Canadian Language Benchmarks from different educational sectors and regions of Canada. The standard setting workshop was conducted following established methodologies as described in the methodology section above. The results of the workshop were analysed using standard statistical methods and show a consistent relationship with other measures.

The outcomes of the workshop provide a robust basis for establishing the cut off points on the PTE A scale for different levels on the Canadian Language Benchmarks.

8. References

CaMLA (2017) Linking the Common European Framework of Reference and the Michigan English Language Assessment Battery. *Technical Report Ann Arbor, Michigan: Cambridge Michigan Language Assessments*

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

CIC, Canada. (2013). *National Language Placement and Progression Guidelines*.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting : a guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: SAGE.

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

PTE Academic Score Guide: Retrieved from https://pearsonpte.com/wp-content/uploads/2017/08/Score-Guide.pdf. Pearson

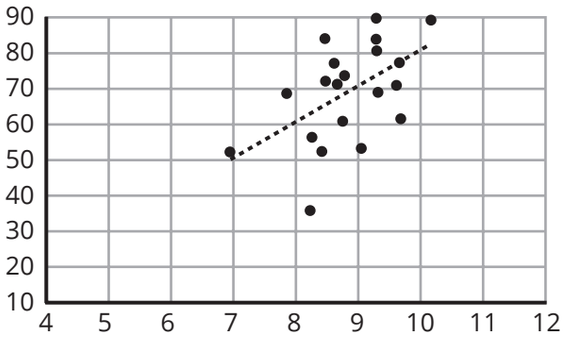
Tannenbaum, R. J. & Wylie, C. (2004) Mapping Test Scores Onto the Canadian Language Benchmarks: Setting Standards of English Language Proficiency on The Test of English for International Communication (TOEIC), The Test of Spoken English (TSE), and The Test of Written English (TWE). *Princeton, NJ: Educational Testing Services (ETS)*

Tannenbaum, R. J. & Wylie, C. (2008) Linking English-Language Test Scores Onto the Common European Framework of Reference: An Application of Standard-Setting Methodology. *TOEFL iBT Research Report TOEFL iBT-06. Princeton, NJ: Educational Testing Services (ETS)*

Woo, A., Dickison, P. & De Jong, J.H.A.L. (2010). Setting an English Language Proficiency Passing Standard for Entry-Level Nursing Practice Using the Pearson Test of English Academic. *NCLEX Technical Brief, National Council of State Boards of Nursing*

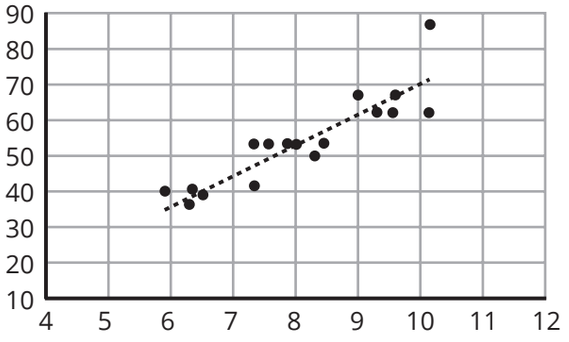
Zheng, Y. & De Jong, J.H.A.L. (2011). Establishing the construct and concurrent validity of Pearson Test of English Academic. *Pearson Research summaries and notes: www.pearsonpte.com/organizations/researchers/research-notes*

Appendix 1: Regression plots



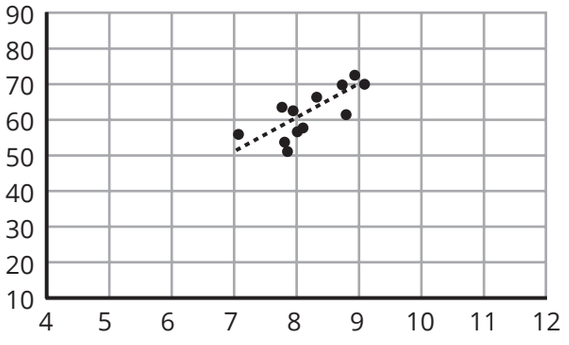
CLB Ratings Predict GSE Reading after Cleaning

$y = 8.184x - 6.301$
$R^2 = 0.4821$



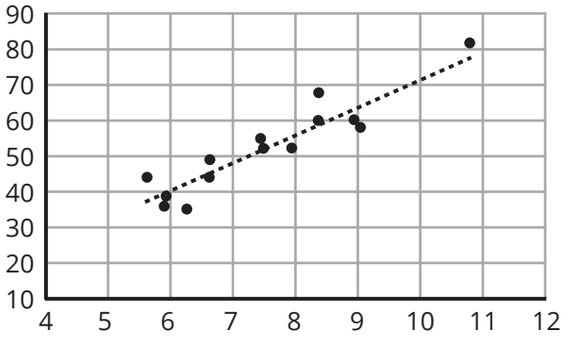
CLB Ratings Predict GSE Speaking after Cleaning

$y = 8.6667x - 10.479$
$R^2 = 0.8081$



CLB Ratings Predict GSE Listening after Cleaning

$y = 9.4097x - 14.643$
$R^2 = 0.603$



CLB Ratings Predict GSE Writing after Cleaning

$y = 7.7247x - 5.0884$
$R^2 = 0.8495$

Appendix 2: Panellists

We would like to acknowledge the vitally important contribution made by the members of the panel.

The members are: Ibitisaam Abboud, (Ottawa-Carleton District School Board (OCDSB), Ottawa, Ontario), Brett Basbaum, (HPL, Hamilton, Ontario), Heidi Bundshcoks, (Columbia College, Calgary, Alberta), Anita Chauduri, (Mount Royal University, Calgary, Alberta), Sandra Christensen, (Camosun College, Victoria, British Columbia), Tracey Derwing, (University of Alberta, Edmonton, Alberta), Sonia Fiorucci-Nichols, (Humber College, Toronto, Ontario), Anna Janik-Kelly, (Red River College, Winnipeg, Manitoba), Ghazal Lotfi, (University of Windsor, Windsor, Ontario), Ann Mackenzie, (Saskatchewan Polytechnic, Saskatoon, Saskatchewan), Chayan Mallick (Saskatchewan Polytechnic, Saskatoon, Saskatchewan), Erin McDonald, (Language Assessment Services of Nova Scotia (LASNS), Halifax, Nova Scotia), Paul Meighan, (Independent Consultant, Toronto, Ontario), Danita Midena, (OCDSB, Ottawa, Ontario), Cynthia Morehouse, (Saskatchewan Polytechnic, Saskatoon, Saskatchewan), Lisa Robertson, (Camosun College, Victoria, British Columbia), Shahrzad Saif, (Université Laval, Quebec), Debra Schweyer, (WELARC, Winnipeg, Manitoba), Andrea Strachan, (Touchstone Institute, Toronto, Ontario), Melissa Taylor, (Dalhousie University, LASNS, Halifax, Nova Scotia), Antonella Valeo, (York University, Toronto, Ontario), Sarah-Jane Williams, (OCDSB, Ottawa, Ontario), and Johan Woodworth, (York University, Toronto, Ontario).

Appendix 3: About the Authors, Reviewers and Support

Glyn Jones is a highly respected test developer and assessment consultant with over 30 years' experience in language education and assessment. Glyn has worked in a number of senior positions for Eurocenters, City and Guilds and Pearson. He is currently completing his Doctoral thesis at the University of Lancaster. For this project Glyn ran the Standards Setting Seminar and wrote much of the report.

Dr. John H.A.L. de Jong is a highly respected academic and testing professional. As Vice-President of Language Testing at Pearson he was a leading figure in the development of PTE A, and from this the Global Scale of English. Currently John is Chair of Language Testing at VU University in Amsterdam and continues to provide consultancy services for a wide range of organisations and national bodies. For this project John analysed the data and wrote certain sections of the report.

Dr. Ying Zheng is a noted scholar in the field of language testing. She is the Associate Professor of Modern Languages at the University of Southampton and the Director of the Confucius Institute. Ying obtained her MEd and PhD in Cognitive Studies from Queen's University in Canada, specializing in Second Language Testing and Assessment. For this project Ying acted as a consultant with specific experience of the Canadian context. She also advised on the data analysis and wrote certain sections of the report.

David Booth is Test Development Director at Pearson. He has over 30 years' experience in teaching, teacher training and language assessment with the British Council, Cambridge Assessment and Pearson. For this project he was responsible for organising the Standard Setting event and the delivery of the final report.

Andrea Strachan is the foremost expert on the Canadian Language Benchmarks (CLB) and a leader in the area of CLB assessments in Canada. Andrea was involved in the early stages of trialling the benchmark scale in 1996-99 and was an expert reviewer for the CLB 2010 CLB document. Since the publication of the original CLB in 2000, Andrea has developed English as a Second Language (ESL) teaching materials, curricula and assessments referenced to the CLB for a range of instructional contexts. She has been a Canadian Language Benchmarks Placement Test (CLBPT) trainer and is currently the Chief Examiner for the Canadian English Language Benchmarks Assessment for Nurses (CELBAN). Andrea applied her knowledge of the CLB and her standard setting experience in the review of the report prepared for this project.

Hassan Fouad Is a Project Support Manager based at Pearson, UK. He provided invaluable support to the project and was instrumental in its success. Hassan contacted potential participants through the appropriate channels observing the relevant protocols and seeking permission in advance of making direct contact. He maintained communication with all panellists to ensure they had the appropriate information before coming to the seminar. Hassan liaised with the staff in the Hyatt, Toronto to ensure the seminar ran smoothly with high quality technical support. Hassan also liaised with Pearson Canada staff to ensure timely payment of suppliers and reimbursement for the panellists' expenses.