Research Note:

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

Summary Report: *How the content of the PTE Academic and its reported scores relate to the reporting scale of the CAEL*

Submitted by:

Janna Fox, Director, Language Assessment and Testing Research Unit, Linguistics & Language Studies, Carleton University, Canada

Wendy Fraser, Testing Manager, Centre for English Language Assessment and Support (CELAS), Carleton University, Canada

June 2014

1. PURPOSE AND RECOMMENDATIONS

This report summarizes recommended cut-scores on the Pearson Test of English Academic (PTE Academic) which relate to the levels of proficiency defined by band scores and performance categories of the Canadian Academic English Language (CAEL) Assessment. Recommended PTE Academic cut-scores are based on:

1) the judgments of expert panelists, who participated in formal standard-setting sessions (Angoff, 1971; Cizek, 2001; Cizek & Bunch, 2007); and

2) the scores of test takers who took both the PTE Academic and the CAEL within the same week.

In undertaking two approaches to setting cut-scores on the PTE Academic, we were aware that the literature suggests there is no one best method for setting cut-scores (Livingston & Zieky, 1982), and different approaches may yield differing results (see Jaeger, 1989, for a review). Thus, it was advantageous to approach the setting of cut-scores using two methods as a means of validation.

Approach 1: Expert judgment
The initial approach to setting PTE Academic cut-scores in relation to the CAEL was drawn from de Jong, Li & Duvin (2010). This approach depended upon data provided to the researchers by Pearson. For the speaking and writing sub-tests, an *analytical judgment* or test taker-centred method (Jaeger, 1989) was used, where the actual responses of PTE Academic test takers on speaking and writing tasks were evaluated for their proficiency in relation to CAEL criterion-referenced scale descriptors and scores. For the listening and reading sub-tests, an *item-centred* approach was used, where the level of ability required to answer an item correctly was evaluated in relation to CAEL criterion-referenced scale descriptors and scores.

We have maintained the distinctions between *test-taker centred* and *item-centred approaches* (cf. Jaeger, 1989) as a means of highlighting differences in the focus of the panelists in judging concordance between the two tests. We recognize, however, that such terminology has been questioned in the literature (e.g., Hambleton & Pitoniak, 2006), because inevitably standard-setting panels utilize both item and person information.

Approach 2: Test Re-test Study
After the initial standard setting, we subsequently recruited a purposive sample of 15 test takers who took both the PTE Academic and the CAEL. The test takers' results were used to triangulate the cut-scores identified by the expert panelists' judgments.

RECOMMENDATIONS

The recommended overall passing score for PTE Academic is 60 (equivalent to CAEL band 70).

The recommended minimum sub-test scores for each sub-test on the PTE Academic are: Listening (60), Reading (60), Writing (60), and Speaking (60).

2. OVERVIEW OF THE CAEL ASSESSMENT AND THE PEARSON TEST OF ENGLISH ACADEMIC

As noted by Cizek & Bunch (2007) before undertaking any standard setting procedure, it is important to "describe the construct(s) or characteristic(s) assessed by the test and to articulate the relationship of the proposed cut-score(s) to the construct(s) of interest and the purpose that the test and performance standards are expected to serve" (p. 41). In the section below, we describe the PTE Academic and the CAEL, and then discuss their comparability in terms of construct, characteristics, and purpose.

THE PTE ACADEMIC is a computer-administered English language test, which, according to the test developer, measures a test taker's "English language proficiency to ensure success in courses and active participation in university and college-level education, where English is the language of instruction" (Official Guide to PTE Academic, 2010, p. 1). It was designed for use by "universities, higher education institutions, government departments and other organizations requiring academic-level English ... to determine the actual English language skills of applicants when making admission decisions" (p. 1).

The PTE Academic takes approximately three hours to administer (see Table 1, below), and provides information on test takers' proficiency in listening, reading, writing and speaking as well as an overall score.

Table 1. PTE Academic by skill focus, section item/task type and time

SKILL FOCUS	SECTION	ITEM/TASK TYPE	TIME
Introduction			Not timed (not marked)
Speaking	1	Personal Introduction	1 minute
	2	Text read aloud	30-35 minutes
		Sentence Repetition	
		Description (e.g., describe an image)	
		Lecture re-tell	
		Question short answer	
Writing	3-4	Summarize written text	20 minutes
	5	Summarize written text or Write essay	10 or 20 minutes
	6	Write essay	20 minutes
Reading		Multiple-choice (single answer)	32-42 minutes
		Multiple-choice (multiple answers)	
		Re-order paragraphs	
		Fill in the blanks (reading only)	
		Fill in the blanks (reading & writing)	
Listening	1	Summarize spoken text	20 or 30 minutes
	2	Multiple-choice (single answer)	23-28 minutes
		Multiple-choice (multiple answers)	
		Fill in the blanks	
		Highlight correct summary	
		Select missing word	

		Highlight incorrect words	
		Write from dictation	

Different versions of the test are balanced for total length. PTE Academic features integrated tasks in which, for example, reading or listening is followed by writing or speaking. Some sections are speeded (e.g., reading, speaking). Partial credit is awarded for some items/tasks (e.g. fill in the blanks, multiple choice with multiple answers, etc.). In addition to reporting an overall score (from 10 to 90), test takers are provided with a *skills profile* which indicates both raw scores and their relative position on a graph which illustrates the test taker's performance. Relative position (from 10 to 90) is reported for all of the sub-tests of listening, reading, speaking and writing (from 10 to 90), as well as for enabling skills (i.e., in grammar, oral fluency, pronunciation, spelling, vocabulary, and written discourse).

THE CAEL ASSESSMENT is a criterion-referenced, topic-based performance test, comprised of an integrated set of language activities (Fox, 2003, 2004, 2009). The language tasks and activities in the CAEL Assessment are systematically sampled from those that are commonly undertaken within the university academic community. The content for the tasks on the CAEL Assessment is drawn from introductory university courses at times when professors are introducing new topics to their students with the expectation that the students know little or nothing about the content. The test is comprised of representative tasks and performances that characterize academic study (see Table 2, below), for example:

- speaking about academic experience, information, or understanding,
- listening to, taking notes, and transferring or applying information on a topic introduced or extended by an academic lecture,
- reading and selectively applying information from academic articles and texts about a topic introduced or extended by a lecture, and
- incorporating what has been learned from the lecture and readings in writing a formal, academic response to an academic task.

Table 2. CAEL Assessment by skill focus, item/task type and time

SKILL FOCUS	ITEM/TASK TYPE	TIME
Speaking	Personal Introduction	25 minutes
	Lecture re-tell	
	Question short answer	
	Text read aloud	
	Impromptu mini presentation	
Writing	Write essay based on information in reading and listening lecture	45 minutes
Reading	Multiple-choice (single answer)	55 minutes
	Multiple-choice (multiple answers)	
	Short answer	
	Fill Charts and tables	
	Fill in the blanks	
	Label diagrams	
Listening	Summarize spoken text	25 minutes
	Multiple-choice (single answer)	
	Multiple-choice (multiple answers)	
	Fill in the blanks	
	Short answer response	
	Fill in tables or charts/information	

	transfer	
	Take notes on spoken text	
	Extended response	

COMPARABILITY

Similarities between the two tests are important in relating the scores on the PTE Academic to the CAEL Assessment. Both tests operationalize a construct of English for academic purposes (EAP) at the level of undergraduate/first-year university/college admission. Both tests feature integrated tasks and partial scoring where appropriate and both tests provide a proficiency profile to test takers and test users/decision makers, which supplements scores with a finer grain of information to enhance the validity of inferences drawn from test performances. Both tests are used for a similar range of decisions in high-stakes contexts, and report on four sub-skills.

There are, however, important differences between the two tests. With the exception of the speaking component, the CAEL Assessment is not administered by computer; all speaking and writing performances are marked by human raters; the test report does not provide a finer-grain (i.e., enabling skills) profile. Further, although both CAEL and PTE ACADEMIC tasks are integrated, they differ in that PTE Academic items/tasks sample from a wide range of topics and contexts; CAEL items/tasks are *fully integrated* within a single topic. CAEL test takers are provided with the essay prompt for the writing sub-test at the beginning of the test, introduced to the topic through the readings (two-three) with items/tasks that are used for the reading sub-test scores. The listening sub-test consists of an extended lecture on the same topic with items/tasks that are used for the listening sub-test scores. Test takers use the information from the readings and lecture in responding to the prompt in the writing sub-test at the end of the test.

PTE Academic reading and listening texts and tasks are shorter with mainly multiple choice or fill-in the blanks responses; CAEL reading and listening texts are much longer and involve extended reading and listening with such tasks such as written summaries, information transfer, and short answer responses.

All performance (sub-tests and overall score) on the CAEL Assessment is defined by criterion-referenced band scores ranging from 10 to 90.**These criteria served as performance level descriptions (Cizek & Bunch, 2007, p. 46) for categorizing PTE Academic performance/proficiency and/or item/task difficulty.**

Although CAEL Assessment proficiency standards are set internally by tertiary institutions in relation to their own programs, **in general, most institutions in Canada require a band 70 on the CAEL Assessment for admission**; **a number of institutions accept band 60**. Only two institutions in Canada require proficiency above band 70 for admission to their first-year, undergraduate programs.

3. APPROACH 1: EXPERT JUDGMENT

3.1 SELECTION FOR THE PTE ACADEMIC STANDARD SETTING PANEL

In planning the standard setting workshops, we were keenly aware of the importance of selecting participants: "Participants in the standard-setting process are critical to the success of the endeavor and are a source of variability of standard setting results" (Cizek & Bunch, 2007, p. 49). We initially sought to recruit 10 highly skilled experts for the workshop, but given the potential of a single individual to influence outcomes when the

panel is so small (John de Jong, personal communication, August, 2010) we sought to recruit as large a panel as possible, without sacrificing the quality of expertise. A core of 20 panelists was recruited for the two-day standard setting workshops. Because the workshops occurred over two days, we were concerned about attrition. Therefore, we also recruited back-up panelists (n=5). These panelists also attended the sessions and participated in the workshops.

Panelists were carefully selected to provide a sufficiently large and representative sample of expertise drawn from the following key groups:

1) **Admissions officers/registrars**, who routinely review test scores for university admission and credentials;

2) **Certified language proficiency raters** (CAEL, DELNA, IELTS, CANTest, CET);

3) **EAP specialists** who currently teach in concurrent EAP programs (i.e., programs which grant admission to students with a minimum threshold of proficiency, but require EAP support during the first term/terms of study. Students must achieve a level of proficiency equivalent to that established by standardized test scores before they are deemed to have satisfied the language admission requirement);

4) **Graduate students in Applied Linguistics**, with backgrounds in language teaching in College/University programs in Canada or abroad, who were specializing in language testing/assessment at the MA or PhD level;

The standard setting panel represented key expertise in the evaluation of English language proficiency relevant to the main selection context of both the PTE Academic and the CAEL Assessment. Faculty or staff from four Canadian universities and three colleges participated in the standard setting sessions.

3.2 STANDARD SETTING PROCESS AND PROCEDURES

3.2.1PREPARATION FOR STANDARD SETTING

Once panelists were recruited, they were provided with an information package, which included initial instructions for the session, information about the PTE ACADEMIC, information about the CAEL Assessment. Participants also consulted the websites for the two tests.

At the first session participants were presented with a review of key tasks and item types on the PTE ACADEMIC, followed by a question and answer session in which key similarities and differences between the CAEL Assessment and PTE ACADEMIC were discussed. Of particular interest were the CAEL criterion descriptors for each of the skills of reading, speaking, writing, and listening, which had been circulated in advance of the session, and served as the performance level descriptions (PLDs)/critical referent criteria for the participants' judgments of tasks/items or performances/proficiencies during the standard setting. In all cases, the PLD were considered in relation to minimally adequate levels of performance required for engaging in university level study in English.

3.3STANDARD SETTING METHODS AND PROCEDURES

In order to relate PTE ACADEMIC items, tasks. or performances to the CAEL passing or cut-score/performance band descriptors, an extended Angoff (Cizek & Bunch, 2007) or *iterative* approach (Jaeger, 1989) was used, wherein judges were given the opportunity to discuss their initial ratings, provided with additional statistical information on the group's assessment, and offered a second opportunity to record their judgments.

As noted in de Jong et al. (2010) for productive skills such as speaking and writing the test taker-centred method asks standard setting participants to consider actual test taker performances on the PTE ACADEMIC tasks or items and evaluate the level of proficiency demonstrated by the test taker in relation to the CAEL criterion-referenced scale descriptors and scores. On the other hand, for receptive skills (i.e., listening and reading) the item-centred method asks participants to evaluate the level of ability required to provide a correct response in relation to the CAEL standard (i.e., the minimally adequate level of proficiency in English required to engage in university-level study).

3.4ANALYSIS

Following the work of de Jong et al. (2010), rather than providing full test forms, Pearson provided the researchers with a subset of items from the test, drawn from the PTE ACADEMIC item bank, and "scaled on a single IRT scale" (p. 11). de Jong et al. argue that such calibrated items are "sufficient to predict scores on any test form …[because] the subset is sufficiently representative …[and] stratified in such a way that test forms have equivalent test information functions" (p. 11). This approach is further explained by de Jong et al.:

> In the perspective of Item Response Theory both the ability demonstrated by test takers when solving an item correctly and the difficulty of an item refer to the same point on an interval scale, because the difficulty of an item is identical to the ability required to solve an item correctly. As all scores on the PTE Academic reporting scale, including the overall score, are expressed on the same interval scale, [and]…lead to an estimate of the score that is required to meet the linguistic demands as an entry-level [student]. (p. 9)

Specifically, a three-step standard setting process was used for each skill:

Step 1: Each judge evaluated the PTE ACADEMIC *performance* (writing/speaking) or *task/item difficulty* (listening/reading) in relation to CAEL band scores and descriptors. The performance or task/item was assigned a CAEL band level based on the mean proportion of agreement. If the mean proportion of agreement was above .8, we proceeded to step 2. If not, statistical information was presented to the judges about the group evaluation, the group discussed the items, and the judges participated in a second round of evaluation.

Step 2: The band level judgments of all panelists were transformed into a scaled value using the de Jong et al. formula (Table 3). The scale is a 7-point *CAEL scale* defined by band levels (1 = CAEL band 30; 2 = CAEL band 40 etc.) (see Table 3 for an example of how the transformation to the CAEL scale was calculated).

Step 3: The judges reviewed each performance or task/item again and voted whether the performance or difficulty was 'too low', 'just right', or 'too high' for minimum entry/CAEL band 70. The votes were transformed to a scaled position on a 3-point scale (1=too low; 2=just right; 3=too high).

TABLE 3 : Transformation calculation sample (PTE ACADEMIC Item 1; speaking)

Step 1 Mean proportion of agreement: .95

The mean proportion of agreement is based on adjacent categories, i.e., the sum of the number of votes in the two most frequently selected categories, divided the total number of votes over the three adjacent categories. In the example below, (6 + 14)/(6 + 14 + 1) = .95.

Step 2 Speaking Transformations from judgments of PTE ACADEMIC speaking performance (item 1) to 7-point CAEL scale (based on CAEL band criteria).

ITEM NUMBER 1 (20 core panelists + 1 back-up)

CAEL Bands (using a 7-point scale) →	**30 (1)**	**40(2)**	**50(3)**	**60(4)**	**70(5)**	**80(6)**	**90(7)**	
Number of panelists' votes	**6**	**14**	**1**	**0**	**0**	**0**	**0**	**N/A**
Percentage (agreement)	**.29**	**.66**	**.05**					**N/A**
Total → Calculations= CAEL Band level (expressed on a 7-point scale) x percentage of agreement (based on distribution over three adjacent categories)	**1 x (.29) = [.29] +**	**2 x (.66) = [1.32] +**	**3 x (.05) = [.15] +**	**4 x ( ) = [ ] +**	**5 x ( ) = [ ] +**	**6 x ( ) = [ ] +**	**7 x ( ) = [ ] +**	**1.76 + .5* = <u>2.26(Band 40)</u>**

***.5 is added as an adjustment (de Jong et al. , 2010)**

Step 3: Transformations from judgment to CAEL level cut-off score (CAEL band 70 = 'just right' or minimally adequate required level of performance for entry-level admission represented by scalar point 2)

Decision level	(1) too low	(2) just right	(3) too high	Total
Number	**21**			
Percentage	**100.**			
Calculations Scale position (expressed on a 3-point scale) x percentage of agreement in each of the three categories)	**1 x (1) = [1] +**	**2 x ( ) = [ ] +**	**3 x ( ) = [ ] +**	**__1_ + .5 = 1.5 (scale position)**

***.5 is added as an adjustment by (de Jong et al., 2010)**

3.5 RESULTS

In this section, we report the results by sub-test skill.

Writing

Step 1: In order to determine a cut-off score on the PTE ACADEMIC that reflected the level of performance required to meet the standard set by the CAEL Assessment (i.e., band 70), expert judges assigned a CAEL performance band score category to previously rated PTE ACADEMIC writing samples. Proportions of agreement were then calculated (Table 4).

As noted above, if the proportion of agreement for an item was above .80, the item was not considered in a second round. In order to reach a satisfactory level of agreement, however, 13 items were reconsidered in round 2. The overall proportion of agreement was .89 (Table 4) with a standard deviation (s.d.) of .07.

Step 2: All resulting judgments were transformed to scaled values (see example in Table 3) using the formula provided by de Jong et al. (2010). Table 5 below provides a summary of CAEL scaled values (i.e., scaled value) for each of 20 PTE ACADEMIC tasks/items.

Step 3: The standards or cut-off decisions (i.e. assigning 'too low', 'just right', or 'too high') were also transformed using the de Jong et al. (2010) formula. Table 4 below provides an overview of the task judgments and transformations in relation to PTE ACADEMIC total test scores (i.e., level value).

Table 4. Overview of Task Judgments for the Writing Subtest

Item	PTE ACADEMIC **Total Test Score**	**Proportion of Agreement**	**Scaled Value**	**Level Value**	**STANDARD**
1	35.00	1.00	2.00	1.50	too low
2	36.00	1.00	2.00	1.50	too low
3	39.00	.83	3.38	1.54	too low
4	39.00	.96	3.79	1.85	too low
5	44.00	.87	2.14	1.54	too low
6	44.00	.91	3.05	1.59	too low
7	49.00	.83	2.81	1.58	too low
8	52.00	.80	3.05	1.67	too low
9	52.00	.92	3.92	1.96	too low
10	55.00	.87	3.22	1.67	too low
11	58.00	1.00	4.33	1.89	too low
12	58.00	.91	4.94	2.55	just right
13	60.00	.88	4.75	2.25	just right
14	60.00	.83	4.59	2.37	just right
15	60.00	.92	5.91	2.71	too low
16	68.00	.83	3.53	1.67	too low
17	68.00	.83	5.80	2.76	too low
18	82.00	.87	6.24	2.84	just right
19	82.00	.80	6.04	3.11	too high
20	68.00	.91	3.61	1.85	too low

In order to test the relationship between the PTE ACADEMIC total scores on the writing tasks and the judgments of the panel, we analyzed the data using Pearson's r. A significant correlation was found between the Standard and Total Test Score (r=.61, p<.01).

Given that these findings indicated a significant relationship, we next performed a regression analysis to predict PTE ACADEMIC Scores based on the CAEL Standard identified by the panelists. A regression analysis predicting PTE Academic scores from the CAEL Standard was statistically significant. The regression model accounted for 37% of the variance, with Standard as predictor and PTE ACADEMIC Total Score as dependent variable (β=.61, t=3.26, p <.01). PTE ACADEMIC scores for the writing tasks ranged from a low of 35 to a high of 82. The analysis suggests that Standard levels (i.e., too low, just right, too high), increase by approximately 15 point increments on PTE ACADEMIC test scores. Using only the regression analysis, this would suggest that level one — too low (i.e., band 50 or lower on the CAEL Assessment) -- would encompass PTE ACADEMIC scores from 35 to 50 points; level 2 -- just right (i.e., bands 60 or 70 on CAEL) -- would encompass scores from 51-66; level 3-- too high (i.e., bands 80 or 90 on CAEL) -- would encompass 67-82 on the PTE ACADEMIC. However, the judgment transformations, calculated based on the formula provided by de Jong et al., provided a more definitive cut-score for the "just right" classification, with level value difficulty ranging from 2.25 to 2.55 [see Table 4, above] (in CAEL terms, the cut-point between bands 60 and 70), or from 58-60 on the PTE ACADEMIC.

Based on a review of all results, we recommended that the cut-score for the PTE ACADEMIC in writing be set at 60 (which most closely corresponds to CAEL band 70).

Reading

Step 1: In order to determine a cut- score on the PTE ACADEMIC that reflected the level of difficulty required to meet the standard set by the CAEL Assessment (i.e., band 70), expert judges assigned a CAEL band score to previously calibrated PTE ACADEMIC reading items. Proportions of agreement were then calculated (Table 5). In total, 11 of the items/tasks were reconsidered in a second round which resulted in an overall proportion of agreement of .91, s.d = .06.

Step2: All resulting judgments were transformed to scaled values (see example of transformation in Table 3) . Table 5 below is a summary of CAEL scaled values in relation to reading items on the PTE ACADEMIC.

Step 3: The standards or cut-off decisions (i.e. assigning 'too low', 'just right', or 'too high') were also transformed using the de Jong formula. Table 5 below provides an overview of the judgments and transformations in relation to the difficulty estimates, in the form of PTE ACADEMIC total test scores, provided by Pearson.

Below is an overview of task judgments in relation to item difficulty estimates provided by PTE ACADEMIC(see de Jong for further information re. "Ability Delta", which he uses instead of theta for person ability).

Table 5. Overview of Task Judgments for the Reading Subtest (N=20 core panelists)

Item	Ability Delta	Estimated Total Score	Proportion of Agreement	Scaled Value	Level Value	STANDARD
1	**-.45**	**46**	**.95**	**3.95**	**1.80**	too low
2	**-.57**	**43**	**.95**	**5.35**	**2.65**	just right
3	**-.42**	**45**	**.90**	**5.05**	**2.50**	just right
4	**.73**	**67**	**1.00**	**4.25**	**2.15**	just right
5	**-.35**	**47**	**.95**	**4.00**	**1.85**	too low

6	**.80**	**69**	**.95**	**5.05**	**2.75**	too high
7	**.11**	**74**	**.85**	**5.85**	**2.95**	too high
8	**-.07**	**52**	**.85**	**3.95**	**1.75**	too low
9	**-.89**	**37**	**.90**	**4.05**	**2.00**	too low
10	**-.02**	**53**	**.85**	**4.55**	**2.40**	just right
11	**.03**	**54**	**.95**	**4.70**	**2.50**	just right
12	**.14**	**56**	**.90**	**4.90**	**2.65**	just right
13	**.85**	**69**	**1.00**	**4.20**	**2.10**	just right
14	**.17**	**57**	**.85**	**5.60**	**3.15**	too high
15	**.26**	**58**	**.85**	**5.10**	**3.70**	too high
16	**.27**	**59**	**.85**	**6.50**	**3.35**	too high

In order to analyze the relationship between the CAEL Band levels and the estimated total score, we ran a Pearson's r correlation with significant results ($r = .53$, $p < .05$). We next performed a regression analysis to predict PTE ACADEMIC based on the Standard identified by the panelists. A regression analysis predicting the PTE Academic total scores with the Standard was statistically significant. The regression model accounted for 36% of the variance, with Standard as predictor and PTE ACADEMIC total estimated score as dependent variable ($\beta = .53$, $t = 2.343$, $p = .034$). Total estimated scores for the items/tasks considered by the panel ranged from 37-74.

The analysis suggests that Standard levels (i.e., too low, just right, too high), increase by approximately 7.625 point increments. Using only the regression analysis, this would suggest that level one — too low (i.e., band 50 or lower on the CAEL Assessment) -- would encompass PTE ACADEMIC estimated total scores from 37 to 45; level 2 -- just right (i.e., bands 60 or 70 on CAEL) -- would encompass PTE ACADEMIC total score estimates from 46 to 54; level 3-- too high (i.e., bands 80 or 90 on CAEL) -- would encompass PTE ACADEMIC total score estimates above 55. Given the transformed level difficulty estimates, however, based on the de Jong et al. formula, the exact cut point between band 60 and 70 is not located at 2 (i.e., 'just right), but rather between 2.65 and 2.74.Based on the consideration of these results, the recommended cut or passing score on the PTE ACADEMIC is 60, which approximates CAEL band 70.

Speaking

Step 1: In order to determine a cut-score on the PTE ACADEMIC that reflected the level of performance required to meet the standard set by the CAEL Assessment (i.e., band 70), expert judges assigned a CAEL performance band score category to previously rated PTE ACADEMIC speaking samples. The overall proportion of agreement based on two rounds of evaluation was .89, sd = .06.

Step 2: All resulting judgments were transformed to scaled values using the formula provided by de Jong et al. . See Table 6 below for a summary CAEL scaled values in relation to speaking performance samples on the PTE ACADEMIC.

Step 3: The standards or cut-off decisions (i.e. assigning 'too low', 'just right', or 'too high') were also transformed using the de Jong formula. Table 6 below provides an overview of the task judgments and transformations in relation to PTE ACADEMIC total test scores, provided by Pearson.

Table 6. Overview of Task Judgments for the Speaking Subtest (N=20)

Item	Total Test Score	Proportion of Agreement	Scaled Value	Level Value	STANDARD

1	36.00	.95	2.36	1.50	too low
2	39.00	.81	2.98	1.55	too low
3	40.00	.90	2.80	1.50	too low
4	40.00	.85	2.30	1.55	too low
5	41.00	.80	2.85	1.60	too low
6	41.00	.85	3.55	1.90	too low
7	50.00	1.00	4.05	2.40	just right
8	53.00	.80	3.64	1.85	too low
9	53.00	.90	4.00	2.39	just right
10	53.00	1.00	3.92	2.19	just right
11	53.00	.80	3.85	2.08	just right
12	53.00	.85	3.25	1.76	too low
13	67.00	.95	5.00	2.82	just right
14	67.00	.85	4.00	2.50	just right
15	67.00	.85	4.55	2.55	just right
16	62.00	.95	4.95	2.92	just right
17	62.00	.85	5.40	2.97	just right
18	62.00	.95	5.70	2.93	just right
19	87.00	.90	6.16	2.92	just right
20	87.00	.90	5.40	2.95	just right

In order to test the relationship between the PTE ACADEMIC total scores on the speaking tasks and the judgments of the panel, we analyzed the data using Pearson's r .Significant relationships were found between the CAEL Standard and PTE ACADEMIC Total Test Score ($r=.73$, $p<.01$).

We next performed a regression analysis to predict PTE ACADEMIC scores based on the CAEL standard identified by the panelists. It was significant ($\beta=.73$, $t=4.486$, $p<.01$) and accounted for approximately 53% of the variance. PTE ACADEMIC scores for these speaking tasks ranged from a low of 36 to a high of 87. The analysis suggests that CAEL Standard levels (i.e., too low, just right, too high), increase by approximately 20 point increments on PTE ACADEMIC. Using only the regression analysis, this would suggest that level one — too low (i.e., band 50 or lower on the CAEL Assessment) -- would encompass PTE ACADEMIC scores from approximately 36(and below) to 56 points; level 2 -- just right (i.e., bands 60 or 70 on CAEL) -- would encompass scores from 57-77; level 3-- too high (i.e., bands 80 or 90 on CAEL) -- would encompass 77 (and higher).

As demonstrated in Table 6 above, there is a gap of 14 points (i.e. 53 to 67 between items #12 and #13) in the Total Test Scores provided by PTE ACADEMIC. This gap is situated at the key separation point between minimally adequate and too low levels of speaking proficiency – the critical cut-score location. Therefore, we were unable to estimate the cut-score between CAEL bands 60 and 70 with the same degree of precision as was possible for the writing and reading analysis.

The scores that were provided ranged from 36 to 87 points. Based on our analysis, these scores correspond to bands 50 through 90 on CAEL. The mean of the transformed level values (i.e. transformed values for 'too low', 'just right', 'too high') is 2.24 (s.d. .56). This suggests that the cut-score separating CAEL band 60 from 70 is located between level value 1.68 and level value 2.80(i.e., 2.24 +/- .56). However, the threshold, where the items/tasks demonstrate a level that is 'just right' (Table 6) is at 2.08 (approximately 10% more of the 'just right' band width from 57-77 (PTE scores).This suggests that the critical cut-point between CAEL band 60 and CAEL band 70 is 10% of the 20 point spread or 2 points. Thus, after a review of the results, we recommended that60 be set as the cut-point on the PTE ACADEMIC, which is equivalent to band 70 on CAEL.

Listening

Step 1: In order to determine a cut-off score on the PTE ACADEMIC that reflected the level of difficulty required to meet the standard set by the CAEL Assessment (i.e., band 70), expert judges assigned a CAEL band score to previously calibrated PTE ACADEMIC listening items. Proportions of agreement were then calculated (Table 7), with the overall proportion of agreement after two rounds at .88, s.d.=.04.

Step 2: All resulting judgments were transformed to scaled values using the formula provided by de Jong et. al. (2010).Table 7provides a summary of CAEL scaled values in relation to listening items on the PTE ACADEMIC.

Step 3: The standards or cut-off decisions (i.e. assigning 'too low', 'just right', or 'too high') were also transformed using the de Jong formula. Table 7 below provides an overview of the judgments and transformations in relation to the difficulty estimates, in the form of PTE ACADEMIC total test scores, provided by Pearson.

Table 7. Overview of Task Judgments for the Listening Subtest (core panelists= 20 + 5 back-up)

Item	Ability Delta	Estimated Total Score	Proportion of Agreement	Scaled Value	Level Value	STANDARD
1	.83798	69	.96	4.73	2.50	just right
2	-.40523	45	.88	4.64	2.59	just right
3	-.08573	52	.88	4.69	2.61	just right
4	.21721	57	.92	3.92	2.21	just right
5	.66335	66	.88	3.96	2.42	just right
6	.10432	55	.92	4.88	2.50	just right
7	-.09588	53	.88	4.20	2.04	just right
8	.18049	57	.92	4.58	2.16	just right
9	.32269	59	.88	5.71	2.65	just right
10	.45884	62	.88	4.47	2.73	just right
11	.69382	66	.88	4.84	2.50	just right
12	.77953	68	.85	4.89	2.11	just right
13	.50048	63	.81	4.66	2.48	just right
14	.02461	54	.88	4.39	2.27	just right
15	.63043	65	.81	5.20	2.27	just right
16	.67063	66	.81	4.73	2.23	just right

Although transformation of panelists' decisions indicates a small degree of discrimination between the tasks/items in terms of difficulty, it was clear, based on a follow-up discussion, that they were systematically applying the criteria from the CAEL, and following the advice provided by PTE ACADEMIC in terms of text/task characteristics. The panelists concluded that all of the items/tasks met the 'just right' criteria. We conducted a second, independent session (N=10) to evaluate each of the test tasks/items again, with the same outcome.

When the second panel arrived at the same result, we contacted Pearson to confirm that the calibrations for 'Delta ability' were indeed accurate, and that the items/tasks reflected a range of ability levels (as was the case with the speaking, writing and reading items). However, discussion with PTE ACADEMIC revealed that there was actually very little variation amongst the items considered by the panelists. However, we did run a Pearson r correlation between scale, level, Delta ability estimate and estimated total score. No significant correlations were identified other than the correlation between the Delta ability estimates and the PTE ACADEMIC estimated total score. Thus, we could not run a regression using the CAEL standard to predict PTE ACADEMIC total score because there was no variation in the judgment.

The panelists' perceptions of difficulty did not agree with those of PTE ACADEMIC (based on the ability estimates). Indeed, there was very limited agreement between the panelists and PTE ACADEMIC on rank order of difficulty (see Table 8 below), from low difficulty level (1) to high difficulty level (16).

Table 8. Relationship between CAEL and PTE ACADEMIC based on rank order of difficulty

CAEL Rank	PTE ACADEMIC Rank	CAEL Level	Estimated Total Points	PTE Delta (ability)	Item/Task
1.00	2.00	2.04	53	-.09588	**7**
2.00	15.00	2.11	68	.91194	**12**
3.00	6.00	2.16	57	.18049	**8**
4.00	7.00	2.21	57	.21721	**4**
5.00	13.00	2.23	66	.67063	**16**
6.00	4.00	2.27	54	.02461	**14**
7.00	11.00	2.27	65	.63043	**15**
8.00	12.00	2.42	66	.66335	**5**
9.00	10.00	2.48	63	.50048	**13**
10.00	16.00	2.50	69	.83798	**1**
11.00	5.00	2.50	55	.10432	**6**
12.00	14.00	2.50	66	.95763	**11**
13.00	1.00	2.59	45	-.40523	**2**
14.00	3.00	2.61	52	-.08574	**3**
15.00	8.00	2.65	59	.41442	**9**
16.00	9.00	2.73	62	.48190	**10**

For example, as the information presented in Table 8 indicates, according to the panelists item #7 was considered the easiest (see CAEL Level and Rank), whereas according to the delta statistic/total score provided by PTE ACADEMIC, it was considered the second easiest (see PTE ACADEMIC Rank). In this case it is not a serious discrepancy; however, in the case of item # 12, for example, the panelists considered it the second easiest, whereas it was ranked at number 15 (one of the most difficult items) according to PTE ACADEMIC analyses.

Although no significant correlations were found between total score and CAEL levels, Table 20 does provide a view of what the panelists considered the 2 range of 'just right', which in PTE ACADEMIC estimated total scores ranges from a low of 45 to a high of 69 and covers CAEL bands 60 and 70. This would suggest that the mid-point in the point spread from 45 to 69 might be an appropriate cut-off for the division between bands 60 and 70 or 57. Taking standard deviation into account, and also the consistent calibration of PTE ACADEMIC scores across the subtests, we would recommend (albeit with less confidence than with the other sub-test cut-score recommendations) that 60 be considered the cut-off for listening on the PTE ACADEMIC which reflects band 70 on CAEL.

3.6 SCORE REQUIREMENT RECOMMENDATIONS

Based on the results obtained above, the recommended cut points on the PTE Academic for entry into Canadian tertiary institutions should be as follows.

Table 9. PTE Academic Score Requirement Recommendations

TEST	READING	LISTENING	SPEAKING	WRITING	OVERALL
Estimated PTE Academic Total Score	60	60	60	60	60
CAEL Band Score	70	70	70	70	70

Caveats

As noted above, we have less confidence in the listening score recommendation because of the restricted range of difficulty and a restricted range of item types provided by PTE ACADEMIC for consideration by the panel. There is also the possibility that the listening construct may not be operationalized in a similar enough manner to afford precise comparison: the CAEL listening comprehension task involves multiple items based on a single, extended lecture (which is fully integrated with an introductory reading), whereas the PTE ACADEMIC uses multiple, stochastically independent lectures with item bundles. Further, the PTE ACADEMIC is nearly twice as long as the CAEL (PTE ACADEMIC= up to 58 minutes; CAEL = up to 25 minutes).

4. APPROACH 2: TESTRE-TEST

Subsequent to the analysis of the expert panelists, a test re-test study was conducted to analyze CAEL and PTE ACADEMIC test performance. This test re-test study drew a purposive, stratified random sample of 15 test takers who took both the PTE ACADEMIC and the CAEL within one week, but not in the same order, to avoid an order effect.

4.1 THE SAMPLE (N=15)

The sample consisted of 4 female and 11 male test takers, from 10 different countries of origin; ranging in age from 41 to 17 years. They had been in Canada from a few months to five years. All of the students were enrolled at the time of the study in English for Academic Purposes (EAP) courses. They had tested into these courses as basic, intermediate and advanced levels of proficiency in English. Five students were drawn from each level for the study, however, one completed only the CAEL before withdrawing. There were 2 graduate students; 13 undergraduates;8 Arabic speakers; 3 Chinese speakers; 2 Farsi speakers; 2 others; 8 majoring in Engineering (incl. architecture and industrial design); 3 Business; 3 in Social Sciences; 1 in science. Science.

4.2 RESULTS

The overall distribution of CAEL and PTE ACADEMIC scores is illustrated below in Figure 1.

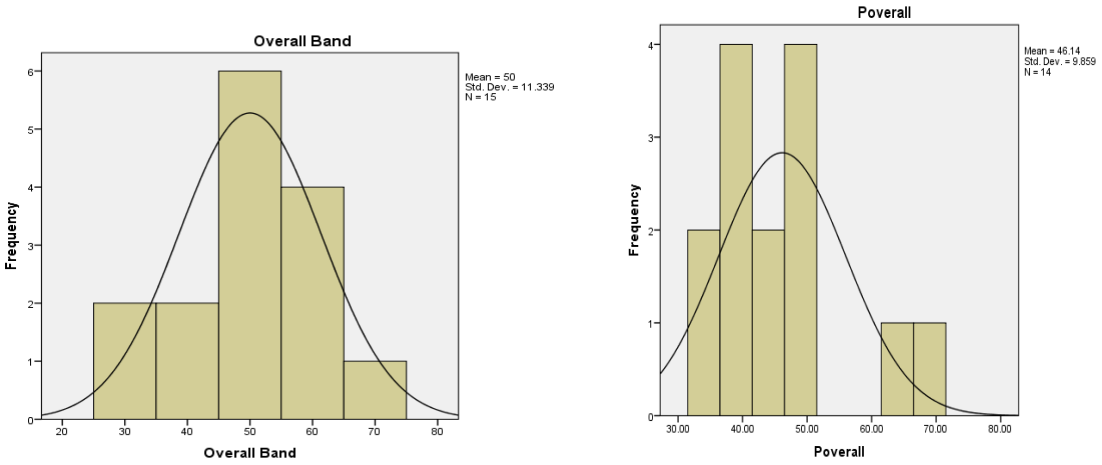

Figure 1. Overall CAEL Band distribution (on the left) and PTE ACADEMIC Score distribution (on the right)

Although the CAEL Bands distribute in a roughly normal curve, the PTE ACADEMIC scores(on the right) do not. This, however may be due to the small number of cases considered in the analysis.

Correlations between the sub-tests and overall scores between the two tests were significant, based on Pearson's r:

Overall, r=.58, significant, p<.05

Listening, r=.66, significant, p<.01

Reading, r=.63, significant, p<.05

Speaking, r=.65, significant, p<.05

Writing, r=.54, significant p<.05

Although the sample was arguably representative of the CAEL test taker population, the number of cases used for analysis was small (only 15 seats were available for the test comparison, and one of the test takers failed to submit their PTE ACADEMIC scores for analysis). Given the trends in the data, if the sample were larger we would expect the correlations to increase.

Paired samples t-test

Comparing means between Overall scores on PTE ACADEMIC and CAEL, using a paired samples t-test, there was no significant difference in performance between the two tests.

CAEL Scores as Indicators

Working with CAEL scores as indicators, we calculated the mean scores for PTE ACADEMIC with the following results:

Overall Score Comparisons

CAEL band 30-40 (EAP level basic) = PTE ACADEMIC overall 34.
CAEL band 50 (EAP intermediate level) = PTE ACADEMIC overall 40.

CAEL band 60 (EAP advanced level) = PTE ACADEMIC overall of 52.
CAEL band 70-90 (full-time university admission) = PTE ACADEMIC overall of 60.

The tester-test approach provided triangulating evidence to support the findings of the expert panelists.

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.

Cizek, G. (2001). *Setting performance standards.* NJ: Lawrence Erlbaum Associates.

Cizek, G. & Bunch, M. (2007).*Standard Setting*. London: Sage Publications.

de Jong, J., Li J. & Duvin, J. (2010). *Setting Requirements for Entry Level Nurses on PTE Academic.* Internal Report, Pearson Test of English Academic.

Fox, J. (2003). From products to process: An ecological approach to bias detection. *International Journal of Testing, 3*(1), 21-48.

Fox, J. (2004). Test decisions over time: Tracking validity. *Language Testing, 21*(4), 437-465.

Fox, J. (2009). Moderating top-down policy impact and supporting EAP curricular renewal: Exploring the potential of diagnostic assessment. *Journal of English for Academic Purposes*, 8, 26-42.

Hsieh, M. (2013).Comparing yes/no Angoff and Bookmark standard setting methods in the context of English assessment. *Language Assessment Quarterly, 10*(3), 331-350.

Jaeger, R. M. (1989).Certification of student competence. In R. L. Linn (Ed.), *Educational Measurement, 3rd edition*, pp. 485-513. Washington, DC: Macmillan.

Livingston, S. & Zeiky, M. (1982).*Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.

Pearson Test of English Academic (2010).*The Official Guide: PTE Academic*. Hong Kong: Pearson Longman Asia ELT.

Zeiky, M. (2001).*So much has changed: How the setting of cut scores has evolved since the 1980s.* In G. J. Cizek (Ed.), *Setting performance standards* (pp.19-51). Mahwah, NJ: Lawrence Erlbaum.

Zeiky, M. & Perie, M. (2004).*A primer on setting cut-scores on tests of educational achievement. P*rinceton, NJ: Educational Testing Service.