Linking the Common European Framework of Reference and the CaMLA English Placement Test

Technical Report

CONTACT INFORMATION

All correspondence and mailings should be addressed to:

CaMLA
Argus 1 Building
535 West William St., Suite 310
Ann Arbor, Michigan
48103-4978 USA

T: +1 866.696.3522
T: +1 734.615.9629
F: +1 734.763.0369

info@cambridgemichigan.org
www.CambridgeMichigan.org

TABLE OF CONTENTS

LIST OF TABLES

1. INTRODUCTION

This report presents the results of a project to link CaMLA English Placement Test (EPT) scores to the proficiency levels of the Common European Framework of Reference (CEFR, Council of Europe, 2001). Since its introduction, the CEFR has become widely used to interpret test scores. Test users and other stakeholders find CEFR levels useful for decision-making. It is hoped that, by linking CaMLA EPT scores to the CEFR, the test results will be more meaningful to test users.

1.1 The Common European Framework

The CEFR scales and their constituent descriptors were developed during a large research project (North, 2000; North & Schneider, 1998). They describe what learners can do with language at six main levels (A1, the lowest, to C2, the highest).

1.2 The CaMLA English Placement Test

The CaMLA English Placement Test (EPT) is a test for institutions (TFI) product that is designed to quickly and reliably place English as a Second Language (ESL) students into homogeneous ability levels. By using this exam, teachers and program administrators will be able to confidently place ESL students into appropriate levels and classes based on a CaMLA EPT score. It provides an accurate assessment of a test taker's general receptive language proficiency by measuring performance in the key areas of: listening comprehension, grammatical knowledge, vocabulary range, and reading comprehension.

The CaMLA EPT has three unique forms (D, E, and F), constructed so that they are parallel in content and difficulty. The exam contains 80 multiple choice items, and takes about 60 minutes to administer. It consists of two sections, listening and GVR (Grammar, Vocabulary, and Reading). The items are situated in a variety of language domains: educational, social, occupational, and personal. Each item type targets different language interactions and contexts, enabling the test takers to demonstrate a range of receptive language skills.

All parts of the examination are written following specified guidelines, and items are pretested to ensure that they function properly.

1.3 Standard Setting

Standard setting is defined as the decision-making process of classifying candidates into a number of levels or categories (Kane, 2001: 53). The "boundary between adjacent performance categories" (Kane et al., 1999: 344) is defined by a cut score. In other words, a cut score is "a point on a test's score scale used to determine whether a particular score is sufficient for some purpose" (Zieky et al., 2008: 1). For example, when determining whether candidates have passed or failed an exam, a cut score functions as the boundary between the pass and fail category.

During a standard-setting meeting, a panel of expert judges (commonly referred to as panelists) makes judgments on which examination providers will base their final cut score decisions. Under the guidance of one or more meeting facilitators, statistical information about test items and the distribution of scores are provided to help panelists with their judgment task. More than one round of judgments is usually organized to allow panelists to discuss their decisions, take into account the relevant statistical information, and revise their judgments. Once the standard-setting activities are complete, the meeting is evaluated in terms of three main categories (Council of Europe, 2009: ch. 7):

- Procedural validity: examining whether the procedures followed were practical and implemented properly, that feedback given to the judges was effective, and that documentation has been sufficiently compiled.
- Internal validity: addressing issues of accuracy and consistency of the standard setting results.
- External validation: collecting evidence from independent sources which support the outcome of the standard setting meeting.

2. METHODOLOGY

This section discusses the selection of the judges, the standard-setting method, the material prepared, and the tasks the panel engaged in during the meeting.

2.1 Selection of Judges

The panel consisted of seven participants that are experts in the field of language assessment. All are CaMLA staff. The judges were chosen to ensure that the panel represented a variety of different educational backgrounds and work experiences. The panelists' job

functions are varied and include assessment managers and developers, as well as the business and assessment directors. They have graduate qualifications in education, linguistics, and applied linguistics, and have a range of experience as language teachers.

As CaMLA employees, all of the judges had prior knowledge of the EPT and CEFR. However, to ensure that all judges were equally familiar with the CEFR levels, familiarization activities were included to calibrate the panel. It should be noted that one of the judges (J1) was unable to attend the second day, so cut score estimates for the GVR section were provided by six judges.

2.2 Standard-Setting Method

The standard-setting method is a modification of that proposed by Angoff (Angoff, 1971). The Angoff method is commonly phrased as the probability of an imaginary, borderline candidate answering an item correctly (Angoff, 1971: 515). However, this approach is considered to be challenging because judges may have difficulty in understanding the notion of probability.

Therefore, the task was modified: judges were asked to think of 100 borderline candidates at each CEFR level (e.g. A2 borderline candidates). These borderline candidates were defined as having just passed the border between one level and the higher adjacent level (e.g. between A1 and A2). For each exam item, the judges were asked to state how many of these 100 candidates would answer each item correctly (cf., Cizek & Bunch, 2007: 85). In principle, the panelists would have to work through the test four times, each time focusing on borderline learners at one specific CEFR boundary i.e. A2, B1, B2, and C1. As can be imagined, it would have been very time-consuming and tiring to make four separate judgments on each item. It is therefore attractive to combine levels, asking panelists to make judgments on two nonadjacent levels at a time (e.g., A2 and C2). This was the approach taken: panelists were asked to work through the EPT section looking first at only B1 and C1 borderline candidates. They then repeated that process for A2 and B2 borderline candidates. Panelists worked with nonadjacent levels in order to minimize the possibility of a judge's decision being influenced by a neighboring level's score.

This procedure was conducted separately for each section of the EPT. We believe that this approach ensured clarity of the judgment task and also helped to reduce the cognitive load the judges would have experienced had they been asked to make judgments for all four CEFR boundaries at once.

2.3 Material

In order to familiarize the panel with the CEFR, each panelist received "atomized" CEFR descriptors for each language skill: listening, reading, grammar, and vocabulary. The "atomization" of the descriptors into short statements, based on Kaftandjieva and Takala (2002), aimed to familiarize the panelists with all constituent statements of the descriptors. For each skill, judges had to read the individual statements and place them at the CEFR level they belong to (A1–C2). This is a challenging task because a number of the statements were quite short, without a detailed description of the context of language use. However, it encourages close reading of (and familiarization with) the details of the CEFR.

Upon completion of this task, the panelists were given a handout that had the descriptor statements ordered according to the level they belonged to. This handout was used in the subsequent tasks as a set of Performance Level Descriptions (PLD, see Cizek & Bunch, 2007: 44–47) according to which panelists should make their cut score decisions.

In order to help panelists obtain a better understanding of the difficulty of test items and how this relates to the judgment task, the training material asked them to rank a number of EPT listening and GVR items from easiest to most difficult

Finally, using the Angoff method (Section 2.2) the panelists were asked to estimate for each item (EPT form D) the number out of 100 candidates at the border between two CEFR levels that would answer the item correctly. The panelists entered their judgments into preprepared excel files, so that the data could be quickly analyzed and discussed. Once entered, each panelist's estimates were then added and divided by 100. This gave, for each panelist, the total number of items that would be answered correctly by test takers at each of the levels being defined (A1/A2, A2/B1, B1/B2, and B2/C1) i.e. each panelist's proposed cut scores.

2.4 Tasks During the Meeting

Both days of the standard setting meeting followed the same overall structure for linking the EPT to the

CEFR levels. The first day focused on the listening section, while the second day focused on the GVR section.

Aside from a brief discussion of the prereading and an introduction to the linking process on day one and an exit survey on day two, both days were identical, and began with the familiarization task for their respective sections (i.e., listening on day one, GVR on day two). The panelists worked individually on this task, assigning the CEFR level of the descriptors, and filling in an excel spreadsheet with their answers. Upon completion, the data were condensed into one excel file and analyzed, and the panelists were shown how many descriptors they placed at the correct level, how well their answers correlated with the true levels, and how their mean level compared with the true mean level. The panelists received a handout that listed the descriptors with the correct CEFR levels, and they discussed their answers and the actual level of the CEFR descriptors. The moderator invited the participants to explain the reasons for choosing a particular level when a descriptor had a median judgment that did not agree with the correct CEFR level, a range of judgments that was too wide, or too many judges with the incorrect level. The discussion moved on to the next descriptor after all of the panelists felt that they understood the correct CEFR level of a descriptor statement.

The CEFR familiarization task was followed by the training task. Here, the panelists ranked a sample of EPT items from easiest to most difficult. Once this was done, their rankings were compared to the rankings of the items based on difficulty. Particular attention was paid to the different clusters of item difficulties to make sure that that the panelists could correctly differentiate between items at various difficulty levels. This task helped the panelists to understand that it is difficult to predict item difficulty. It also helped to orient them to the item characteristics that might contribute to the challenge of an item. Finally, this task helped panelists to understand the link between item difficulty and CEFR level (i.e., the harder an item, the higher the CEFR level).

The final task on both days was the judgment task. This task comprised two identical judgment rounds. In each judgment round, the panelists first took the exam as a test taker, answering each item. They then expressed judgments on how many candidates out of 100 at the given CEFR level would get each item correct. Both of the judgment rounds were done in two sets. The first set asked the panelists to look at only B1 and C1 borderline candidates, while the second asked them to look at A2 and B2 borderline candidates. The panelists entered their judgments into an excel file and cut score estimates were calculated. The cut score recommendations were discussed between judgment rounds. Item statistics (facility values and biserial correlations for each item) were also provided between rounds, so that the panelists could see how their responses compared with actual data on item difficulty. For the second round, the same process was repeated, only now the panelists had more knowledge to apply to their judgments. At the conclusion of the second judgment round, the resulting cut scores were again presented and discussed. The overall recommended EPT cut scores were obtained by adding together the mean cut scores for both sections.

3. RESULTS OF THE CEFR FAMILIARIZATION ACTIVITIES

An important first step in the postmeeting analysis of the data is to establish the panelists' familiarity with the CEFR. If judgments are made by participants who do not have a good understanding of the CEFR levels, this will cast doubt on the validity of the recommended cut score because, when the panelists recommend cut scores, the underlying premise is that they fully and completely understand the CEFR levels and can rank the CEFR descriptors in the correct order and also assign random descriptors to the correct CEFR level. If they cannot assign descriptors consistently to the correct CEFR level, then they are likely to provide inconsistent judgments when setting cut scores.

The postmeeting analysis examined the number of descriptors that were placed at the correct level. Correlations of the panelists' level placements and the correct levels were also calculated. High correlations indicate that the panelists understand how the descriptors progress from lower to higher levels. However, correlations do not show how many descriptors were placed at the correct level, thus they should be consulted along with the number of correct level placements. Finally, each panelist's mean level was calculated to establish their tendency to put descriptors at lower or higher levels.

Tables 3.1, 3.2, 3.3, and 3.4 present information about the panelists' performance on the familiarization tasks. Similar analyses can be found in other relevant studies, such as Kaftandjieva and Takala (2002), the Ministry of Education in Catalonia, Spain (Generalitat de Catalunya, 2006), and the Trinity College London CEFR Project Report (Papageorgiou, 2007).

Table 3.1 shows the results of the listening descriptor familiarization task. Table 3.2 shows the results of the reading descriptor familiarization task. Tables 3.3 and 3.4 show the results of the vocabulary and grammar descriptor familiarization tasks respectively. The first row of each table shows the number of descriptors placed at the correct level by each panelists (J1–J6 or J7). The second row presents the Spearman correlation between each panelist's descriptor placement and the correct CEFR levels. The correlation was calculated by comparing a panelist's level placements with the correct ones. The final row in each table presents the mean level of all level choices by each panelist. This can be compared to the mean CEFR level for the descriptors (indicated in the parentheses of the caption for each table). If a panelist's mean is higher than the CEFR mean, this indicates a tendency to assign descriptors to a higher level than the correct one. If a panelist's mean is lower than the CEFR mean, this indicates a tendency to assign descriptors to a lower level than the correct one. These tendencies could result in inconsistent judgment when setting cut scores, so it is important to identify and correct these issues if they exist.

Table 3.1 Listening Familiarization Task Results (71 Descriptors, 3.56 mean level)

	J1	J2	J3	J4	J5	J6	J7
Correct	42	38	33	26	38	37	39
Spearman	0.92	0.85	0.87	0.86	0.92	0.86	0.90
Mean	3.79	3.79	3.39	3.70	3.45	3.45	3.35

Table 3.2: Reading Familiarization Task Results (56 descriptors, 3.25 mean level)

	J2	J3	J4	J5	J6	J7
Correct	34	33	37	43	33	38
Spearman	0.88	0.93	0.91	0.95	0.90	0.93
Mean	3.57	3.68	3.38	3.34	3.45	3.29

Table 3.3: Vocabulary Familiarization Task Results (25 descriptors, 3.40 mean level)

	J2	J3	J4	J5	J6	J7
Correct	16	20	19	21	14	18
Spearman	0.92	0.97	0.95	0.97	0.95	0.93
Mean	3.20	3.52	3.24	3.24	2.96	3.08

Table 3.4: Grammar Familiarization Task Results (17 descriptors, 3.41 mean level)

	J2	J3	J4	J5	J6	J7
Correct	15	6	8	12	11	10
Spearman	0.97	0.94	0.93	0.93	0.90	0.91
Mean	3.41	4.12	3.59	3.59	3.29	3.47

As can be seen from the tables, the correlations[1] are very high, which suggests that the panelists had a good understanding of how language proficiency progresses from lower to higher CEFR levels. However, the relatively low number of correct descriptor placements suggests that they had difficulty with placing the descriptors at the exact levels. Comparing the panelists' mean assigned CEFR level with the true mean CEFR level for each task we can see that the majority of the panelists differed from the true CEFR level in the average CEFR level that they assigned. Larger mean values suggest that the panelist tends to be more lenient, while smaller mean values indicate that a panelist is more severe. The tables above show that the severity (or leniency) of the panelists depends largely on both the panelist and the language skill being analyzed. The implications of this are important for setting cut scores, since a panelist may apply their leniency or severity from the familiarization tasks to the judgment task, possibly skewing the results. In order to avoid this issue, leniency and severity of the panelists was pointed out during the meeting, and the descriptor statements were discussed to ensure that the judges all had a good understanding of the correct CEFR levels. It is also important to note that these findings are typical, and that they are similar to the results obtained by Papageorgiou (2010) in his standard-setting study for the Michigan English Test (MET).

So far the analysis has focused on the individual panelists' understanding of the CEFR levels. However, cut scores are not based on an individual panelist's decisions, but on the judgments of the entire panel. Because of this, further analysis was performed to establish the consistency of the panel. The results of which are presented in Table 3.5.

Table 3.5: Agreement and Consistency of the Group

	Grammar	Listening	Reading	Vocabulary
Alpha	0.979	0.977	0.978	0.983
ICC	0.973	0.975	0.977	0.981
W	0.862	0.851	0.872	0.888

The table presents three different measures of agreement and consistency. Cronbach's Alpha, is an internal consistency index usually used in item-based tests. It is reported here to indicate "the consistency of the reliability of ratings in terms of rater consistency" (Generalitat de Catalunya, 2006: 62). The intraclass correlation coefficient, ICC (Generalitat de Catalunya, 2006: 62), is calculated in order to demonstrate how the average rater agreed with all others. The ICC two-way mixed model was used and average measures for exact agreement are reported. Kendall's W was also used to investigate rater agreement in similar contexts (Generalitat de Catalunya, 2006: 112; Kaftandjieva &Takala, 2002). Kendall's W can be interpreted as a coefficient of agreement among raters. The coefficient ranges from 0 to 1, with 1 indicating complete inter-rater agreement, and 0 indicating complete disagreement among judges. As can be seen from Table 3.5, these measures are high, which suggests agreement and consistency among the judges.

The analysis presented in this section suggests that the panel had a good overall understanding of how language ability progresses from lower to higher levels in the CEFR scales. However, as mentioned before, individual panelists had difficulty with placing the descriptors at the correct level, and were often more severe or lenient than the mean level. While occasionally placing a descriptor at an adjacent level is not unreasonable, if the panelists systematically misunderstood some of the differences between these levels they may suggest cut scores whose validity should be questioned. As stated before, in order to avoid this issue, leniency and severity of the judgments was pointed out during the meeting, and the descriptor statements were discussed to ensure that the panelists all had a good understanding of the correct CEFR levels. The panelists felt that the discussion of the familiarization task was very useful. They believed that it helped to clarify the differences between adjacent levels, as well as allowed them to look at the descriptors thoroughly.

4. CUT SCORE RESULTS AND VALIDITY EVIDENCE

4.1 Cut Score Validation

Standard setting validation includes three main areas of validation: Procedural, Internal, and External. Several arguments supporting procedural validity have already been presented: the methodology of this standard setting study was based on recommendations in the relevant literature; and it has been documented

1 All correlations are significant at the 0.01 level (two-tailed)

in detail in the previous sections. This section will focus on the cut scores produced by the linking meeting, and will provide evidence of internal and external validation. It will also discuss the panels' feedback as part of the procedural validation.

4.2 Initial Cut Score Estimates

Tables 4.1 and 4.2 show the cut scores recommended by each of the panelists in the listening and GVR section, respectively. Both tables show the four cut scores recommended for both rounds of the judgment procedure. Note that the cut scores are expressed as the number of items correct in each EPT section (out of 25 for Listening, 55 for GVR). The

Table 4.1: Cut Score Judgments for EPT Listening Section

	Round 1				Round 2			
Judge ID	**A1/A2**	**A2/B1**	**B1/B2**	**B2/C1**	**A1/A2**	**A2/B1**	**B1/B2**	**B2/C1**
J1	6.80	9.70	17.44	21.34	5.40	8.75	16.85	20.75
J2	12.13	14.41	21.15	22.98	12.02	15.12	18.49	22.03
J3	11.08	16.62	21.31	23.10	12.61	17.09	19.95	22.09
J4	5.35	8.90	16.10	19.00	12.35	14.25	20.35	21.20
J5	9.82	11.59	18.09	19.28	9.41	12.27	18.29	20.93
J6	9.80	11.15	19.60	21.45	9.75	11.10	20.95	21.35
J7	9.99	15.20	18.26	20.17	14.65	15.54	17.27	20.09
Mean	9.28	12.51	18.85	21.05	10.88	13.45	18.88	21.21
Median	9.82	11.59	18.26	21.34	12.02	14.25	18.49	21.20
SD	2.38	2.93	1.93	1.65	3.00	2.89	1.57	0.71
Min	5.35	8.90	16.10	19.00	5.40	8.75	16.85	20.09
Max	12.13	16.62	21.31	23.10	14.65	17.09	20.95	22.09

Table 4.2: Cut Score Judgments for EPT GVR Section

	Round 1				Round 2			
Judge ID	**A1/A2**	**A2/B1**	**B1/B2**	**B2/C1**	**A1/A2**	**A2/B1**	**B1/B2**	**B2/C1**
J2	20.60	37.11	44.53	50.39	21.92	27.61	35.58	42.92
J3	23.79	29.17	40.70	47.66	21.11	24.12	33.69	39.63
J4	18.50	26.05	42.95	46.00	15.82	19.95	32.75	39.80
J5	17.75	24.94	38.43	44.06	16.56	23.16	32.03	41.18
J6	16.85	21.80	46.00	49.00	16.95	21.65	45.20	48.00
J7	24.03	29.45	40.78	46.69	24.93	27.75	31.91	34.75
Mean	20.25	28.09	42.23	47.30	19.55	24.04	35.19	41.05
Median	19.55	27.61	41.87	47.18	19.03	23.64	33.22	40.49
SD	3.09	5.25	2.79	2.24	3.65	3.15	5.09	4.36
Min	16.85	21.80	38.43	44.06	15.82	19.95	31.91	34.75
Max	24.03	37.11	46.00	50.39	24.93	27.75	45.20	48.00

Table 4.3: Initial Cut Score Estimates

	A1/A2	A2/B1	B1/B2	B2/C1
Raw Cut Score	30.43	37.49	54.07	62.25
Rounded Cut Score	31	38	55	63

descriptive statistics for these judgments are summarized in the bottom half of the tables. The Round 2 cut scores were made after the panelists compared their cut scores to those recommended by other panel members as well as the item analysis data.

Table 4.3 presents the initial cut score estimates for the EPT as a whole. These overall cut scores were obtained by adding together the Round 2 mean cut scores of the listening and GVR sections. These numbers were rounded up in order to minimize any false positive classifications (Cizek & Bunch, 2007: 25). For example, the original B1 cut score was 37.49 (See Table 4.3), and rounding to the nearest whole number would result in a cut score of 37. However, test takers who answer 37 items correctly do not demonstrate the ability depicted by a cut score of 37.49. Therefore, since the number of correctly answered items can either be 37 or 38, 38 was chosen as the cut score.

4.3 Method Consistency Analysis and Finalization of Cut Scores

This analysis examines method consistency by estimating the standard error of judgment (SEj). This is calculated by dividing the standard deviation of the judgments with the square root of the number of judges (Norcini et al., 1987). According to Cohen et al. (1999), SEj should be less than or equal to half of the standard error of measurement (SEM) of the test. EPT Form F (an equivalent test form to the one used in this linking study) has a SEM of 1.913 for the listening section, and a SEM of 2.894 for the GVR section. Thus, in order to argue for the validity of the cut score, the SEj should be less than or equal to 0.957 and 1.447 for their respective sections. However, as can be seen from Table 4.4, this is not the case for many of the cut scores.

Table 4.4: Comparison of SEj Before and After Excluding Extreme Ratings

Section	Cut Score	SEj (all ratings)	SEj (exclude extremes)
Listening	A1/A2	1.135	0.738
	A2/B1	1.092	0.834
	B1/B2	0.594	0.594
	B2/C1	0.268	0.268
GVR	A1/A2	1.490	1.152
	A2/B1	1.288	1.288
	B1/B2	2.076	0.617
		1.780	1.243

In order to reduce the SEj, extreme ratings (too low or too high cut scores) were excluded from the calculation of the cut score (see also calculation of the trimmed mean in Zieky et al., 2008: 38–39). As Table 4.4 shows, removing the extreme scores has reduced the SEj to the desired level. Each of the five cut scores with a high SEj had only one extreme score removed. The resulting mean cut scores can be found in Tables 4.6 and 4.7 (on the following page).

Table 4.5 shows the recommended overall cut scores after the extreme ratings are excluded. These cut scores are very similar to the initial ones (see 4.2, above). However, by excluding the extreme judgments, and lowering the SEj we expect that the cut scores are more dependable. Therefore, these scores will be used as the final cut scores for the EPT.

Table 4.5: Cut Score Estimates (w/o extreme)

	A1/A2	A2/B1	B1/B2	B2/C1
Raw Cut Score	30.27	38.27	52.07	60.86
Rounded Cut Score	31	39	53	61

Table 4.6: Cut Score Judgments for EPT Listening Section (w/o extreme)

	Round 1				Round 2			
Judge ID	**A1/A2**	**A2/B1**	**B1/B2**	**B2/C1**	**A1/A2**	**A2/B1**	**B1/B2**	**B2/C1**
J1	6.80	9.70	17.44	21.34	(excl.)	(excl.)	16.85	20.75
J2	12.13	14.41	21.15	22.98	12.02	15.12	18.49	22.03
J3	11.08	16.62	21.31	23.10	12.61	17.09	19.95	22.09
J4	5.35	8.90	16.10	19.00	12.35	14.25	20.35	21.20
J5	9.82	11.59	18.09	19.28	9.41	12.27	18.29	20.93
J6	9.80	11.15	19.60	21.45	9.75	11.10	20.95	21.35
J7	9.99	15.20	18.26	20.17	14.65	15.54	17.27	20.09
Mean	9.28	12.51	18.85	21.05	11.80	14.23	18.88	21.21
Median	9.82	11.59	18.26	21.34	12.19	14.69	18.49	21.20
SD	2.38	2.93	1.93	1.65	1.95	2.21	1.57	0.71
Min	5.35	8.90	16.10	19.00	9.41	11.10	16.85	20.09
Max	12.13	16.62	21.31	23.10	14.65	17.09	20.95	22.09

Table 4.7: Cut Score Judgments for EPT GVR Section (w/o extreme)

	Round 1				Round 2			
Judge ID	**A1/A2**	**A2/B1**	**B1/B2**	**B2/C1**	**A1/A2**	**A2/B1**	**B1/B2**	**B2/C1**
J2	20.60	37.11	44.53	50.39	21.92	27.61	35.58	42.92
J3	23.79	29.17	40.70	47.66	21.11	24.12	33.69	39.63
J4	18.50	26.05	42.95	46.00	15.82	19.95	32.75	39.80
J5	17.75	24.94	38.43	44.06	16.56	23.16	32.03	41.18
J6	16.85	21.80	46.00	49.00	16.95	21.65	(excl.)	(excl.)
J7	24.03	29.45	40.78	46.69	(excl.)	27.75	31.91	34.75
Mean	20.25	28.09	42.23	47.30	18.47	24.04	33.19	39.66
Median	19.55	27.61	41.87	47.18	16.95	23.64	32.75	39.80
SD	3.09	5.25	2.79	2.24	2.82	3.15	1.51	3.04
Min	16.85	21.80	38.43	44.06	15.82	19.95	31.91	34.75
Max	24.03	37.11	46.00	50.39	21.92	27.75	35.58	42.92

4.4 Decision Consistency Analysis

After excluding the extreme ratings, we examined decision consistency (Cizek & Bunch, 2007: 307) using the following equation:

$$|Z|=(CX-M-0.5)/SX$$

where Cx is the cut score for the test, M is the observed test mean, and Sx is the standard deviation (SD) of the observed test scores. This equation, combined with the reliability of the test, was then used to obtain the estimates of agreement coefficient (p0) and kappa (k) from two tables in Subkoviak (1988), reproduced in Cizek and Bunch (2007: 310–311). Table 4.8 presents the results of this analysis for each cut level.

Table 4.8: Agreement Coefficient (p_0) and Kappa (k) for the EPT Cut Scores

Cut Level	p_0	k
A1/A2	0.95	0.63
A2/B1	0.91	0.68
B1/B2	0.86	0.71
B2/C1	0.86	0.71

When interpreting these statistics, it should be noted that the agreement coefficient (p0) is a measure of overall consistency, and kappa (k) is a measure of the test's contribution to that consistency (Subkoviak, 1988: 54). In Subkoviak's table, the maximum values are 0.98 for p_0 and 0.71 for k. It could therefore be argued that the EPT recommended cut scores for the CEFR levels demonstrate satisfactory decision consistency.

4.5 Intra-Judge and Inter-Judge Consistency

In order to examine intra-judge consistency, the panelists' mean ratings for each item were correlated with the items' empirical difficulties. These correlations[2] are shown in Table 4.9. Bearing in mind that expert judges cannot easily estimate the difficulty of test items (cf. Alderson, 1993), correlations above 0.30 are considered satisfactory. All the panelists exceeded this minimum expectation, showing moderate to excellent correlations between their average item judgments and the items' empirical difficulty.

Inter-judge consistency in the judgment rounds was examined using three measures of rater consistency: Cronbach's alpha, Intra-class correlations, and Kendall's W (see section 3 for explanations of these measures). Table 4.10 shows that all of the measures are high, which provides additional cut score validity evidence.

Table 4.10: Agreement and Consistency of the Group

	Listening	GVR
Alpha	0.957	0.962
ICC	0.883	0.948
W	0.807	0.867

4.6 External Validation

The analyses presented in sections 3 and 4 have offered evidence in terms of internal and procedural validity. In this section, evidence for external validity will be presented. For external validation, the manual prepared by the Council of Europe to support CEFR standard-setting studies (Council of Europe, 2009: Ch.7) suggests collecting evidence from independent sources which support the outcome of the standard setting meeting. One example would be to analyze test data for students who took both the EPT and another CEFR linked exam. Another would be to use a second standard setting method. Unfortunately test takers' scores on both the EPT and another CEFR linked exam are not available. Also, having the panelists make judgments using a second method would increase their cognitive load, and could potentially cause fatigue which, in turn, could affect the judgments. As a result,

Table 4.9: Correlation (Spearman) Between Average Item Judgment and Empirical Difficulty for Judges

Judge	J1	J2	J3	J4	J5	J6	J7
Listening	0.630	0.939	0.973	0.882	0.935	0.765	0.987
GVR	NA	0.916	0.936	0.987	0.994	0.692	0.997

2 All correlations are significant at the 0.01 level (two-tailed)

Table 4.11: Classification of Form F Candidates (N=312) into CEFR Levels Based on the Recommended Cut Scores

CEFR Level	A1	A2	B1	B2	C1
Percent of Test Takers	10.90	7.05	20.83	13.78	47.44

external validation was attempted by exploring the reasonableness of the cut scores.

This was investigated by looking at how these cut scores would group 312 candidates that took the parallel EPT Form F into levels. It should be noted that all of these candidates were from the same school. The results of this are presented in Table 4.11.

As can be seen from the table, the test takers are distributed throughout the five CEFR levels, however there is a very large percentage categorized as C1. This result was not entirely unexpected, since the students attending the school that provided this data tend to be more proficient. The judges felt that the recommended cut scores yielded a reasonable classification of the test takers.

As an additional piece of external validity evidence, the recommended CEFR cut scores were compared to the guidance cut scores offered in the EPT Administration Manual. These guidance cut scores are based on a retired form of the EPT (Form A) and were prepared and used by the University of Michigan's Intensive English Program. While the preexisting cut scores do not link to the CEFR and were only ever intended as a guide, they still provide a good means of comparison. Table 4.12 presents both scales for comparison.

Table 4.12: Comparison of EPT Cut Scores

CEFR Level Scores		Skill Level Scores	
Level	**Score Range**	**Level**	**Score Range**
A1	0–30	Beginner	0–26
A2	31–38	Beginner (High)	27–40
B1	39–52	Intermediate (Low)	41–50
B2	53–60	Intermediate	51–61
C1	61–80	Advanced (Low)	62–68
		Advanced	69–80

The table shows that the CEFR cut scores are very similar to the skill level cut scores. Most of the CEFR levels match up with the corresponding skill level, with the exception of Advanced, which is merged together with Advanced (Low) into the C1 level.

4.7 Panel Feedback

Section 4 of this report has provided a variety of sources to support the validity of the recommended cut score. This section presents the results of the anonymous feedback questionnaire completed by the judges at the conclusion of the standard setting meeting. The questionnaire collected data using a four-point Likert scale (1 – *Strongly Disagree* to 4 – *Strongly Agree*). The results are summarized in Table 4.13.

Table 4.13: Results of Exit Survey

No.	Question	1	2	3	4
1	The prereading helped me to understand the background to the CEFR?	-	1	3	3
2	The EPT sample test helped me to understand the structure and level of the test.	-	1	3	3
3	The introductory presentation helped me to understand the linking process.	-	-	2	5
4	The discussion of the prereading answered my questions.	-	-	3	4
5	The familiarization tasks helped me to understand the CEFR levels.	-	-	3	4
6	The training items helped me to understand the judgment process.	-	1	-	6
7	I understood the instructions for each judgment round.	-	-	-	7
8	I understood the discussion of item statistics.	-	-	5	2
9	I had enough time to complete my individual tasks.	-	1	1	5
10	I had enough time to participate in the discussions.	-	-	-	7
11	I am confident in the decisions I have made.	-	-	6	1

The table shows that the majority of the ratings were positive, and that the panelists understood all of the tasks performed during the meeting. All of the panelists felt confident in the decisions made on the cut scores, and felt that the familiarization task helped with their understanding of the CEFR levels.

While most of the responses were positive, there were a few negative ones. Two of the panelists did not feel that the prereading and sample tests were helpful. However, it appears that this was not an issue in the standard setting process; once the information was discussed in more detail during the meeting all the panelists agreed that the activities (questions 4 and 5) helped them to understand the reading and the CEFR. Also, since all of the panelists were selected from within CaMLA, they already had an understanding of the structure of the EPT and the background to the CEFR.

One panelist felt that the training task did not help him/her to understand the judgment process. While this is not an issue, since all of the panelists understood the judgment round, in future standard setting meetings it may be useful to modify the training task. Only one panelist felt that they did not have enough time to complete their tasks. This is contrasted by the five panelists who felt strongly that they had enough time to complete the task. Although the panelist claimed to not have enough time, it does not appear to have negatively affected their understanding of the tasks, or their confidence in the decisions they made. Overall, since the results of the survey are mostly positive, this questionnaire helps to provide further validity evidence for the cut scores.

5. CONCLUSION

This technical report has presented the setting of CEFR cut scores for the CaMLA EPT. This report has summarized the methodology used in obtaining the cut scores, as well as provided validity evidence supporting these scores. The final recommended CEFR cut scores for the CaMLA EPT are presented in Table 5.1.

Table 5.1: Final EPT Cut Scores

	A1/A2	A2/B1	B1/B2	B2/C1
Cut Scores	31	39	53	61

6. REFERENCES

Alderson, J.C. (1993). Judgments in language testing. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research* (pp. 46–57). Alexandria, VA: TESOL.

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R.L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington: American Council on Education.

Cizek, G. J., & Bunch, M. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. London: Sage Publications.

Cohen, A. S., Kane, M. T., & Crooks, T. J. (1999). A generalized examinee-centered method for setting standards on achievement tests. *Applied Measurement in Education*, *12*(4), 343–366.

Council of Europe(2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.

Council of Europe (2009). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment. A Manual*. Retrieved from http://www.coe.int/t/dg4/linguistic/Source/ManualRevision-proofread-FINAL_en.pdf

Generalitat de Catalunya (2006). *Proficiency scales: The Common European Framework of Reference for Languages in the Escoles Oficials d'Idiomes in Catalunya.* Madrid: Cambridge University Press.

Kaftandjieva, F., & Takala, S. (2002). Council of Europe scales of language proficiency: A validation study. In J. C. Alderson (Ed.), *Common Europen Framework of Reference for Languages: Learning, teaching, assessment. Case studies*. (pp. 106–129). Strasbourg: Council of Europe.

Kane, M. (2001). So much remains the same: conception and status of validation in setting standards. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 53–88). Mahwah, N.J.: Lawrence Erlbaum Associates.

Kane, M., Crooks, T., & Cohen, A. S. (1999). Validating measures of performance. *Educational measurement: Issues and Practice*, *18*(2), 5–17.

Norcini, J. J., Lipner, R. S., Langdon, L. O., & Strecker, C. A. (1987). A comparison of three variations on a standard-setting method. *Journal of Educational Measurement*, *24*(1), 56–64.

North, B. (2000). *The development of a common framework scale of language proficiency*. New York: Peter Lang.

North, B. & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing*, *15*(2), 217–262.

Papageorgiou, S. (2007) *Relating the Trinity College London GESE and ISE examinations to the Common European Framework of Reference: Final Project Report, February 2007*, Trinity College London. Retrieved from http://www.trinitycollege.co.uk/resource/?id=2261

Papageorgiou, S. (2010) *Setting cut scores on the Common European Framework of Reference for the Michigan English Test*, CaMLA Technical Report, CaMLA. Retrieved from http://www.cambridgemichigan.org/sites/default/files/resources/MET_StandardSetting.pdf

Subkoviak, M. J. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement*, *25*(1), 47–55.

Zieky, M. J., Perie, M., & Livingston, S. A. (2008). *Cutscores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.