CaMLA
Speaking Test

Linking the Common European Framework of Reference and the CaMLA Speaking 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
CambridgeMichigan.org

CAMBRIDGE ENGLISH Language Assessment Part of the University of Cambridge
UNIVERSITY OF MICHIGAN

12/2015

TABLE OF CONTENTS

LIST OF TABLES

1. INTRODUCTION

1.1 OVERVIEW

This report summarizes the results of a standard-setting study conducted in October 2015. The purpose of this study was to link scores on the CaMLA Speaking Test to the proficiency levels of the Common European Framework of Reference. This study utilized the Council of Europe's (2009) manual supporting standard setting and Tannenbaum and Cho's (2014) article on critical factors to consider in standard-setting studies as guidelines for the study. This report documents the standard-setting study and provides validity evidence to support its quality.

1.2 COMMON EUROPEAN FRAMEWORK OF REFERENCE

The Common European Framework of Reference (CEFR) provides a common basis for the elaboration of language syllabuses, curricula, examinations, and textbooks (Council of Europe, 2001, p. 1). The framework comprehensively describes "what language learners have to learn to do in order to use a language for communication and what knowledge and skills they have to develop so as to be able to act effectively" (Council of Europe 2001, p. 1). The CEFR defines six main proficiency levels: A1 and A2 (basic users), B1 and B2 (independent users), and C1 and C2 (proficient users). The CEFR is widely used by test developers and other stakeholders to interpret test scores and make decisions, so linking the CaMLA Speaking Test to the CEFR will assist the users of the test in interpreting its results.

1.3 STANDARD SETTING

Standard setting can be defined as the process of identifying minimum test scores that separate one level of performance from another (Tannenbaum, 2011). These minimum test scores are generally referred to as cut scores, and are defined as the points on a score scale that act as boundaries between adjacent performance levels (Cohen, Kane, & Crooks, 1999). An example of a cut score would be the lowest passing score on a test. Anyone scoring at or above this value would be classified as having passed the test, while anyone scoring below this value would be classified as having failed the test.

A key aspect of the standard-setting process is the standard-setting meeting. During this meeting, a panel of experts typically goes through the test that is the focus of the study and makes judgments that are used to inform cut score recommendations for a particular population of test takers. Panelists are guided through the process of determining cut scores by the meeting facilitators. The first stage of the standard-setting meeting is known as familiarization. The purpose of this stage is to ensure that the panelists understand the standards and/or performance descriptors to which the test is being linked (e.g., the CEFR, as was the case for this study). The second stage of the standard-setting meeting is known as training. It provides panelists with the opportunity to practice making judgments, and ensures that they understand the procedure prior to making their cut score recommendations. The final stage of the standard-setting meeting is known as judgment. This stage is when the panelists make their individual cut score recommendations. There are often two or more rounds of judgment, which allow the panelists to discuss their individual cut score decisions and, if necessary, adjust their decisions. The final product of the standard-setting meeting is a recommended cut score (or scores) that links scores on the test to the standards or performance descriptors in question.

Once the standard-setting meeting has concluded, both the meeting itself and the resulting cut scores are examined for procedural, internal, and external validity (Council of Europe, 2009, Ch. 7; Tannenbaum & Cho, 2014). Procedural validity evidence shows that the study plan was implemented as intended, while internal validity evidence shows that the judgments were consistent (Tannenbaum & Cho, 2014). External validity evidence refers to any independent evidence that supports the outcomes of the current study (Council of Europe, 2009, Ch. 7).

1.4 CAMLA SPEAKING TEST

The CaMLA Speaking Test is an assessment of spoken language proficiency developed by Cambridge Michigan Language Assessments (CaMLA). It is designed to measure the speaking proficiency of English language learners from the upper beginner (A2) to lower advanced (C1) levels of the CEFR. The test evaluates a test taker's ability to produce comprehensible speech in response to a range of tasks and topics. Speaking performances are scored locally by the speaking examiner using the rating scale and evaluation criteria provided by CaMLA. Speaking examiners are also provided with training materials and instructions on how to assess test-taker performances, which must be successfully completed before they are certified to administer the test by their local testing coordinator.

The CaMLA Speaking Test is a structured, one-on-one interaction between the examiner and the test taker. The test lasts between 6 and 10 minutes and consists of five distinct tasks. These tasks are specifically designed to elicit language representative of the target ability levels. Task 1 asks the test taker to describe a picture, Task 2 asks the test taker to talk about a personal experience related to this picture, Task 3 asks the test taker to give his or her opinion and provide support on a topic related to Task 2, Task 4 asks the test taker to explain the advantages and disadvantages of a specific situation, and Task 5 asks the test taker to give an opinion on a specific issue and convince the examiner to agree with him or her. Each task receives a score from 1 to 5, and the final score for the test ranges from 5 to 25.

2. METHODOLOGY

2.1 PANELISTS

The panel of experts that make judgments on the location of the cut scores is one of the most important features of any standard-setting study. It is essential to include participants who have good knowledge of the examination in question, the test-taking population, and the performance level descriptors (Mills, Melican, & Ahluwalia, 1991; Papageorgiou, 2010). A total of twelve participants were selected for this study; seven from within CaMLA and five from outside of CaMLA. All of the panelists had experience (ranging from 1 to 22 years) as both ESL/EFL teachers and speaking examiners. As a group, they had an average of 8.42 years of ESL/EFL teaching experience and an average of 5.42 years of speaking examiner experience. The panelists also had a wide variety of other language testing experience, including experience in test development, test administration, and test scoring. The panelists' experience with and understanding of the CEFR prior to the standard setting study was varied, so the familiarization activities were particularly important. Overall, the panelists selected for this study provided a diverse representation of experienced professionals from the field of ESL/EFL.

2.2 STANDARD-SETTING METHOD

A variety of standard-setting methods exist in the field of educational measurement, with different methods selected for different types of tests. This standard-setting study utilized the bookmark method. The bookmark method is a procedure for establishing cut scores that was developed in 1996 in order to address perceived limitations (i.e., too cognitively challenging) of other standard-setting methods (Cizek, G. J., Bunch, M. B., & Koons, H., 2004; Mitzel, Lewis, Patz, & Green, 2001). This procedure is centered on the use of an ordered item booklet, which is the primary tool used to facilitate the panelists' cut score judgments. The ordered item booklet consists of test items listed in order of increasing difficulty, from the easiest item to the most difficult, and panelists make their judgments by going through the booklet and placing a 'bookmark' at the location where they believe the cut score is located (i.e., the point where one level of test taker is separated from the next). A key feature of the bookmark procedure is that it can be applied to both dichotomous (e.g., multiple-choice) and polytomous (e.g., constructed response) item responses (Council of Europe, 2009, Ch. 6). This is useful because it provides a means of setting cut scores for writing and speaking performances.

For this study a modification of the bookmark method was applied to the CaMLA Speaking Test in order to make three cut score judgments (A2/B1, B1/B2, and B2/C1). This method was selected because it allowed panelists to make multiple cut score judgments relatively quickly, and because it could be applied to constructed response items. The first step in applying the bookmark method was the creation of an ordered item booklet. It should be noted that because the speaking performances were audio recordings, the ordered item booklet was actually a digital folder of audio files, not a physical booklet. In practice, the digital folder is used in the same way as the ordered item booklet, so for the sake of simplicity, this folder will be referred to as an ordered item booklet throughout this report. Audio clips from recorded CaMLA Speaking Test performances for each possible score point, ordered from lowest (5) to highest (25), were selected to create the ordered item booklet. Each speaking performance included in the ordered item booklet was scored by at least two certified raters who worked to build consensus scores for each performance. Because of the time constraints of the standard-setting meeting, it was impractical to have the panelists listen to the entirety of each speaking performance. Instead, the raters selected audio clips of the task most representative of the total score awarded for each speaking performance. During this selection process, the raters listened to the performance very carefully and came to a consensus on the task to select for inclusion in the ordered item booklet. To make their cut score judgments, the panelists listened to the ordered audio clips and placed their bookmarks at the first performance that they felt could have been produced by a just-qualified B1-, B2-, or C1-level candidate.

2.3 MEETING PROCEDURES

This section provides an outline of the standard-setting meeting and summarizes the various activities performed during the meeting. The standard-setting meeting took place over the course of one nine-hour day in October 2015 and was conducted by two meeting facilitators who were experienced with the CEFR, speaking assessment, and standard setting.

Prior to the standard setting meeting, the panelists were required to complete pre-study activities to begin familiarizing (or, as was the case for many panelists, re-familiarizing) themselves with the CaMLA Speaking Test and the CEFR. In addition to a brief background questionnaire, the panelists also completed a pre-study CEFR quiz to assess their understanding of the CEFR prior to the standard-setting meeting. Both the background questionnaire and the pre-study CEFR quiz were administered through the online survey software tool, Qualtrics. This quiz required panelists to assign CEFR levels to 18 speaking descriptors selected from six scales related to speaking. Once the quiz was completed, the panelists were then asked to familiarize themselves with the CaMLA Speaking Test by reading information on the CaMLA website and by watching a sample test performance. They were also asked to familiarize themselves with the CEFR by reading an article on the CEFR by Morrow (2004) and by reviewing the CEFR self-assessment grid (Council of Europe, 2001, p. 26–27). Finally, the panelists were asked to read through the global scale (Council of Europe, 2001, p. 24), self-assessment grid (Council of Europe, 2001, p. 26–27), and qualitative aspects of spoken language use table (Council of Europe, 2001, p. 28–29) and describe their initial impressions of the characteristics of a just-qualified B1-, B2-, and C1-level candidate. They were asked to bring their "just-qualified" descriptions with them on the morning of the meeting.

The standard-setting meeting began with a brief introduction to the standard-setting procedure and the goals of the study. The pre-study materials were then reviewed and discussed to address any of the panelists' questions. The discussion primarily focused on the panelists' descriptions of the just-qualified candidates. This helped the panelists to understand the characteristics of just-qualified candidates, and helped to highlight the importance of those characteristics.

This study utilized two different familiarization activities, both of which aimed to familiarize the panelists with the CEFR levels and descriptors. Both activities asked the panelists to assign CEFR levels to a set of descriptors selected from the CEFR speaking scales. While sorting activities are rather challenging due to the decontextualization of the descriptors, they help to encourage familiarization with the CEFR by forcing participants to fully read and deeply consider the language of each descriptor. For the first familiarization activity, the panelists began by reviewing and discussing the overall oral production (Council of Europe, 2001, p. 58) and overall spoken interaction (Council of Europe, 2001, p. 74) scales. The discussion focused on understanding how the descriptors defined each CEFR level, as well as what features a just-qualified B1-, B2-, and C1-level speaker would exhibit. After the discussion the panelists were given a set of 28 descriptors from these scales, and were asked to individually assign CEFR levels to each descriptor. The results were then discussed as a group to help clarify any misclassified descriptors and to ensure that the panelists understood the CEFR levels. The second familiarization activity was similar to the first; however, it did not include an initial review or discussion of the scales. The panelists began the activity by individually assigning CEFR levels to a set of 49 descriptors from the sustained monologue (Council of Europe, 2001, p. 59), conversation (Council of Europe, 2001, p. 76), and spoken fluency (Council of Europe, 2001, p. 129) scales. Because these scales were not discussed prior to the activity, panelists needed to use their knowledge and understanding of the CEFR to help them complete the activity. As before, the results of this activity were then discussed as a group to ensure that the panelists understood the descriptors for each CEFR level.

Once the familiarization activities were finished, the participants completed a training task. The training task provided the panelists with the opportunity to practice making cut score judgments prior to the actual judgment round. The panelists were provided with 11 speaking performance excerpts, ordered by score, from the middle range of CaMLA Speaking Test scores (i.e., the scores ranged from 10–20 rather than the full range of 5–25). These performances were selected in the same way as those selected for the ordered item booklet (see Section 2.2), and a more narrow score range was selected to help reduce the panelists' workload for the training activity so that they could focus on understanding the judgment process. The panelists then practiced making their cut score judgments at the A2/B1 and B1/B2 boundaries. To make their practice judgments, the panelists listened to the audio clips using laptops and headphones, marking their cut score decisions on a spreadsheet. While making their decisions, the panelists had access to their notes and the CEFR scales, and they had the opportunity to listen to the audio clips multiple times. It should be noted that the panelists were instructed to think of the just-qualified candidate at each level when making their decisions. Once the panelists were finished making their practice judgments, the procedure was discussed as a group to address any questions or concerns. When the discussion of the training task had concluded, the panelists were given a prejudgment questionnaire to assess their understanding of the procedures and their willingness to proceed with the judgment task.

For the judgment task, the panelists followed the procedures practiced in the training task to determine the recommended cut scores for the CaMLA Speaking Test. The panelists were provided with the ordered item booklet, which consisted of 21 speaking performance excerpts from the entire range of CaMLA Speaking

Test scores (5–25). The panelists then listened to each performance and made their cut score judgments at the A2/B1, B1/B2, and B2/C1 boundaries. The procedure for making the judgments was the same as for the training task. The panelists listened to the audio clips using their laptops, and marked their decisions on a spreadsheet. The judgment task consisted of two judgment rounds, each of which was followed by a group discussion where the panelists discussed the results and shared their reasoning. The discussion of the first judgment round allowed the panelists to discuss the reasoning behind their cut score decisions. During this discussion the audio recordings for each cut score selection were replayed to the group so that the panelists could discuss the factors of the performance that influenced their decisions. The second judgment round utilized the same ordered item booklet and allowed the panelists the opportunity to make adjustments to their cut score decisions. The discussion of the second judgment round focused on finalizing the panel's cut score recommendations. The end result of the judgment task was a set of recommended cut scores. Once the discussion of the judgment task had concluded the panelists were given a post-judgment questionnaire to collect their opinions on the quality of the meeting and their confidence in the recommended cut scores. They also completed a post-study CEFR quiz through Qualtrics to assess how much their knowledge of the CEFR descriptors had improved throughout the study.

The procedures and results of the standard-setting meeting were documented throughout the meeting using Google spreadsheets, and analyzed after the meeting to help provide evidence of procedural, internal, and external validity to support the recommended cut scores.

3. RESULTS

3.1 SPECIFICATION

The first stage of a standard-setting study, known as specification (Council of Europe, 2009) or construct congruence (Tannenbaum & Cho, 2014), provides evidence that the skills and abilities measured by the test are "consistent with those described by the framework" (Tannenbaum & Cho, 2014, p. 237). This step is often done prior to the standard-setting meeting. It requires that the test developers justify the need for a linking study by showing that the test content is aligned with the target framework. This justification is necessary because, as Tannenbaum and Cho note: "If the test content does not reasonably overlap with the framework of interest, then there is little justification for conducting a standard-setting study, as the test would lack content-based validity" (2014, p. 237).

The linking of the CaMLA Speaking Test to the CEFR is justifiable because the test was specifically designed to assess spoken English language proficiency at the A2 through C1 levels of the CEFR. The CEFR was used throughout the development of the test as a reference to help define the test construct. The 'overall oral production' (Council of Europe, 2001, p. 58) scale was particularly relevant, and its descriptors were used to help inform the task design of the CaMLA Speaking Test.

The different tasks on the CaMLA Speaking Test were specifically designed to elicit a range of language at the A2 through C1 levels of the CEFR. The descriptors of these CEFR levels were used to determine the linguistic functions that would be elicited in the test, with a focus on the linguistic functions that help distinguish between CEFR levels. Table 3.1 provides a summary of the CEFR levels and linguistic features targeted by each of the tasks on the CaMLA Speaking Test. It shows that Tasks 1, 2, and 3 were aimed at beginner and low-intermediate level test takers, while Tasks 4 and 5 were aimed at intermediate to advanced level test takers.

Overall, the information summarized in this section provides evidence that the skills and abilities measured by the CaMLA Speaking Test are consistent with those described by the targeted CEFR levels, and that it is

Table 3.1: CEFR Level and Linguistic Functions Targeted by Each Task on the CaMLA Speaking Test

Task	Description	CEFR Level	Linguistic Functions
1	Describe a picture	A2	Describe people, places and possessions in simple terms.
2	Describe a personal experience	A2	Give short, basic descriptions of events and activities.
3	State and explain an opinion	B1	Briefly give reasons and explanations for opinions, plans, and actions.
4	Discuss advantages and disadvantages of various options	B2	Explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.
5	Argue for or against a point of view or proposal	C1	Expand and support points of view at some length with subsidiary points, reasons, and relevant examples.

(CaMLA, 2015)

justifiable to link the CaMLA Speaking Test to the CEFR. Additional information about the development of the CaMLA Speaking Test can be found in the CaMLA Speaking Test Development Report (CaMLA, 2015) located on the CaMLA website.

3.2 FAMILIARIZATION

This section summarizes the results of the two familiarization activities performed during the standard-setting meeting in order to establish the panelists' familiarity with the CEFR. This step is important because if the panelists do not understand the CEFR levels and their descriptors, then the validity of the recommended cut scores would be jeopardized since their judgments may also reflect this lack of understanding.

Recall from Section 2.3 that this study utilized two different familiarization activities. Table 3.2 summarizes the results of the first familiarization task, which required the panelists to assign CEFR levels to a set of 28 descriptors that the group had previously discussed and reviewed. Table 3.3 summarizes the results of the second familiarization task, which required the panelists to assign CEFR levels to a set of 49 descriptors that had not been previously discussed. The tables show the number and percentage of descriptors correct in the first and second row, respectively, the Spearman correlation (ρ) between the panelists' assigned CEFR levels and the correct levels in the third row, and the average assigned CEFR level for each panelist in the fourth row. The correlation coefficient shows the degree to which the panelists understand the progression of the CEFR levels, and should be interpreted in conjunction with the number and percentage of descriptors correct to understand the panelists' performance on the familiarization tasks. The average assigned CEFR level for each panelist was calculated by transforming their assigned CEFR levels to numbers (A1–1, A2–2, B1–3, B2–4, C1–5, C2–6) and taking the average. The panelists' averages can be compared with the average level of the descriptors (3.54 and 3.29 for familiarization activities 1 and 2, respectively) to assess the overall severity or leniency of the panelists. Panelists with average assigned CEFR levels higher than the actual average were generally more lenient, while panelists with average assigned CEFR levels lower than the actual averages were generally more severe.

Assigning the exact CEFR level to individual descriptors is a rather challenging task, but Tables 3.2 and 3.3 show that the panelists performed reasonably well on both familiarization activities. The panelists all assigned the correct CEFR level to over 40% of the descriptors for both activities, and analysis of the individual responses reveals that the majority of incorrect descriptors were placed at adjacent CEFR levels. The high correlation coefficients (≥ 0.834) also provide evidence that the panelists understood the progression of language proficiency across the different CEFR levels. Finally, the averages of the assigned CEFR levels show that the leniency and severity of the panelists were varied. Table 3.2 shows that most panelists were lenient on familiarization activity 1, while Table 3.3 shows that there was an even distribution of lenient and severe panelists on familiarization activity 2. Overall, the results summarized in these tables suggest that the panelists had a very good understanding of the CEFR descriptors. This understanding was strengthened through group discussion of the descriptor statements following each familiarization activity. These discussions were held to correct any misunderstandings and to ensure that the panelists understood the correct CEFR level for each descriptor.

Table 3.2: Familiarization Activity 1 Results (28 Descriptors, 3.54 Average CEFR Level)

	J1	J2	J3	J4	J5	J6	J7	J8	J9	J10	J11	J12
# Correct	18	14	19	15	23	21	18	13	19	21	13	16
% Correct	64.29	50.00	67.86	53.57	82.14	75.00	64.29	46.43	67.86	75.00	46.43	57.14
ρ	0.912	0.837	0.942	0.910	0.968	0.949	0.923	0.867	0.899	0.951	0.870	0.900
Average	3.89	3.71	3.71	3.46	3.71	3.57	3.86	3.68	3.61	3.46	3.79	3.50

Table 3.3: Familiarization Activity 2 Results (49 Descriptors, 3.29 Average CEFR Level)

	J1	J2	J3	J4	J5	J6	J7	J8	J9	J10	J11	J12
# Correct	23	29	20	34	33	31	28	23	28	37	29	28
% Correct	46.94	59.18	40.82	69.39	67.35	63.27	57.14	46.94	57.14	75.51	59.18	57.14
ρ	0.879	0.908	0.834	0.922	0.929	0.927	0.849	0.895	0.865	0.929	0.920	0.884
Average	3.00	3.25	2.93	3.14	3.29	3.50	3.21	3.64	3.29	3.46	3.50	3.29

The analysis of panelist familiarity with the CEFR has thus far been centered on the panelists' individual understandings of the descriptors. However, it is also important to assess the consistency of the panel as a whole since the cut score decisions will be based on the decisions of the entire panel. Table 3.4 presents three measures of internal consistency for each familiarization activity: Cronbach's alpha (α), the intraclass correlation coefficient (ICC), and Kendal's coefficient of concordance (W). These indices are three of the most frequently used measures of internal consistency (Kaftandjieva, 2010, p. 96). Cronbach's alpha (α) measures internal consistency by estimating the proportion of variance due to common factors in the items (Davies, Brown, Elder, Hill, Lumley, & McNamara, 1999, p. 39), the ICC measures internal consistency by taking into account both between- and within-rater variance (Davies et al., 1999, p. 89), and Kendall's W is a nonparametric measure of internal consistency that measures the level of agreement between three or more raters that rank the same group of items (Davies et al., 1999, p. 100). These three indices range from 0 to 1, with a value of 1 indicating complete agreement among panelists. Table 3.4 shows that all three indices are very high for each familiarization activity, with Cronbach's alpha (α) and ICC values very close to 1. This suggests that there is a very high level of agreement and consistency between the panelists for both familiarization activities.

Table 3.4: Panel Agreement and Consistency

	Activity 1	Activity 2
α	0.985	0.986
ICC*	0.985	0.984
W	0.832	0.842

* ICC values obtained using a two-way mixed model and average measures for exact agreement.

The familiarization stage was meant to expose panelists to the relevant CEFR descriptors and ensure that they all had an accurate understanding of each CEFR level. While the analysis of the familiarization activities above demonstrates that the panelists had a good understanding of the CEFR descriptors, it is important to note that these were learning activities, so some inaccuracies and inconsistencies from the panelists were expected at this stage. The descriptor statements were thoroughly discussed after each familiarization task, and any questions on the levels of the descriptor statements were addressed to ensure that the panelists understood the correct level of each descriptor.

One measure of the effectiveness of the familiarization tasks can be obtained through analysis of the pre- and post-study CEFR quizzes. Recall from Section 2.3 that the panelists were given a short CEFR quiz with their pre-study materials to assess their initial understanding of the CEFR, and that they were given another version of this quiz at the conclusion of the study to assess whether their understanding of the CEFR had improved. Table 3.5 summarizes the results of both quizzes for each panelist. It reveals that the scores improved after the standard-setting meeting for all but one panelist. Furthermore, it shows that the post-study quiz had an average score 5.75 points higher than the pre-study quiz, and analysis of the data with a paired t-test confirmed that this difference in scores was statistically significant ($t = 5.93$, $df = 11$, $p < 0.001$). These results provide evidence that the familiarization activities and their discussions helped to improve the panelists understanding of the CEFR descriptors.

Table 3.5: Summary of Pre- and Post-Study CEFR Quiz Results

Panelist ID	Pre-Study	Post-Study
J1	8	17
J2	9	12
J3	10	17
J4	2	14
J5	10	14
J6	13	12
J7	6	14
J8	5	9
J9	3	8
J10	9	17
J11	8	12
J12	7	13
Average	7.50	13.25
SD	3.12	2.93

Overall, the analysis of the familiarization activities reveals that the panelists had a good understanding of the CEFR levels and that the activities and discussions were successful in helping them understand the CEFR descriptors. The comments made throughout the discussion of the familiarization activities, the responses to the pre- and post-judgment surveys (see Section 4.1), and the low variability of the judgment task (see Section 3.3) also suggest that the panelists understood the CEFR levels and the differences between adjacent levels.

3.3 JUDGMENT

This section summarizes the results of the judgment task. Table 3.6 presents each panelist's cut score recommendations for both judgment rounds and provides summary statistics for the panel as a whole. It shows that the panelists' cut score recommendations were all very similar. The relatively small standard deviations and the small ranges of recommended cut score values provide evidence that there was very little variation in the panelists cut score recommendations for each level. Furthermore, the table shows that there was a decrease in the variability between judgment rounds. This decrease was expected (Tannenbaum & Katz, 2008) and is a result of the group discussion of the cut score recommendations after the first judgment round.

The average cut scores presented in Table 3.6 represent the panel's initial cut score recommendations for the CaMLA Speaking Test. The average values were used as the initial recommendations because they allow each panelist's recommendation to have equal weight. However, using the average values results in noninteger cut score recommendations and CaMLA Speaking Test scores are reported as integers, so the initial estimates needed to be slightly modified. Generally, cut scores are rounded up to the nearest score point to minimize the chance of false positive classifications (Cizek & Bunch, 2007, p. 25). After discussing the results of the second judgment round, the panel decided to round the B1/B2 and B2/C1 cut score recommendations up to 17 and 21, respectively, however, the A2/B1 cut score recommendation was rounded down to 11 since it was more representative of the panel's recommendations.

4. VALIDITY EVIDENCE

4.1 PROCEDURAL

The above documentation of the standard-setting study provides procedural validity evidence to support the quality of the standard-setting meeting and the recommended cut scores. This section summarizes the results of two surveys that were given during the standard-setting meeting in order to provide additional procedural validity evidence. The first survey was given immediately after the training task and prior to the judgment rounds, and it focused on the panelists' understanding of the familiarization and training tasks. The second survey was given at the conclusion of the standard-setting meeting and focused on the panelists'

Table 3.6: Panelist Cut Score Judgments

Panelist	Judgment Round 1			Judgment Round 2		
ID	A2/B1	B1/B2	B2/C1	A2/B1	B1/B2	B2/C1
J1	12	16	21	12	16	21
J2	12	16	21	12	16	21
J3	12	17	21	11	17	21
J4	10	16	20	12	17	20
J5	11	16	20	11	17	20
J6	12	17	20	11	17	20
J7	10	17	23	11	18	21
J8	11	17	21	11	17	21
J9	11	18	23	11	17	21
J10	11	17	20	11	17	20
J11	11	16	19	11	17	21
J12	11	17	21	11	17	21
Average	11.17	16.67	20.83	11.25	16.92	20.67
Median	11	17	21	11	17	21
SD	0.72	0.65	1.19	0.45	0.51	0.49
SEj	0.21	0.19	0.34	0.13	0.15	0.14
Min	10	16	19	11	16	20
Max	12	18	23	12	18	21

understanding of the judgment rounds and their thoughts on the recommended cut scores. Both surveys utilized a four-point Likert scale (1–strongly disagree to 4–strongly agree) to collect most of this information. Tables 4.1 and 4.2 present the questions and summarize the results for the pre- and post-judgment surveys, respectively. In addition to these questions, the pre-judgment survey also asked panelists if they were ready to proceed with the judgment phase (yes or no), and the post-judgment survey asked panelists their opinion of the recommended cut scores (too low, about right, or too high).

The tables show that panelists generally responded favorably to the surveys. The panelists indicated that they understood the familiarization, training, and judgment tasks, and that they felt they had enough time to complete the required tasks and participate in group discussions. On the pre-judgment survey all twelve panelists indicated that they felt ready to continue to the judgment phase, and on the post-judgment survey they all responded that they felt the cut score recommendations were about right. Only two panelists responded negatively to any of the survey statements.

One panelist disagreed with pre-judgment statements 3 and 6, indicating that she felt the introductory presentation did not help her to understand the linking process and that the training task did not help her to understand the judgment process. It should be noted that while the introductory presentation may not have addressed all of the panelist's questions about the linking process, her responses to the remaining statements on the pre- and post-judgment surveys suggest that she was able to learn and understand the process throughout the meeting. Additionally, despite indicating on the pre-judgment survey that the training task did not help her to understand the judgment process, the panelist also indicated on the post-judgment survey that the training task did help her to understand the judgment process. This inconsistency suggests that the panelist may have had some initial questions applying the method, but that she became more comfortable with it throughout the study.

A different panelist disagreed with post-judgment statement 7, which meant that she did not feel confident in the decisions she made. Unfortunately, the panelist did not provide any additional comments explaining why she lacked confidence; however, her responses to all the other survey statements indicate that she understood the standard setting method and that she felt the recommended cut scores were about right. This suggests that while this panelist lacked confidence in her decisions, she still understood the procedure and felt that the panel arrived at appropriate cut scores.

Table 4.1: Summary of Pre-Judgment Survey Results

No.	Question	1	2	3	4
1	The pre-reading helped me to understand the background to the CEFR.	-	-	6	6
2	The sample test helped me to understand the structure and level of the CaMLA Speaking Test.	-	-	3	9
3	The introductory presentation helped me to understand the linking process.	-	1	5	6
4	The discussion of pre-reading answered my questions.	-	-	6	5
5	The familiarization tasks helped me to understand the CEFR levels.	-	-	4	8
6	The training task helped me to understand the judgment process.	-	1	3	8
7	I had enough time to complete my individual tasks.	-	-	1	11
8	I had enough time to participate in the discussions.	-	-	-	12

Table 4.2: Summary of Post-Judgment Survey Results

No.	Question	1	2	3	4
1	The familiarization tasks helped me to understand the CEFR levels	-	-	3	9
2	The training task helped me to understand the judgment process.	-	-	3	9
3	I understood the instructions for each judgment round.	-	-	4	8
4	I understood the group discussion of our judgments.	-	-	1	11
5	I had enough time to complete my individual tasks.	-	-	-	12
6	I had enough time to participate in the discussions.	-	-	-	12
7	I am confident in the decisions I have made.	-	1	2	9

Overall, the responses to the pre- and post-judgment surveys were very positive. They indicate that, as a whole, the panelists understood the standard-setting procedure and were satisfied with the cut score recommendations. These results provide procedural validity evidence that supports the quality of the cut score recommendations.

4.2 INTERNAL

This section aims to provide internal validity evidence to support the CaMLA Speaking Test's recommended cut scores. One piece of internal validity evidence can be obtained by examining the likelihood that the recommended cut scores can be replicated. This can be estimated using the standard error of judgment (SE_j) of the panel's cut score recommendations (Tannenbaum & Cho, 2014). Cohen, Kane, and Crooks (1999) suggest that SE_j values that are less than half the test's standard error of measurement (SEM) can be considered reasonable. That is, if the SE_j values are less than half the test's SEM, then the recommended cut scores would likely be replicated in another standard-setting study.

Analysis of pilot test data (n = 67) reveals that the CaMLA Speaking Test has an SEM of 1.21, which means that SE_j values would need to be less than 0.61 for each cut score to provide support for the validity of the recommended cut scores. Table 3.4 (Section 3.3) lists the final judgment round's SE_j values for each cut score (A2/B1 = 0.13, B1/B2 = 0.15, B2/C1 = 0.14). It clearly shows that the SE_j values are much less than half the SEM for each cut score, which suggests that the recommended cut scores would likely be replicated in another standard-setting study.

Another piece of internal validity evidence can be obtained through analysis of decision consistency. To measure this consistency, this report utilizes the methods and tables presented in Subkoviak (1988) to estimate the agreement coefficient (p_0) and kappa coefficient (κ) for each cut score. Both coefficients measure classification consistency in slightly different ways. The agreement coefficient is a measure of overall consistency that can be interpreted as the proportion of test takers that would be consistently classified on two administrations of the same test (Subkoviak, 1988). The kappa coefficient also provides an estimate of the proportion of test takers who would be consistently classified by two administrations after accounting for the proportion who would be classified consistently by chance (Subkoviak, 1988).

Reliability estimates and summary statistics from the CaMLA Speaking Test pilot data were used in conjunction with the tables from Subkoviak (1988) to obtain estimates for the agreement and kappa coefficients. Table 4.4 summarizes these estimates for each cut score. To help interpret these statistics, Subkoviak (1988) suggests that tests used to make important decisions should have agreement coefficients larger than 0.85 and kappa coefficients larger than 0.60. Table 4.3 shows that the agreement and kappa coefficients exceed these values, which suggests that the recommended cut scores demonstrate good decision consistency.

Table 4.3: Agreement Coefficient (p_0) and Kappa (κ) for Panel Cut Scores

Cut Score	p_0	κ
A2/B1	0.88	0.70
B1/B2	0.87	0.71
B2/C1	0.93	0.66

Overall, this section has provided two important pieces of internal validity evidence. The analysis of the SE_j values provides evidence that the recommended cut scores are replicable, and the decision consistency analysis provides evidence that the test can provide consistent classification with these recommended cut scores. These two pieces of internal validity evidence work to support the overall quality of the cut score recommendations.

4.3 EXTERNAL

This section summarizes the available external validity evidence to provide support for the recommended CaMLA Speaking Test cut scores. It should be noted that external validity evidence is often the most difficult kind to obtain (Council of Europe, 2009, Ch. 7). It generally consists of independent evidence that supports the results of the standard-setting study (Council of Europe, 2009, Ch. 7), such as cut score recommendations obtained using a different standard-setting method or the results from an external measure of the test takers' speaking ability (e.g., another CEFR-linked speaking test, CEFR judgments by teachers) to compare with CaMLA Speaking Test results. Unfortunately no external measures of the speaking ability were available, and applying a second standard-setting method would have increased the complexity of the judgment task, making it more difficult and time consuming for the panelists.

Despite this lack of data, this report attempts to provide some external validity evidence by exploring the reasonableness of the recommended cut scores. This was examined by comparing the CEFR distribution of the pilot CaMLA Speaking Test data (n = 67) with the CEFR distribution of the 2014 MET Speaking Test population. The demographics distribution of the pilot CaMLA Speaking Test was comparable to that of the 2014 MET Speaking Test population, so comparisons between the two tests may provide useful information. Table 4.4

presents the CEFR distributions for both exams. It shows that CEFR distribution of the pilot CaMLA Speaking Test data is very similar to the CEFR distribution of the 2014 MET Speaking Test population, with the majority of test takers scoring at the B1 and A2 levels for both exams. This similarity helps to provide some external validity evidence for the recommended cut scores.

Table 4.4: Distribution (in %) of Test Takers by CEFR Level

Test	A2	B1	B2	C1
CaMLA Speaking Test (Pilot Data)	28.36	34.33	22.39	14.93
MET Speaking Test (2014 Data)	30.04	32.24	22.38	15.34

5. CONCLUSION

Overall, this report has summarized the standard-setting study to link scores on the CaMLA Speaking Test to the CEFR. It documents both the procedures and results of the study, including the standard-setting meeting. It also provides procedural, internal, and external validity evidence to help support the quality of the panelists' cut score recommendations. Table 5.1 presents the final recommended CEFR cut scores for the CaMLA Speaking Test that have resulted from this study.

Table 5.1: Final CaMLA Speaking Test CEFR Cut Scores

CEFR Level	A2/B1	B1/B2	B2/C1
Total Score	11	17	21

6. REFERENCES

CaMLA (2015). *CaMLA Speaking Test Development Report*, CaMLA Technical Report, CaMLA. Retrieved from https://www.cambridgemichigan.org/wp-content/uploads/2015/08/CaMLA-Speaking-Test-Dev-Report.pdf

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

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting Performance Standards: Contemporary Methods. *Educational Measurement: Issues and Practice,* Winter 2004, 31–50.

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.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing.* Cambridge: Press Syndicate of the University of Cambridge.

Kaftandjieva, F. (2010). *Methods for Setting Cut Scores in Criterion-referenced Achievement Tests: A comparative analysis of six recent methods with an application to tests of reading in EFL*. Arnhem: Cito.

Mills, C. N., Melican, G. J., & Ahulwalia, N. T., (1991). Defining Minimal Competence. *Educational Measurement: Issues and Practice*, *10*(*2*), 7–10, 14.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The Bookmark procedure: Psychological perspectives. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 249–281). Mahwah, NJ: Erlbaum.

Morrow, K. (2004). Background to the CEF. In K. Morrow (Ed.), *Insights from the Common European Framework.* (pp. 3–11). Oxford: OUP.

Papageorgiou, S. (2010). Investigating the decision-making process of standard setting participants. *Language Testing, 27(2)*, 261–282

Tannenbaum, R. J. & Katz, I. R. (2008). *Setting Standards on the Core and Advanced iSkillsTM Assessments*, ETS RM-08-04, Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/RM-08-04.pdf.

Tannenbaum, R. J. (2011). Standard setting. In J. W. Collins & N. P. O'Brien (Eds.), *Greenwood dictionary of education* (2nd ed., p. 441). Santa Barbara, CA: ABC-CLIO.

Tannenbaum, R. J. & Cho, Y. (2014). Critical factors to consider in evaluating standard-setting studies to map language test scores to frameworks of language proficiency. *Language Assessment Quarterly*, *11*(*3*), 233–249.