CaMLA Speaking Test

Development 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

08/2015

TABLE OF CONTENTS

LIST OF TABLES

1. INTRODUCTION

The CaMLA Speaking Test, developed by Cambridge Michigan Language Assessments, assesses spoken language proficiency from the high beginner to low advanced levels, targeting the A2 to C1 ability levels of the Common European Framework of Reference (CEFR).

The purpose of the CaMLA Speaking Test is to evaluate a test taker's ability to produce comprehensible speech in response to a range of tasks and topics. Topics call upon a test taker's experiences, attitudes, or opinions about general, educational, or professional topics. These topics do not require any specialized topical knowledge.

The CaMLA Speaking Test is useful in a variety of settings for various users to assess spoken English-language proficiency. Educational institutions can use the CaMLA Speaking Test as a placement tool for English as a Second Language (ESL) or English as a Foreign Language (EFL) courses, to show student progress throughout a period of instruction, and/or as an exit test at the end of a course of study. Organizations can use the test to confirm readiness for work-related tasks at a range of levels and/or to check progress in employees' spoken English-language proficiency.

This report describes the development of the CaMLA Speaking Test. It provides a description of the test construct, task and rating scale development, and advice on how to set cut scores.

2. TEST CONSTRUCT

The CaMLA Speaking Test is targeted at levels A2 to C1 on the CEFR. The CEFR offers illustrative scales, or can-do statements, for overall oral production and overall oral interaction as well as specific speaking activities. The can-do statements for overall oral production as well as specific speaking activities are particularly relevant; these descriptors informed the task design of the CaMLA Speaking Test.

The progression in overall spoken production from levels A2 to C1 is provided in Table 1. As learners progress through each CEFR level they are expected to have mastered abilities described under lower levels of competence (A1–C1).

The table demonstrates that A2 level test takers are able to give simple descriptions, using short phrases or sentences, on topics that are very familiar. More able test takers are able to speak on an increasing range of topics using increasingly complex language (Council of Europe, 2001).

Table 1: Overall Oral Production

C1	Can give clear, detailed descriptions and presentations on complex subjects, integrating subthemes, developing particular points and rounding off with an appropriate conclusion.
B2	Can give clear, systematically developed descriptions and presentations, with appropriate highlighting of significant points, and relevant supporting detail.
	Can give clear, detailed descriptions and presentations on a wide range of subjects related to his/her field of interest, expanding and supporting ideas with subsidiary points and relevant examples.
B1	Can reasonably fluently sustain a straightforward description of one of a variety of subjects within his/her field of interest, presenting it as a linear sequence of points.
A2	Can give a simple description or presentation of people, living or working conditions, daily routines, likes/dislikes, etc. as a short series of simple phrases and sentences linked into a list.

Council of Europe, 2001, p. 58

The CaMLA Speaking Test measures a test taker's ability to:

- understand and use linguistic information (i.e., grammatical, lexical, phonological) in a variety of academic, social, and/or occupational situations;
- perform a variety of functions, including describing, making suggestions, stating opinions, narrating, explaining, speculating, and arguing a position or viewpoint;
- verbally respond automatically, in real time, and after reading printed prompts;
- produce responses that are grammatical and lexically accurate;
- produce responses that are intelligible, fluent, and relevant; and
- follow politeness principles and sociolinguistic norms.

3. TASK DEVELOPMENT

3.1 TASK THEORY

Tasks on the CaMLA Speaking Test are designed to elicit spoken language representing a range of ability levels from upper beginner to advanced (A2 through C1 on the CEFR). Descriptors of these levels determined the linguistic functions that would be elicited in the test. Specifically, those linguistic functions that distinguish one level from another are targeted by the CaMLA

Speaking Test. For example, while a learner at the B1 level can state an opinion, this learner cannot yet highlight significant points and provide relevant supporting details. Production of this latter function is expected to appear in the language of candidates at the B2 level (Council of Europe, 2001).

Examiner behavior can have an effect both on the amount and type of language that a candidate produces as well as on the score awarded to the candidate (Brown, 2005). Plough, MacMillan, and O'Connell (2010) stress that asymmetry between the interlocutor and the test taker has the potential to limit the language functions elicited in a speaking test. This can ultimately have severe consequences for the test taker's final score. In addition, each examiner has "distinct and individual styles which they tend to employ across interviews" (Brown, 2003, p. 2). Bearing these warnings in mind, the CaMLA Speaking Test design adopts a semidirect format in which the examiner uses a predetermined script to deliver instructions and task prompts. The advantages of this test format include the standardization of test content and delivery, an increase in the number of situations that can be presented to the test taker (Luoma, 1997), and an increase in the kind of linguistic functions elicited (Shohamy & Inbar, 1991; Luoma, 1997). These characteristics benefit the test takers by providing them with equal opportunities to demonstrate the extent of their proficiency.

3.2 TASK DESIGN

The CaMLA Speaking Test is a face-to-face test of spoken production, administered by one examiner to one test taker, and scored in real time. The test consists of five distinct tasks accessible to both lower- and higher-level test takers. Table 2 describes the purpose of each task on the CaMLA Speaking Test, the CEFR level targeted, and the corresponding linguistic functions.

Tasks 1–3 are aimed at beginner and low-intermediate speakers (A2 and B1 on the CEFR). They are thematically related and based around a topic that is introduced in a picture in Task 1. Tasks 4 and 5 are aimed at more proficient speakers (B2 and C1 on the CEFR). Tasks 4 and 5 ask test takers to discuss new topics, as more proficient speakers are expected to be able to handle a range of different topics without noticeable strain on their linguistic resources. An example of the CaMLA Speaking Test (both the examiner and the test taker prompt) is available on the CaMLA website.

The tasks are presented to the test taker both orally by the examiner and textually on a prompt sheet to maximize the likelihood that the test taker will clearly understand the task. Furthermore, the examiner discourse is entirely scripted. This approach ensures standardization of input. The benefits of this approach for the examiner include minimal examiner training and examiner focus on scoring and completion of the administrative procedures, rather on how to properly elicit test taker talk. The benefits for the test takers are that they are all presented with the same challenges and opportunities to give a good sample of their speaking ability.

4. RATING SCALE DEVELOPMENT

4.1 RATING SCALE THEORY

Table 2: CaMLA Speaking Test Tasks, CEFR Levels Targeted and Linguistic Functions

Part 1

Task	Description	Level Targeted	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.

Part 2

Task	Description	Level Targeted	Linguistic Functions
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.

Council of Europe, 2001, pp. 59–60

As suggested in Luoma (2004, pp. 80–81), test construct, tasks, and evaluation criteria for the CaMLA Speaking Test were developed concurrently. The decisions specific to the test construct and tasks are described in sections 2 and 3. Both intuitive and empirical methods to rating scale design were used to create the CaMLA Speaking Test rating scale (cf. Fulcher, 2003). Existing rating scales were used as a guideline to write the scale along with a discussion between a committee of experts to determine the wording of the descriptors and levels in the scale. In this section we describe the decisions taken, looking first at the number of levels to be described and then at the language features to be defined within each level.

According to McNamara (2000), deciding on the number of rating scale levels can be more a matter of practical utility than of theoretical validity. Many test-specific scales have four to six levels (Luoma, 2004, p. 80). In deciding on the number of levels in the CaMLA Speaking Test rating scale, it was important to consider the number of distinctions raters could reasonably be expected to make consistently. It was also important to consider the meaningfulness of the number of levels in the scale in terms of the degree to which they would correspond to the levels of ability that were being targeted by the test. In this respect, the number of levels on the CaMLA Speaking Test rating scale was influenced by the number of ability CEFR levels that were targeted; that is, A2–C1. This led to the initial creation of four levels. After analyses of sample CaMLA Speaking Test performances an additional level was added to account for test takers who produced little to no meaningful language.

The CEFR and other relevant speaking assessment literature were consulted to identify the criteria to be applied. The CEFR presents five qualitative aspects of spoken language use—range, accuracy, fluency, interaction and coherence (Council of Europe, 2001, pp. 28–29)—that is, what a test taker at each CEFR level "can do" when speaking. In addition, the scale was data-driven; an analysis of performance on tasks and descriptions of key features of performance were used to make decisions about the criteria that were relevant for the CaMLA Speaking Test (Fulcher, 2003, pp. 91–92). Three evaluation criteria for the scale emerged: Task Completion, Vocabulary & Grammar, and Intelligibility & Fluency.

The level descriptors for each criterion were designed to be brief, clear, concrete, and detailed enough (with the absence of field-specific jargon) to sufficiently guide raters from varying backgrounds to rate speaking performances consistently, and also allow them to make quick scoring decisions. In line with Luoma (2004), the rating scale levels were carefully defined and did not rely on evaluative labels such as a range from "excellent" to "poor"; such vagueness can cause raters to have difficulties making consistent scoring decisions. Word count and length of each level's performance descriptors were also considered, as the descriptors have to be concrete yet practical to be useful for raters (Luoma, 2004, p. 81).

Task Completion

Task Completion refers to the degree to which the test taker addresses the task presented in the prompt, that is, the relevance of the response to the task. This criterion also focuses upon both cohesion and coherence; that is, the extent to which it is an organized response that progresses in a logical order. At lower levels indicators of cohesion and coherence include the presence of simple connectors like "and," "but," and "because" (Council of Europe, 2001, p. 29). At higher levels, more complex discourse connectives are employed.

Vocabulary & Grammar

This criterion refers to how test takers use their lexical and syntactic resources to convey meaning. With respect to lexis, increases in proficiency level are associated with an increase in the number of words produced, or tokens, and a wider range of words, or types. Significant differences have been found for both token and type (Iwashita et al., 2008, pp. 37–38). Since speaking on the topic of the prompt is part of the construct of this test, relevance of the vocabulary to task is also evaluated. The Council of Europe (2001) states that test takers at the C1 level of the CEFR have a "good command of idiomatic expressions and colloquialisms" (Council of Europe, 2001, p. 112). Indeed, an analysis of speaking test performances revealed that advanced-level test takers typically use highly idiomatic vocabulary and rich formulaic sequences. Therefore, idiomatic expressions are included at the highest score point on the rating scale.

The indicators of syntactic complexity adopted include the use of clauses, verb phrases and length of utterance (Iwashita et al., 2008, p. 32). This is operationalized in the scale as the use of simple versus complex structures. Lower-level test takers are expected to have difficulty forming sentences and fragments accurately; verbs marked with tense and aspect and embedding are expected in the performances of more advanced-level speakers (Upshur & Turner, 1995, p. 9).

Intelligibility & Fluency

Intelligibility & Fluency refer to the clarity and delivery of the test taker's response. This criterion includes the notion of "listener effort"; that is, how hard the

listener has to work to understand the speaker (Brown et al., 2005). Listener effort in the CaMLA Speaking Test is often required for speech including frequent pauses, many attempts at repair (trying to self-correct language), inappropriate rate—either too slow or too fast—and/or where the pronunciation, intonation, and rhythm of speech is unclear. It is typical to categorize the concepts of "pronunciation, intonation, and rhythm" as "phonology" (Brown et al., 2005; Iwashita et al., 2008, pp. 38–40). However, to accommodate scale users from outside of the fields of linguistics and TESOL, the more accessible term "intelligibility" has been used. The pronunciation of words and syllables has been classified as "target-like," "marginally target-like," or "clearly non-target like" where the target is defined as "English-like." This is in line with Iwashita et al.'s (2008) finding that higher-level learners have more English-like intonation and lower-level learners have more non-English-like intonation. For instance, target-like syllables at both the word and subword level show more noticeable differences across levels.

Regarding the fluency of test taker's speech, three features have been found to discriminate among different English-language proficiency levels: speech rate, unfilled pauses, and total pause time (broadly categorized as hesitations). At lower levels of fluency, overly fast or slow speech rate have been found to cause problems for the listener; at higher levels of fluency, speech rate is typically consistent and appropriate (Brown et al., 2005, p. 38). Unfilled pauses have been found to characterize low-level learners; filled pauses and other types of hesitations (such as searching for content words or ideas) were shown to be markers of higher-level learners (Iwashita et al., 2008, p. 41). In addition, instances of repair have been shown to contribute to fluency judgments. Therefore, "fluency" is defined as "hesitations or pauses," "repair," and speech rate."

4.2 PILOTING THE RATING SCALE

The rating scale underwent a small-scale pilot (N = 28) to ensure that it was usable and meaningful, and that it distinguished appropriately between test takers at different levels. Three raters with experience in the TESOL profession scored sample videos of complete CaMLA Speaking Test performances. Each test performance was scored by two raters who assigned a holistic score of 1–5 (i.e., performance on all three evaluation criteria was considered) for each task. No single evaluation criterion or descriptor in the rating scale was weighted more than others. In the case of a discrepancy between raters, the performance was reviewed by all three raters, and an overall consensus score was reached.

The pilot test population consisted of 16 females and 12 males whose ages ranged from 18 to 46, with an average age of 26. The test takers represented a wide range of nationalities and first languages (L1). The largest language groups were Mandarin, Arabic, and Spanish, which is representative of the L1s of most learners in ESL language programs in the United States.

As a result of the piloting process, the rating scale was revised slightly. For example, several advanced level test takers used highly idiomatic vocabulary and rich formulaic sequences. The C1 level of the CEFR states that test takers have a "good command of idiomatic expressions and colloquialisms" (Council of Europe, 2001, p. 112). Therefore, idiomatic expressions were included at the highest score point on the rating scale in the Vocabulary & Grammar criterion. Additionally, our analysis showed that higher-scoring test takers were rarely hesitant in their speech. Instead, any pauses that occurred "allow(ed) speakers to continue the conversation if they wish to" (Fulcher, 1996, p. 218). At the lower levels of fluency, pauses occurred frequently because the test takers were not able to continue speaking (Fulcher, 1996, p. 217). Finally, "attempts at repair" were added to the Intelligibility & Fluency criteria, because this feature was evident at the mid-levels and repair impacts the effort expended by the listener in order to understand the speaker.

4.3 RATING SCALE DESIGN

Table 3 presents the evaluation criteria for the CaMLA Speaking Test. It identifies the language features that are relevant for each criterion.

Examiners are provided with a rating scale that is presented as a five by three grid; that is, 5 score points (1–5) and three evaluation criteria. Summative statements (in bold) are included within each score point; all other performance descriptors within the score point support this overarching statement. Additionally, the rating scale includes both positive (what test takers can do) and negative (what test takers cannot do) descriptors. This combination of information helps the examiner evaluate the test taker within the 10 minutes it takes to administer the test.

4.4 SCORING

The CaMLA Speaking Test is scored holistically by task, that is, for each of the five tasks, examiners give a holistic score (1–5) that takes into account the test taker's performance in relation to all three evaluation criteria. A test taker receives a total reported score ranging from 5–25. From this total score, inferences are made about the test taker's ability to use spoken English.

Table 3: Evaluation Criteria for the CaMLA Speaking Test

Main Criteria	Description of Features
Task Completion	Relevance of response to task • Coherence and cohesion: organized response that progresses in a logical order • Quantity of language, elaboration, and relevant supporting details in the response
Vocabulary & Grammar	Use of appropriate vocabulary and grammar to add meaning • Vocabulary: variety of vocabulary and relevance to task • Grammar: complexity and accuracy
Intelligibility & Fluency	Clarity and fluency of speech • Intelligibility: pronunciation of words and phrases; intonation; rhythm of speech and stress placement on syllables in words and phrases • Fluency: speech hesitations or pauses; occurrences of repair; speech rate

5. INTERPRETING TEST SCORES

CaMLA Speaking Test scores can be used as information for ESL/EFL or content-area departments, as a diagnostic tool to recommend placement into the appropriate ESL/EFL classes, to gain additional information about the test taker's strengths and weaknesses, or for eligibility for employment.

CaMLA recommends that each institution conduct a standard setting study to determine appropriate cut scores for use. Standard setting enables test users to interpret the meaning of the scores received on the CaMLA Speaking Test within the context of their institution. For instance, if a college ESL program has four levels of ESL courses, the institution will need to decide which cut scores place students in each level of their program. As part of standard setting, the institution will need to identify the cut scores, or scores below which a student should be placed in a less advanced group and above which they should be placed into a more advanced group. It is typically most straightforward for the test to be administered to a group of learners whose ability is already known. Cut scores can then be based on test scores for learners who are already in certain levels.

6. REFERENCES

Brown, A. (2003). Interviewer Variation and the Co-construction of Speaking Proficiency. *Language Testing,* 20(1), 1–25. doi: 10.1191/0265532203lt242oa

Brown, A. (2005). *Interviewer Variability in Oral Proficiency Interviews.* Frankfurt, Germany: Peter Lang.

Brown, A., Iwashita, N., & McNamara, T. (2005). An Examination of Rater Orientations and Test-Taker Performance on English-for-Academic-Purposes Speaking Tasks. *ETS Report. MS-29*. Retrieved from http://www.ets.org/Media/Research/pdf/RR-05-05.pdf

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

Fulcher, G. (1996). Does Thick Description Lead to Smart Tests? A Data-based Approach to Rating Scale Construction. *Language Testing,* 13(2). 208–238. doi: 10.1177/026553229601300205

Fulcher, G. (2003). *Testing Second Language Speaking.* Harlow, UK: Pearson Longman.

Iwashita, N., Brown, A., McNamara, T. & O'Hagan, S. (2008). Assessed Levels of Second Language Speaking Proficiency: How Distinct? *Applied Linguistics,* 29(1), 24–49. doi.10.1093/applin/amm017

Luoma, S. (1997). *Comparability of a tape-mediated and a face-to-face test of speaking.* (Unpublished Licentiate thesis), University of Jyvaskyla, Finland.

Luoma, S. (2004). *Assessing Speaking.* Cambridge, UK: Cambridge University Press.

Plough, I., MacMillan, F. M., & O'Connell, S. P. (2011). Changing Tasks…Changing Evidence: A Comparative Study of Two Speaking Proficiency Tests. In Granena, G., Koeth, J., Lee-Ellis, S., Lukyanchenko, A., Botana, G. P., and Rhoades, E. (Eds). *Proceedings of the 2010 Second Language Research Forum* (pp. 91–104). College Park, MD: Cascadilla Press.

Shohamy, E. & Inbar, O. (1991). Validation of Listening Comprehension Tests: The Effects of Test and Question Type. *Language Testing.* 8(1), 23–40. doi: 10.1177/026553229100800103

Upshur, J. & Turner, C. (1995). Constructing rating scales for second language tests. *ELT Journal,* 49(1), 3–12. doi: 10.1093/elt/49.1.3