CaMLA Working Papers

2016–02

The CaMLA Speaking Test: Face-to-Face vs. Audio Delivery

Ildiko Porter-Szucs
Cynthia Macknish
Barry DeCicco

The CaMLA Speaking Test: Face-to-Face vs. Audio Delivery

Authors

Ildiko Porter-Szucs
Cynthia Macknish
Barry DeCicco

About the Authors

Ildiko Porter-Szucs is ESL/TESOL faculty in the Department of World Languages at Eastern Michigan University, Ypsilanti, Michigan, USA. She received her Ph.D. in Higher, Adult, and Lifelong Education with an emphasis in TESOL at Michigan State University. Her teaching and research interests lie in second-language assessment and teacher formation. She is a certified International Teaching Assistant Speaking Assessment/ Graduate Student Instructor Oral English Test (ITASA/GSI OET) examiner.

Cynthia Macknish received her Ed.D in TESOL and Applied Linguistics from the University of Leicester, UK. She currently teaches in graduate and undergraduate TESOL programs, and the ESL program at Eastern Michigan University. Her research interests include language pedagogy and language assessment. She is a certified IELTS examiner.

Barry DeCicco was employed as a statistician in the University of Michigan Health System at the time this study was conducted. Since then he has moved to the automotive industry. He has a MS in Probability and Statistics from Michigan State University and a MS in Industrial and Operations Engineering, University of Michigan - Ann Arbor, with Ph.D. level coursework in biostatistics. He has published in the *American Journal of Critical Care*, *Urology Practice*, and *The Gerontologist*.

Table of Contents

Abstract

This study examines whether the mode of delivery—direct (face-to-face) or semi-direct (computer-mediated audio-recorded)—influences test takers' scores on the CaMLA Speaking Test (CST). A mixed-methods design was employed for data analysis. The results were analyzed to answer four research questions:

1) Do CST test takers receive different scores on face-to-face vs. audio-delivered modes of the speaking test?
2) Do CST test takers prefer one mode of test delivery over the other?
3) Is there any alignment between test takers' scores and preferred mode of test delivery?
4) What factors do test takers claim affect their performance on both modes of the speaking test?

The first question was answered by fitting a linear mixed model to the data and examining parameter estimates for the factors and their interactions. The second question was answered by examining the percentages of test takers who stated the different pre-test preferences and post-test beliefs. The third question was answered using ANOVAs to compare differences between mean outcomes by mode vs. the test takers' answers to the preference/belief questions. The fourth research question was answered by analyzing the open-ended comment section on the post-study survey for themes pertaining to perceived factors influencing test takers' performance on the two modes of test delivery.

Background

With the recent launch of the CaMLA Speaking Test (CST), CaMLA has responded to a need in the field of English-language assessment for a stand-alone speaking test that can be administered by institutions themselves. According to the CaMLA website, the CST is quick, reliable, easy to administer and score, and provides meaningful results to assess a person's ability to produce comprehensible spoken English (CaMLA, n.d.).

As the CST assesses test takers' (TTs') ability to produce spoken English, rather than two-way communication, the question arises as to whether there is a need to use live examiners to administer the test when a voice talent could record an audio-delivered test instead. Cost-conscious test users may find an audio-delivered option of the CST attractive[1].

Literature Review

Spoken language proficiency is typically evaluated through, what Clark (1979) calls, direct or semi-direct tests. Direct tests are conducted by a live examiner, often one-on-one and face-to-face. Semi-direct tests, on the other hand, utilize a series of prerecorded prompts for TTs to respond to. These responses are recorded to be scored at a later time (Ginther, 2012). Computer-delivered human-scored tests have resulted in scores that are comparable to live tests (Stansfield & Kenyon, 1992). The correlation between these two modes of delivery has been found to be high (.89 to .95), strengthening the claim that these modes are comparable (Ginther, 2012; Thompson, Cox, & Knapp, 2016). In contrast with the favorable statistical findings, most TTs prefer direct tests because of the artificiality of interacting with a recorder or computer (Brown, 1993; Ginther, 2012; Qian, 2009).

Studies investigating face-to-face vs. telephone interviews have yielded similar results. During face-to-face communication, the speakers offer each other paralinguistic clues, such as body language,

1 Throughout the paper, direct tests with prompts delivered by a live examiner will be referred to as 'face-to-face', while semi-direct tests with prompts delivered by computer will be referred to as 'audio-delivered' or 'audio tests'. In both versions of CST delivery, TTs could also read the print copy of the prompt.

for understanding (Vogl, 2013). Such clues often entail emotional reactions to what has been said. In response, the speakers then adjust what is said next. These paralinguistic clues are absent in telephonic communication. Qualitative analyses of face-to-face vs. telephone interviews have produced mixed results. In some studies, researchers have found that participants speaking on the telephone spoke less fluently and produced statistically significantly more filled pauses—*ums* and *uhs* (8.0/100 words on the phone vs. 6.1/100 words face-to-face)—when they found it difficult to answer a question (Conrad et al., 2007, as cited in Vogl, 2013). Yet in another study, Sykes and Collins (2001, as cited in Vogl, 2013) found that participants spoke faster and produced fewer total pauses because they felt awkward pausing when they could not see their interlocutor's face. Though various studies have produced seemingly contradictory results, what is consistent is that the ability to see one's interlocutor alters the behavior of the study participants and yields qualitatively different speech.

The effectiveness of spoken performance typically depends on the ability to comprehend the input. In the context of listening, the difficulty of a task is influenced by numerous variables. Of particular relevance to the present study of a speaking test is the contribution to task difficulty made by low-frequency formulaic expressions and other vocabulary in the initial tasks of the input material. Several studies of listening comprehension have revealed that the proportion and nature of low-frequency lexical items influence listening difficulty (see Brunfaut & Revesz, 2015; Kostin, 2004; Muljani, Koda, & Moates, 1998; Revesz & Brunfaut, 2013). Brunfaut and Revesz's own study revealed that "lexical complexity characteristics of the listening passages significantly correlated with task difficulty" (p. 159). In fact, the "nature of the relationship between task difficulty and individual phrase-related characteristics seemed to depend on the corpus-based frequency of the expressions" (p. 159). Low-frequency words, in particular idioms, appeared to increase the difficulty of tasks. At the same time, the presence of high-frequency formulaic sequences in the input (such as "rely on" and "in a way") had an inverse effect on task difficulty. As Brunfaut and Revesz reason, this can be explained by the fact that formulaic sequences seem to be "prefabricated: that is, stored and retrieved whole from memory" (Wray, 2002, p. 9), thereby reducing the brain's processing load and increasing processing speed.

The speed of delivery has also been found to impact listening comprehension, particularly for lower-level TTs. "Faster delivery of speech is assumed to cause more listening difficulty, because it affords a shorter period of time to process the incoming information" (Brunfaut & Revesz, 2015, p. 145). Rosenhouse, Haik, and Kishon-Rabin (2006) investigated Arabic L1 and Hebrew L2 bilinguals' ability to comprehend speech under a variety of conditions. They varied the speech rate and the background noise. Their findings indicate that under ideal conditions, the bilinguals performed as well in their L2 as in their L1. Under adverse conditions, however, such as increased rate of delivery or background noise, the TTs performed worse in their L2 than in their L1.

Additional factors, such as the use of pauses by the examiner, may influence TTs' listening comprehension. Blau (1990) conducted research on the effect of syntax, speed, and pauses on the listening comprehension of Polish and Puerto Rican ESOL students. In one part of the study, the effect of speed and pauses in the delivery of listening passages was tested. Results show that slower delivery yielded significantly higher listening comprehension scores for the Puerto Rican group that listened to simple sentences at reduced speed (there were no statistically significant differences in the listening comprehension of the Polish group at slow vs. normal speeds). In another part of the study, three longer passages were recorded at 'normal speed' at 200 wpm, 'slowed down speed' at 185 wpm, and with 3-second pauses inserted between sentences, clauses, and phrases, which slowed the average rate down to 150 wpm. Unlike in the first stage of the study, when only the speech rate was manipulated, in the second stage the presence of pauses (and slower speech rate) yielded statistically significantly higher listening comprehension scores. This time both the Puerto Rican and Polish groups listening to the version with pauses performed better on the listening comprehension test. The slowed down version alone (without pauses), however, only improved the listening comprehension of the Puerto Ricans (not the Poles), particularly those at lower proficiency levels. Pausing at constituent boundaries did, however, contribute significantly to comprehension.

Other factors may influence the results of a study investigating the effects of the mode of delivery on the TTs' scores. One such factor often critiqued in the literature is the lack of demonstrated parallel test form reliability in speaking tests. Weir and Wu (2006) cite a number of empirical studies that indicated that manipulating the difficulty of tasks can have an effect

on the TTs' performance. In studies where the goal is to establish task and form equivalency, it makes good sense to administer two forms in quick succession with overlapping facets and tasks between them because the "error variance in this case represents fluctuations in performance from one set of items to another, but not fluctuations over time" (p. 170).

The industry standard of an interrater reliability of 0.8 or higher has been an acceptable measure of raters' ability to assign reliable scores to any given TT's performance. However, studies of the ACTFL (American Council on the Teaching of Foreign Languages) OPI (Oral Proficiency Interview), a direct test, and the Simulated OPI (SOPI), a semi-direct test, reveal that though interrater reliability may be high, trained raters can vary greatly in exact agreement between scores. Across studies perfect agreement has ranged from a low of 25% to a high of 88% (Kenyon & Tschirner, 2000).

Speaking tests, whether direct or semi-direct, depend on standardized examiners and raters. Examiners undergo training to achieve standardization, yet they have been found to behave differently during the test. One such difference is how much they adjust their speech to match the candidate's speaking proficiency (Brown, 2003). In some cases the differences between the interviewers' speech adjustments impact the candidates' scores (Ross, 1992; Cafarella, 1993, as cited by Brown, 2003). Examiners' individual styles are perceivable even in scripted interviews. Particularly in semi-direct tests, the rater may or may not be the same person as the examiner. Brown (2003) reports on studies by Morton et al. (1997) and McNamara and Lumley (1997), where the examiner and rater were decoupled. These studies focused on rater behavior in response to the raters' perceived competence of the examiners. If the raters judged the examiners to lack competence, the raters compensated for this by awarding the TT a higher score. In a semi-direct test the examiner is bypassed and the input material is recorded in a standardized way by voice talent. However, the possibility remains that when scoring, trained examiners make accommodations for factors extraneous to the construct represented in the rating scale, such as the TT's willingness or readiness to communicate (Brown, 2003).

Research Questions

The aim of this study was to determine if the mode of delivery of the CST—which is currently available only as a direct face-to-face test—makes a significant difference in test scores. To this purpose, the study was guided by the following research questions.

1) Do CaMLA Speaking Test test takers receive different scores on face-to-face vs. audio-delivered modes of the speaking test?
2) Do CaMLA Speaking Test test takers prefer one mode of test delivery over the other?
3) Is there any alignment between test takers' scores and preferred mode of test delivery?
4) What factors do test takers claim affect their performance on both modes of the speaking test?

Methodology

To answer these research questions, this study collected both quantitative and qualitative data. The combination of quantitative and qualitative data elicited served to provide evidence of test score differences, as well as help explain those differences (Humphreys et al., 2012).

Description of the CaMLA Speaking Test

According to the company's website, the CaMLA Speaking Test can be administered by institutions themselves (CaMLA, n.d.). It is conducted face-to-face, by one examiner to one TT; is scored by the examiner concurrent to the test's administration; and takes up to 10 minutes from start to finish. The rating scale has been designed to capture four general areas of spoken performance:

- fluency and intelligibility
- vocabulary range and relevance to task
- grammatical complexity and accuracy
- ability to successfully complete a specific task.

It is important to note that the speaking construct behind the aforementioned four areas of spoken English is transactional rather than interactional.

CaMLA recommends that the CST can be used as a placement, progress, and exit test. The test consists of five tasks: picture description, narrative, opinion, comparison of advantages-disadvantages, and persuasion. The tasks are designed to be progressively more challenging, both cognitively and linguistically. The

test is aligned with the Common European Framework of Reference (CEFR) and spans A2 (or high beginner) to C1 (or low advanced) levels of speaking proficiency (for information on the CEFR, see Council of Europe, 2001; for the technical report on the linking study, see CaMLA, 2015). The test is fully scripted for ease of administration and increased reliability. During test administration, TTs respond to one of several entirely scripted test cards. The instructions and tasks on the test cards are both read aloud by the examiner and given to the TTs to read along silently.

Test Takers, Tasks, Examiners, Raters, Procedures

In the present study, after all ethical requirements were met and Institutional Review Board approval granted, participant recruitment began. The participant TTs comprised a purposive (rather than convenience) sample of 106 nonnative English speakers from southeast Michigan with various proficiency levels, first languages (see Appendix A), and genders (see Table 1). All volunteers who qualified (i.e., spoke English as a second language, were over the age of 18, and signed the informed consent form) were included in the study.

Table 1: Test Takers' Gender

n=106	Female	Male
TT GENDER	63 (59%)	43 (41%)

The test-taker pool was fairly diverse: 41% male and 59% female, speaking 21 different native languages, ranging in age from 18 to over 60. Most TTs were students, but some were employed, while others were either out of the workforce or retired. The age and employment range of the TTs is a consequence of the purposive recruitment efforts of the researchers. The vast majority of the TTs was recruited from among students at local universities and language schools. A minority of the TTs comprised faculty and staff of these institutions, who responded to recruitment flyers posted around their institutions. As a result of snowball sampling, members of the community and acquaintances of the TTs also took part. As the CST can be used as a placement, proficiency, and exit test for language schools, institutions of higher learning, and companies, the make-up of the TT pool seems representative of the target TT population.

The TTs were exposed to both modes of delivery, but the order of the modes, the test card, and the examiner were randomized for each TT. This counterbalanced design served to ameliorate any effect of fatigue, or any decrease in stress as the TT took a second test. Each TT was exposed to only one examiner. And each TT received only one test card with one set of tasks for both the face-to-face and audio modes. The TTs had the prompts (physically) in front of them when they heard the audio recorded examiners' speech, so they were able to read the prompts and hear the prompts at the same time. Our study sought to determine whether the two modes of test delivery produced comparable results. Therefore, in order to reduce the potential influence of different test forms within TTs, each TT was given the same exact test form for both modes. In all, four full tests with five tasks each were used in the study; they were named after the picture in the first task on each test forms: Library (#1), Train Station (#2), Restaurant (#3), and Kitchen (#4).

Each TT took a pre- and post-test survey (see Appendixes B and C) to express their preferences about modes of test delivery and factors that they anticipated would impact their performance. Following the pre-test survey, each TT took a 10-minute face-to-face (F2F) CST and a 10-minute audio-delivered CST. The F2F tests were delivered by the two researchers from Eastern Michigan University: Cynthia Macknish (CM) and Ildiko Porter-Szucs (IPS). Both CM and IPS are experienced oral examiners of high-stakes standardized tests. For this study, they participated in the examiner-training session together. They completed the benchmarking, calibration, and qualification stages of the Examiner Training Manual (CaMLA, 2014) and passed the training within the acceptable margin of error, as recommended in the training packet supplied by CaMLA. Throughout the entire research study, the examiner-raters communicated with each other about progress and any issues that arose. In order to eliminate as many variables as possible, the two examiners agreed on clothing (business casual, with minimal jewelry, plain hairstyles) and manner of delivery (friendly but neutral facial expression and tone of voice), and strived to standardize the way they behaved in the audio and live F2F tests. The CaMLA Speaking Test Examiner Training Manual recommends that examiners

> [s]peak naturally and at a normal rate of delivery when possible. Examiners may at times need to make linguistic accommodations to the information delivered to test takers. Accommodations should be limited to repetition, rate, and manner of speech (e.g.,

enunciating more clearly or slowing down your own speech). (CaMLA, 2014, p. 9)

The two examiners adhered to these guidelines. All F2F tests were video-recorded so that they could be viewed and rated by the other researcher, who did not serve as the live examiner.

The audio-delivered tests comprised audio recordings of CM and IPS delivering each of the four sets of speaking tasks. The audio-recorded instructions and tasks were saved on a laptop and then played to the TTs in lieu of a live interlocutor. This was facilitated by a trained research assistant who took the following steps:

- seat the TT in a private room in front of the laptop,
- explain to the TT the testing procedures (including that the computer should not be touched but that the assistant is to be called for help if necessary),
- turn on the Audacity audio capture software,
- turn on the prerecorded instructions and series of prompts,
- hand the TT the preselected test card,
- remain with the TT while the instructions were being played to ascertain that there were no questions about the testing procedure, and
- leave the room before the first task was heard from the recording.

The TTs' responses, along with the prerecorded instructions and tasks, were audio-recorded onto the laptops through the built-in microphone using the open-source audio capture software Audacity (Audacityteam.org).

The audio-delivered test was not interactive. Therefore, in both modes of delivery, the examiners waited the full amount of allowable response time for each question. Exceptions were made in the F2F mode when the TTs indicated verbally or nonverbally that they had completed their response before the allotted time was up.

CM and IPS served as both examiners and raters in all the tests (see Table 2). As examiners, they administered approximately the same number of tests live (CM administered fifty-two and IPS fifty-four). Four of IPS's tests had problems due to either human error (two TTs failing to respond to some tasks) or equipment failure (two incoming video/audio recordings becoming corrupt). Two of CM's tests also had problems due to equipment failure. Therefore, IPS conducted four more tests (skipping over test numbers 103 and 105, which had been randomized to CM) and CM conducted two more so that there would remain fifty complete tests each.

During the audio-recorded test, each TT listened to the voice of the same examiner who administered their F2F test. Each examiner served as live rater for her own F2F test. In order to determine interrater reliability, each examiner rated the other examiner's F2F and audio test blind and her own audio test. The video-recorded F2F tests and all the audio-recorded tests were rated at a later date, separately by each examiner-rater. Video recording occurred with a combination of flip cameras, laptops, and cell phones, all of which have built-in microphones. Table 2 depicts two lines from the scoring scheme. The first column shows the TT number. The second and fourth columns together show the order of the two modes. TT #21 completed the F2F version of the test first and the audio version second. The third column depicts the initials of the live examiner. The fifth column entitled "Speaking Test #" depicts which set of tasks the TT was randomized to. TT #21 spoke about the Train Station. Following the two columns containing the time and date of the tests can be seen five columns containing the scores of the first rater (IPS) and five more columns containing the scores of the second rater (CM). As IPS was the F2F examiner for TT #21, the F2F scores by

Table 2: Scoring Scheme

TT#	TestOrder_ within_ TT	Examiner	Mode	Speaking Test #	Task 1 Score IPS	Task 2 Score IPS	Task 3 Score IPS	Task 4 Score IPS	Task 5 Score IPS	Task 1 Score CM	Task 2 Score CM	Task 3 Score CM	Task 4 Score CM	Task 5 Score CM
21	1	IPS	F2F	Train Station #2	3	3	3	4	3	3	3	3	3	4
21	2	IPS	audio	Train Station #2	3	3	3	3	3	3	3	3	3	4

IPS were awarded during the live test administration. The other three sets of scores displayed in Table 2 – IPS's scores of IPS's audio, CM's scores of IPS's videotaped F2F test, and CM's scores of IPS's audio – were awarded at a later time.

Results

Scoring

The CST is scored holistically on a scale of 1-5 for each task, where 1 is the lowest and 5 is the highest score. Achieving the lowest or the highest score on any task, however, deserves further explanation. TTs whose responses match the descriptors of the lowest or highest end of the rating scale will receive a score of 1 or 5, respectively. However, so will TTs whose performance is below a 1 or above a 5, respectively. Therefore, achieving a minimum or maximum score on the CST may indicate that the test was either too difficult or too easy for the TT (for score breakdown see Appendix E).

The total score on the five tasks of the test ranges from 5 to 25 points. In this study, all tests were double-scored. The live F2F performance was scored by the examiner concurrent to test administration. The video recording of the F2F test and the audio-recorded test were scored asynchronously, at a time removed from test administration. The examiner-raters scored all the tests by themselves blind. During every rating session, they consulted the Evaluation Criteria and Rating Scale, as recommended by the Testing Coordinator's Manual (CaMLA, 2014, p. 5). Interrater reliability, intrarater reliability, and exact agreement by CEFR bands were within industry norms (see Appendixes D1-3 and Appendix E).

Resuslts and Analyses

Research question 1: Do CaMLA Speaking Test test takers receive different scores on face-to-face vs. audio-delivered modes of the speaking test?

Analysis Method: This question was answered by fitting a linear mixed model to the data (to deal with the fact that there were multiple iterations of test administration and scores within each TT). The statistical significance of two-way interactions was assessed; those that were not statistically significant were removed from the model. The statistical significance of selected three-way interactions was then assessed.

Answer: The overall marginal mean scores between the two modes were not statistically significantly different (see Table 3). However, the overall marginal means were misleading; there were multiple statistically significant interactions (see Table 4), which canceled out overall for the modes.

One such interaction comprises Rater by Mode (see Table 5). The differences between the mean scores for Audio and F2F modes differed statistically significantly between the two raters. For CM the difference (Audio minus F2F) was positive; for IPS the difference (Audio minus F2F) was negative. This is indicated by the interaction mentioned below in the list of statistically significant effects from a linear mixed model. The practical implication is that there is a strong rater effect even for the Audio mode, which both raters scored asynchronously (for F2F, the examiner rates the TT synchronously, at the time of administration; the other rater rates the TT asynchronously, working from a video recording).

Table 3: Marginal Means by Task for Modes

Marginal Means by Task for Mode	Mean*	Std. Error	95% Confidence Interval		P-Value
			Lower Bound	Upper Bound	
Audio	3.12	0.03	3.06	3.18	
F2F	3.12	0.03	3.06	3.18	
Difference: Audio minus F2F	0.00	0.04	-0.09	0.08	0.946

**Unless otherwise specified, all mean scores refer to Mean by Task, on a scale of 1-5.*

Interactions: For the audio administration, the means were statistically significantly different, depending on who the examiner or rater was; this could be a difference of up to 0.51 points for the mean overall score (averaged over tasks: 1-5 pts). In Plot 1 below, the horizontal line is the Audio=F2F=0 line, which can be restated as the Audio=F2F line. The boxed and filled in portion of each group represents the 25th – 75th percentiles, the center 50% of the data. TTs for whom

CM was the F2F examiner were all above it, which means that 75% of the differences were positive (i.e., that the Audio score was higher than the F2F score). TTs for whom IPS was the F2F examiner were all below it, which means that 75% of the differences were negative (i.e., that the Audio score was lower than the F2F score). In Plot 2, the diagonal line is the Audio=F2F line. Points above it are TTs with Audio scores higher than F2F scores; points below it are TTs with Audio scores lower than F2F scores. The dots are color-coded by examiner / rater. The overwhelming majority of TTs with CM as F2F examiner lie above the line. The overwhelming majority of TTs with IPS as F2F examiner lie below the line.

In addition, there was a Test Form by Examiner interaction (see Table 4). There was a difference between the two examiners in the Library and Restaurant test forms. The Library test form + CM as examiner was associated with an additional decrease of 0.28 points, compared to IPS as baseline (-0.52, -0.04, p=**0.024**); the Restaurant test form + CM as examiner was associated with a decrease of 0.63 points, compared to IPS as baseline (-0.88, -0.39, p=**0.000**). This poses the practical problem that TTs who received the same sets of tasks could have different scores depending on the examiner (see Table 4 for a summary of statistically significant interactions).

Main Effects in Addition to Mode: Main effects in addition to mode are listed below, including test order within TT, sex, and task number. All figures in parentheses are lower 95% confidence interval bound, upper 95% confidence interval bound, p-value.

- Test order within TT: The first test for each participant was associated with a decrease of 0.11 points (-0.19, -0.023, p=**0.013**), compared to the second test.
- Sex: Female TTs (compared to males) were associated with a decrease of 0.27 points (-0.47, -0.078, p=**0.006**).
- Task number: Compared to the 5th task (baseline for comparisons), Task 2 was associated with a decrease of 0.17 points (-0.30, -0.042, p=**0.010**). This was the case regardless of test form; the interaction between test form and task number was not statistically significant.

Some of the aforementioned main effects were to be expected. There was a statistically significant difference in mean score by task between the first and second test administration within each TT. In other words, the second time that the TTs responded to the same set of prompts within a short period of time yielded higher scores than the first time. One likely reason for this was the practice effect: the TTs' familiarity with the structure of the test, with the instructions, and with the actual prompts. This phenomenon is well established in the literature (for a detailed discussion of this topic in the cognitive testing context, see Hausknecht, Halpert, Di Paolo, and Moriarty Gerrard, 2007, and for licensure contexts, see Raymond, Neustel, and Anderson, 2007). In addition, any anxiety the TTs may have experienced initially may have subsided by the time they repeated the test a few minutes later. Because this effect resulting from the test-retest design was to be expected, randomizing the order of the modes encountered by each TT allowed this effect to be evaluated and controlled for.

A main effect by sex was also found. There was a statistically significant difference in mean score by sex. Female TTs (compared to males) were awarded lower scores on the test overall. During this study, no reliable concurrent evidence of TT proficiency was gathered. Question 12 on the post-test survey did ask TTs to provide self-reported results on other English-language assessments (see Appendix C), but produced sporadic and unreliable responses. These responses offer no insight into the main effect by sex.

The final main effect found pertained to the second task (past-tense narrative) being associated with a lower score than the fifth task (persuasion). Upon further examination, Task 2 in Test forms 1, 2, and 3 was worded similarly: "Tell me about a time when you…" while Test form 4 was worded differently: "Tell me about a meal that you really enjoyed." It was noted during test administration that several less proficient TTs struggled with the wording of the formulaic sequence "Tell me about a time when you…," and frequently misunderstood it as a request for information about the time (hour) when something took place. The prompt "Tell me about a time when you visited a library," for instance, often elicited statements such as "Time? Tell me about time? It was 10 o'clock," followed by silence. Occasionally TTs asked for clarification by repeating the wording with a quizzical look and intonation. According to instructions in the Examiner Manual, however, clarification is not to be provided. TTs who requested clarification only received repetition of the prompt. The Task number effect was associated with all the test forms despite the fact that the wording of the second prompt

in test form 4 is more accessible to less able TTs ("Tell me about a meal that you really enjoyed") than was the wording of the other Task 2 prompts. This is probably due to a second contributor to the task effect: the TTs' avoidance of the simple-past verb tense, which the task had been specifically designed to elicit. Many TTs lost points because they started their responses in the past tense but after a sentence or two switched to a general description in the simple present. For instance, to the prompt "Tell me about a time when you visited a library" TTs would state, "It was yesterday, at 4 o'clock. I like to study at the library" and then proceed to describe the library in general terms. TTs were, therefore, penalized for not answering the question directly.

Additional Analysis for Research Question 1, looking at the 'pure' comparison of F2F and Audio total scores within (F2F) examiner

A follow-up analysis to the first analysis for research question 1 was conducted using a more direct 'pure comparison' between scores for the Audio and F2F modes within TTs. A 'pure' comparison can be made by looking at the differences in scores between the Audio and F2F administrations of the test, by examiner. Both sets of scores used were rated by the F2F examiner. This avoids examiner, rater, and mode interactions.

Analysis Method A paired t-test was conducted for TTs within each F2F examiner. The F2F and audio scores as rated by that examiner were used as the 'official' F2F and Audio scores (see Appendix F). Because task differences are not of interest for this analysis ('pure comparison'), the total score was used, summed over all five tasks (scores ranging from 5-25). In addition, a boxplot and scatterplot of the differences were constructed, again grouped by F2F examiner. The 'difference' was the Audio score minus the F2F score. Positive (negative) differences meant that the audio scores were higher (lower) than the F2F scores, within TT.

Results For both examiners, there was a statistically significant within-TT mean difference between the Audio and F2F scores (see Table 6). The direction of the mean differences varied between the two examiners (as was indicated by the results of the full linear mixed model previously used). For the examiner 'CM', the Audio scores were 1.46 pts (5-25 scale) higher than the F2F scores (95% CI: -0.91, 2.01; p=**0.000**). For the examiner 'IPS', the Audio scores were 1.16 pts (5-25 scale) lower than the F2F scores (95% CI: -1.79, 0.53; p=**0.001**).

Research question 2: Do CaMLA Speaking Test test takers prefer one mode of test delivery over the other?

The second research question asked whether CST TTs would prefer one mode of test delivery over the other. Two questions were asked pre-test about the TTs' preference for talking to a person or a computer (Q6, Q7); two parallel questions were asked post-test about the TTs' opinion on which mode they believed they had better performance (Q9, Q10). A fifth question (Q11) was also asked post-test, asking the TTs whether or not they felt that their performance would be the same for both modes of test administration. For the survey instrument, see Appendixes B and C.

- Q6 'When taking a speaking test, I prefer to talk to a person.' (Pre-Test)
- Q7 'When taking a speaking test, I prefer to talk to a computer.' (Pre-Test)
- Q9 'I think I did better on the test when I spoke to a person.' (Post-Test)
- Q10 'I think I did better on the test when I spoke to a computer.' (Post-Test)
- Q11 'I think my test scores on the two tests will be the same.' (Post-Test)

Compliance was extremely high: all respondents answered Q6; one respondent did not answer each of Q7, Q9, Q10, Q11 (out of 106 TTs). Q6, Q7, Q9 and Q10 allowed for one of three answers: 'yes', 'no', 'I don't know' (some respondents did not answer all questions). The fifth question, Q11 allowed for four answers: 'yes', 'I don't know', 'No-My score will be better when I talked to a computer' and 'No-My score will be better when I talked to a person'.

Answer: Before the test, the overwhelming majority of TTs said they preferred to speak to a person (see Table 7). Post-test a lesser majority claimed that they did better when they spoke to a person. The reason for the slight drop in preference for the F2F mode is unknown. It is possible that in hindsight TTs did not mind the Audio mode as much as they had thought they would. It is also conceivable that they preferred the F2F experience less than they had thought they might.

Research question 3: Is there any alignment between test takers' scores and preferred mode of test delivery?

Analysis Method Analysis of variance (ANOVA) was used to compare means (within examiner) of TT

Table 4: Statistically Significant Interactions

Rater by Mode Interaction				
	Mean	**95% Lower Bound**	**95% Upper Bound**	**P-Value**
Rater by Mode Interaction				
Difference: mean for Audio minus mean for F2F, for CM as rater	0.14	0.02	0.26	**0.019**
Difference: mean for Audio minus mean for F2F, for IPS as rater	-0.15	-0.27	-0.03	**0.014**
Examiner by Mode Interaction				
Audio mode with CM as examiner (vs. IPS as examiner and/or F2F mode)	0.22	0.05	0.38	**0.010**
Examiner by Sex Interaction				
Female with CM as examiner (vs. IPS as examiner and/or male TT)	0.52	0.34	0.69	**0.000**
Test Form by Examiner Interaction				
Test form='Restaurant' with CM as examiner (versus IPS and/or any other test form)	-0.28	-0.52	-0.04	**0.024**
Test form='Library' with CM as examiner (versus IPS and/or any other test form)	-0.63	-0.88	-0.39	**0.000**
Test Form by Sex Interaction				
Test form='Kitchen' with Female TT (vs. other test forms and/or male TT)	0.72	0.47	0.97	**0.000**
Test form='Library' with Female TT	0.51	0.26	0.76	**0.000**

Table 5: Rater by Mode Interaction: Overall Means by Task and Differences for TTs by Mode and Examiner Combination

Rater by Mode				**95% Confidence Interval**		
Rater	**Mode**	**Mean**	**Std. Error**	**Lower Bound**	**Upper Bound**	**P-Value**
CM	Audio	3.35	0.04	3.26	3.43	
	F2F	3.20	0.04	3.12	3.29	
IPS	Audio	2.89	0.04	2.81	2.97	
	F2F	3.04	0.04	2.95	3.12	
Difference: mean for Audio minus mean for F2F, for CM		0.14	0.06	0.02	0.26	**0.019**
Difference: mean for Audio minus mean for F2F, for IPS		-0.15	0.06	-0.27	-0.03	**0.014**

Table 6: Total score for within-TT differences between the Audio and F2F Scores, by Examiner (Audio Score minus F2F Score, each with range 5-25)

	Mean Difference (Audio Score minus F2F Score	St. Dev	Lower 95% Bound	Upper 95% Bound	t	df	Sig. (2-tailed)
Examiner (F2F) = CM	1.46	1.97	0.91	2.01	5.364	**51**	**0.000**
Examiner (F2F) = IPS	-1.16	2.31	-1.79	-0.53	-3.690	**53**	**0.001**

scores by mode (Audio and F2F), as well as the between-mode differences within TTs. The TTs were grouped by their responses to each of the five questions (five different, independent groupings). Because between-task comparisons were not of interest for this analysis, the total scores by mode were used (sum of the scores for each task, with a range of 5-25). The groups (within each examiner) were defined by the responses to the five pre- and post-test questions (with each question independently dividing the TTs into groups for that question). See Table 8 Summary of Scores by Mode, within Examiner CM and Table 9 Summary of Scores by Mode, within Examiner IPS.

Answer: Within examiner, there were no statistically significant differences between the TTs who gave different responses, for any of the five survey questions, for both examiners (see table 10 Summary ANOVA). Neither the preferences nor the predictions of the TTs were statistically significantly associated with differences in mean performance with either mode. The TTs' preferences and predictions were also not statistically significantly associated with relative performance between the two modes.

Research question 4: What factors do test takers claim affect their performance on both modes of the speaking test?

Analysis method: The fourth research question was qualitative. It was answered by analyzing the open-ended comment section on the post-study survey for themes pertaining to perceived factors influencing TTs' performance on the two modes of test delivery (see plots 1 and 2 on page 11). This section did not elicit comments about such factors directly; rather it asked for any: "Do you have any other comments about the two tests you took today?" (see question 13 in Appendix C). To answer the fourth research question, only those comments were analyzed and included in Table 8 that spoke to the TTs' perception of the factors influencing their performance on the direct and semi-direct versions of the test. All comments that were made at least once were reported and categorized by larger themes. Comments that were excluded, for instance, wished the researchers good luck with the study or expressed pleasure to be of assistance.

Answer: Thirty-four TTs volunteered comments about the factors that, in their opinions, affected their performance on the two versions of the CST (see Table 11). The TTs completed the survey without assistance from the researchers or the research assistants and were free to write as much or as little as they chose in the comments section of the post-test survey. Most chose to write brief phrases only. Regrettably, no further information is available from the TTs about the meaning of their comments, and the limited responses to this question make generalization impossible.

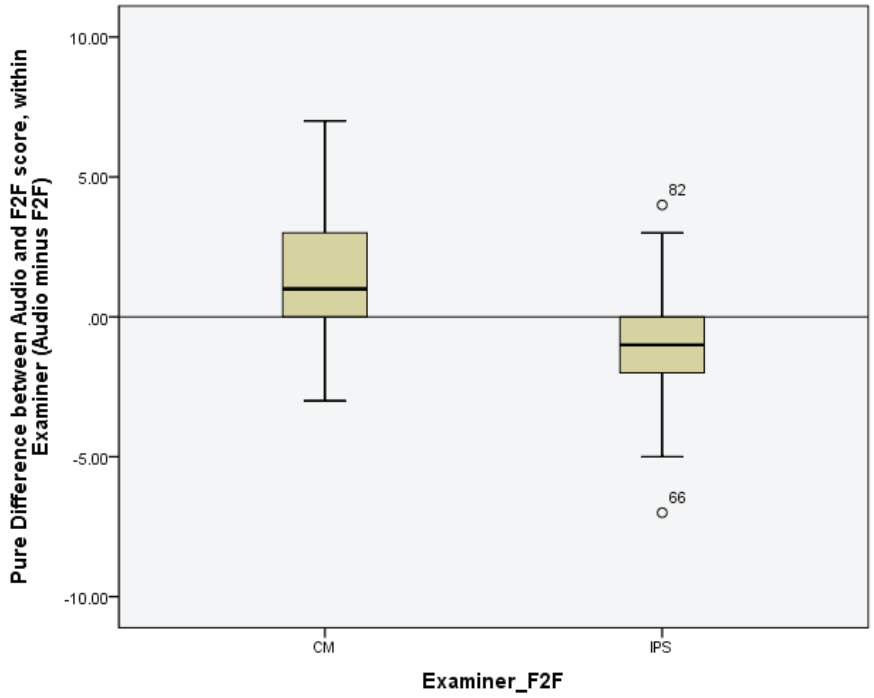
Additional Results

There were additional findings not directly covered under the four research questions. The statistical findings reveal that multiple variables impacted the scores to different degrees. Discussing them from the examiners' perspective enables a deeper exploration.

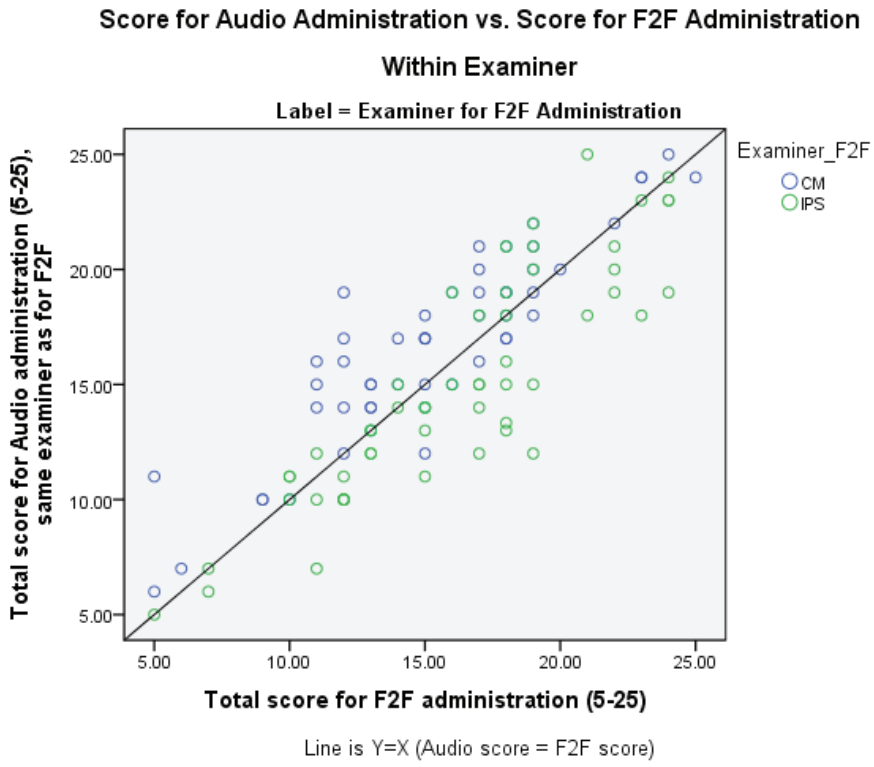
One finding that emerged in the study was the Test Form by Examiner interaction (see Table 4). There was a difference between the two examiners in the Kitchen (#4) and Restaurant (#3) test forms. The examiner-researchers noted during test administration that some TTs experienced difficulty responding to the "Kitchen" test form. Unlike the other stems in Task 1, which ask the TTs to describe the location depicted in the picture, this test form asks the TTs to describe an event: "Describe the family meal." Some TTs commented that the meal itself was not visible in the picture. They then proceeded haltingly to describe aspects of the meal that they could intuit. This discrepancy may have impacted the TTs' scores negatively.

Furthermore, as mentioned before, an Examiner by Mode and a Rater by Mode interaction was found (see Table 4). In order to investigate this further, we

Plot 1: Boxplot of within-TT differences (Audio Score minus F2F score) by Examiner (the '82' and '66' mark the case numbers of two outliers)



Plot 2: Audio Scores vs. F2F Scores within TT, by Examiner



calculated the differences between the examiners in the timing of the delivery of instructions and tasks in the audio-delivered and F2F tests. An examination of all eight audio-delivered tests (four tests with CM reading the instructions and four with IPS doing the same) and fourteen recordings of F2F tests (seven with CM and seven with IPS) reveals no statistically significant difference between the examiners in the timing of prompt delivery in the audio-delivered and F2F tests (see Table 12 'Examiner Speaking Times'). For some tasks CM spoke for a longer time and for others IPS did. However, intuitively there does seem to be a difference between how the two examiners spoke. CM spoke more in bursts, delivering the instructions and prompt more quickly but pausing for a longer period of time. On the other hand, IPS spoke more ploddingly, delivering the instructions and prompt more slowly, with shorter pauses. This also mirrored their respective natural ways of speaking. Quantitatively, the total amount of time each examiner took to speak may have been comparable; yet qualitatively, their speech delivery differed. On the CST, the tasks are not only read aloud by the examiner, but they are also printed on the TT's task card. It was noted, however, that during the test some TTs only listened and did not read along with the examiner. Others may have appeared to be reading along yet failing to keep up with the examiner's pace. In the literature review of this paper, studies were discussed about the speed of the aural stimulus and even one about the impact of pauses on less proficient TTs. However, the researchers have been unable to locate any studies into the exact aspect of the manner of delivery that appears to be present here.

It is important to note that the Examiner by Mode and Rater by Mode interactions occurred despite the fact that in this study the two researcher-examiner-raters were experienced oral examiners, who trained together and who remained in close contact with each other throughout the study. They did briefly discuss the optimal verbal delivery of the speaking prompt and instructions while calibrating. However, as analyses revealed, there were differences that may have contributed to the research findings.

One TT, for instance, noted that the live examiner's reaction propelled them to produce more speech. This could be interpreted in one of two ways. Either the examiner displayed a nonverbal reaction to the TT's response – such as eyebrows raised in surprise, nodding agreement – which the TT may have interpreted as encouragement to elaborate or the examiner remained expressionless despite the TT's answer, which the TT may have interpreted as a sign that the response was insufficient or inadequate. The examiners did attempt to keep the paralinguistic cues to a minimum, yet it is possible that they did reveal some reaction, as would be natural in a face-to-face encounter. Regrettably, all video cameras were directed at the TTs only and the examiners were not recorded. Therefore, confirming either hypothesis or forming a new one based on nonverbal cues is not an option in this study.

The study also revealed an issue with the wording of the instructions, especially evident in audio delivery. The TT instructions to the F2F test in Part II state that TTs will be told when to begin speaking. This, however, never happens. The TT's test card does not contain information about when the TT should begin speaking. This led one participant in the present study to remain silent throughout the fourth and fifth tasks during the audio-delivered test. During the F2F test, TTs had the possibility of confirming with the live examiner whether it was time to start speaking; also the Examiner Training Manual states that if after a specified amount of time TTs do not begin speaking, the examiner should prompt them to do so (CaMLA, 2014, p. 7). This prompting and verifying, however, is only possible F2F. In the present study, one examiner (IPS) prerecorded the audio tasks exactly as written and administered the audio recording to a group of TTs, and one TT remained silent during Tasks 4 and 5, awaiting instructions to begin. Having learned from this, at subsequent test administrations, IPS and her research assistant forewarned IPS's TTs to begin responding to the tasks immediately and not expect to be prompted to begin. The other examiner (CM), who had not yet recorded her outgoing prompts added the command "Begin" at the end of every task. In both modes of delivery, she ended all her prompts with "Begin." CM did not encounter similar problems during the study. It is conceivable that such a difference between the two ways of delivering the instructions could have contributed to the Examiner by Mode interaction. It is also possible that this may have contributed to the Rater by Mode interaction if the two raters interpreted silence differently. In IPS's tests, TTs may have remained silent after the instructions were read either because of preparing a response or because of confusion due to ambiguous instructions preceding Tasks 4 and 5. In CM's tests, TTs were more likely to have remained silent due to preparing a response. It is imaginable that the raters may have held silence against

the TTs and lowered their scores as a result, though neither CM nor IPS recalls doing so.

Under ideal circumstances, both the test delivery and the ratings are standardized: the examiner's speech rate is controlled and the raters are centralized with one rater per institution. Attaining such standardization becomes a challenge in a test for institutions, however, because of limited time and resources. Controlling the mode of delivery with clearly scripted instructions can mitigate some unwanted effects. Further, if CaMLA chose to offer the CST as a semi-direct, audio-delivered rather than direct, F2F test, all prompt delivery could be standardized. The voice talent could be given specific instructions on the precise speed and manner of delivery for all versions of the test, thereby increasing test reliability. This way only the rating of the tests would remain in the hands of institutional users.

Summary of Results

In this study we sought to investigate whether the mode of delivery, i.e., direct (F2F) and semi-direct (audio-delivered), influenced the TTs' scores on the CST. Overall, there is no difference between the modes of delivery, but we were unable to answer research question 1 effectively without considering a number of interactions surrounding the modes. These interactions include Rater by Mode, Examiner by Mode, Examiner by Gender, Test Form by Examiner, Test Form by Gender. On the one hand, the interactions surrounding the mode must be resolved before the test mode can be switched from F2F to audio. On the other hand, some of these interactions would not have occurred had the mode of delivery been limited to audio only. The study also sought to determine if performance correlated to preferred mode of delivery. Prior to taking the two modes of the test, the TTs overwhelmingly stated preference for the F2F version. During the test, the TTs were randomized to take either the F2F or the audio version of the test first. After the two versions of the test when asked to predict whether they would receive a higher score on the F2F or audio version, the largest group of TTs (48% overall) still believed that they would score higher in person. Further analyses revealed that the TTs' responses to the survey questions about their preferences/perceptions of performances had no statistically significant association with their mean performance on either mode, or their relative performances between the two modes. Finally, the study investigated self-reported factors that affected TTs' performance on the CST. In this regard, comments were limited. Factors that were elicited included nervousness, response time and lack of preparation time, sequencing of the two tests, wording of prompts, and examiner influence.

Conclusion, Implications, Future Research

It can be concluded that, while the mode of delivery of the CST does not appear to affect scores overall, various interactions do have an effect on TTs' scores by mode. A number of these interactions could be eliminated if the mode of delivery were limited to audio only. In addition, while the CST is designed to be highly practical and valid, this study found that an audio-delivered CST would increase reliability in terms of delivery.

The findings of this study reveal that even trained and experienced speaking examiners who have undergone the recommended amount of training and calibration can vary statistically significantly in their interpretation of the CST rating scale when scoring the same TT's performance. As a test for institutions, the CST is likely to be rated by a variety of raters at institutions. In educational settings, placement and exit tests are usually given to a large number of students within a short period of time. Administering and scoring the tests in a timely manner will require multiple calibrated raters. It is doubtful that institutional raters would achieve lower Examiner or Rater effect than the two researchers in this study.

Establishing prompt equivalency was not the main focus of this study. The research did reveal a possible prompt inequivalency problem, nevertheless. CaMLA could increase the equivalency of the test forms by eliminating low-frequency formulaic sequences (such as "tell about a time") from tasks aimed at TTs with lower proficiency, such as Tasks 1, 2, and even 3. Robust pilot testing may therefore ensure greater prompt equivalency (see Weir and Wu, 2006).

There appear, thus, to be advantages of making the CST available in a semi-direct format. One such advantage to both the test producer and test users is economic in nature. Audio delivery would potentially make the test less costly to deliver and hence more attractive for institutions. The live examiner can be replaced by a proctor, or invigilator, who can ensure that test delivery is standardized. In fact, with sufficient facilities and equipment, multiple concurrent test administrations can be offered. Sound files can be

transferred to the raters in a remote location, who can then rate at their convenience. Thus, the criterion of feasibility, one of the four considerations for evaluating the choice between direct and semi-direct modes, is met (Shohamy, 1994). In addition to these economic benefits is increased test security. Administering the test to multiple students simultaneously would increase the likelihood that the speaking prompts would remain confidential longer. Although TTs who take the test sequentially can still share answers, TTs who take the test simultaneously cannot, leading to increased confidence in the scores. As the test gains popularity in countries where English is not spoken widely, the semi-direct mode of delivery would allow TTs to listen to the tests being delivered in the way that CaMLA has deemed optimal.

A further potential benefit of an audio-delivered mode is that the roles of the examiner and rater can be decoupled and the reliability of ratings increased. This releases the rater from the duty of multitasking and interpreting examiner instructions. In the current F2F version of the CST, during the test the live examiner must read the instructions and tasks verbatim, monitor the time, and concomitantly assign a holistic score to each task based on a rating scale consisting of multiple evaluation criteria. This multitasking may influence the accuracy of the awarded scores even for trained and experienced examiner-raters. A semi-direct test-delivery format would allow for the standardization of prompt delivery and allow the rater to concentrate only on rating, which too would improve reliability.

Therefore, should CaMLA choose to provide recorded versions of the test, it may be prudent to commission one voice talent to record all the outgoing audio according to carefully crafted recording specifications. We recommend that these instructions extend to not only the total words per minute spoken but also the lengths of pauses between sentences; otherwise, an adjustment factor may need to be made based on who the voice talent is and his/her manner of speaking. On the other hand, the standardization of the accent may hold disadvantages as well if the users' local variety of English differs considerably from the CaMLA standard. This may cause the TTs to have difficulty understanding the recording. Test users will need to consider this possibility and mitigate its effects by exposing TTs to the language variety of the voice talent prior to TTs' taking the test (Winke & Gass, 2013).

This study has the possible benefit of demonstrating that an inexpensive and non-sophisticated audio-delivered version of the CST can replicate the face-to-face test. However, the fact that the scores in this experiment were significantly different does not provide evidence that a more technologically sophisticated mode of delivery or greater consistency between the examiner-raters would not have produced scores that agree much more closely. Since the study yielded different results for the two modes, CaMLA could choose to emphasize the importance of face-to-face delivery in their marketing.

A limitation of the study is that since the protocols for F2F and Audio examinations could not be identical, it is possible that those protocol differences had an effect on the outcomes. Still, the findings should provide insight into the option of offering an audio-delivered mode for the CST. If the audio-delivered mode of test delivery had yielded similar test scores, this would have demonstrated the reliability of the test across modes, akin to studies by Ginther (2012); Stansfield & Kenyon (1992); and Thompson et al. (2016). The findings of this study, however, suggest that further research is needed into several areas.

Future research on the CST can extend in a variety of directions. Both the existing and new datasets would warrant further investigation of the CST. With the existing dataset, this quantitative investigation of the equivalence of TTs' scores can be followed by a qualitative investigation of the language produced by the TTs in each mode. O'Loughlin (2001) found a qualitative difference between the way TTs spoke in the direct and semi-direct versions of the Australian Assessment of Communicative English Skills. With the live examiner present, the TTs produced more conversational, interactional language, with lower lexical density estimates. In the recorded version, however, the language seemed more monologic. The CST is not designed to test interactional ability. Therefore, it is not expected that the lexical density of the TTs' speech would differ by mode. However, if it does, this finding would corroborate O'Loughlin's that such difference is likely due to the mode of delivery. In addition to the lexical density, direct and semi-direct tests have also elicited differences in fluency and filler pauses (Conrad et al., 2007, as cited in Vogl, 2013). A closer look at the language produced by the TTs would yield additional insights. Taking this a step further, it could be very interesting to compare performances of TTs on the CST with a more interactional speaking test to determine if results correlate.

In the literature review of this paper, studies were discussed about the overall delivery speed of the aural

stimulus and one about the impact of pauses on less proficient TTs. In this study, however, the total length of time that each task was read was comparable between examiners, yet the ratio of pause to speech differed between examiners. This line of inquiry was not found in the literature and may be worth investigating further with the remaining videos to understand whether a deeper analysis might shed light on the Examiner by Mode interaction detected in this study.

A further line of inquiry that may prove fruitful entails examining rater drift. Studies have suggested that raters' scores become less accurate over time (see, for instance, Lumley & McNamara, 1995). In the present study, all 106 tests were administered within a three-week window; however, three months passed between rater training and the rating of the final test. The two raters spent the vast majority of this time either examining or rating and relied on the rating scale while rating all the tests. Nevertheless, they may have become uncalibrated. Rescoring the existing tests would yield more information about rater drift.

Finally, in order to examine whether any bias – in terms of gender or L1, etc. – existed in the present study, a future study might investigate automated/computerized scoring of the recorded performances. The results of computer vs. human scoring could be compared to gain further insight into the possibility of implicit bias and any role it may have played.

Extending the research beyond the existing dataset, it could also be useful to repeat this study using video recorded delivery in lieu of audio delivery. As mentioned in the literature review of this paper and corroborated by this study, TTs overwhelmingly prefer to speak to a person. They find talking to a voice recorder unnatural and appreciate the paralinguistic clues of F2F interaction. It might prove to be fruitful, however, to see whether video-recorded input material could contribute to the literature on how body language may affect the interpretation of prompts.

The results of this study contribute as well to the broader language assessment community in that the findings add to the existing literature on feasibility, as well as desirability of delivering speaking tests without a live interlocutor in this context.

Acknowledgements: We gratefully acknowledge Cambridge Michigan Language Assessments and Eastern Michigan University for their financial support; research assistants Wesley Chen, Haley Gardner, Lauren Prebenda, and Martina Syrova for their administrative support; and the 106 test takers for their assistance.

References

Audacity (n.d.) http://audacityteam.org/

Blau, E.K. (1990). The effect of syntax, speed, and pauses on listening comprehension. *TESOL Quarterly*, 24(4), 746-753.

Brown, A. (1993). The role of test-taker feedback in the test development process: test-takers' reactions to a tape-mediated test of proficiency in spoken Japanese. *Language Testing*, 10(3), 277-301.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing*, 20(1), 1-25.

Brunfaut, T. & Revesz, A. (2015). The role of task and listener characteristics in second language listening. *TESOL Quarterly*, 49(1), 141-168.

CaMLA (n.d.) http://www.cambridgemichigan.org/institutions/products-services/tests/placement-progress/camla-speaking-test/ Accessed November 29, 2015.

CaMLA (2014). CaMLA Speaking Test: Test Materials. Testing Coordinator's Manual and Examiner Training Manual.

CaMLA (2015). Linking the Common European Framework of Reference and the CaMLA Speaking Test. CaMLA Technical Report. https://www.cambridgemichigan.org/wp-content/uploads/2015/12/CaMLA-Speaking-Test-LinkingStudy.pdf. Accessed January 15, 2015.

Clark, J.L.D. (1979). Direct vs. semi-direct tests of speaking ability. In E.J. Briere & F.B. Hinofotis (Eds.), *Concepts in language testing: some recent studies* (pp. 35-49). Washington, DC: TESOL.

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

Ginther, A. (2012). Assessment of speaking. In C.A. Chapelle (Ed), *The Encyclopedia of Applied Linguistics*. Wiley-Blackwell.

Hausknecht, J.P., Halpert, J.A., Di Paolo, N.T., & Moriarty Gerrard, M.O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology*, 92(2), 373-385.

Humphreys, P., Haugh, M., Fenton-Smith, B., Lobo, A., Michael, R., & Walkinshaw, I. (2012). Tracking international students' English proficiency over the first semester of undergraduate study, *IELTS Research Reports Online Series*, 1, 4-41. Retrieved from: http://gallery.mailchimp.com/d0fe9bcdc8ba233b66e1e0b95/files/Humphreys2012_ORR.pdf

Kenyon, D. M. & Tschirner, E. (2000). The Rating of Direct and Semi-Direct Oral Proficiency Interviews: Comparing Performance at Lower Proficiency Levels. *The Modern Language Journal*, (84), 85–101. doi: 10.1111/0026-7902.00054

Kostin, I. (2004). *Exploring item characteristics that are related to the difficulty of TOEFL dialogue items* (TOEFL Research Report No. RR-79). Princeton, NJ: Educational Testing Service.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: implications for training. *Language Testing*, 30(1), 53-72.

Muljani, D., Koda, K., & Moates, D.R. (1998). The development of word recognition in a second language. Applied Psycholinguistics, 19, 99-113.

O'Loughlin, K. (2001). *The equivalence of direct and semi-direct speaking tests*. Cambridge: Cambridge University Press.

Qian, D.D. (2009). Comparing direct and semi-direct modes for speaking assessment: Affective effects on test takers. *Language Assessment Quarterly*, 6(2), 113-25.

Raymond, M.R., Neustel, S., & Anderson, D. (2007). Retest effects on identical and parallel forms in certification and licensure testing. Personnel Psychology, 60, 367-396.

Revesz, A., & Brunfaut, T. (2013). Text characteristics of task input and difficulty in second language listening comprehension. *Studies in Second Language Acquisition*, 35, 31-65.

Rosenhouse, J., Haik, L., & Kishon-Rabin, L. (2006). Speech perception in adverse listening conditions in Arabic-Hebrew bilinguals. *International Journal of Bilingualism*, 10(2), 119-135.

Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing*, 11(2), 99-123.

Thompson, G.L., Cox, T.L., & Knapp, N. (2016). Comparing the OPI and the OPIc: The effect of test method on oral proficiency scores and student preferences. *Foreign Language Annals,* 49(1), 75-92.

Vogl, S. (2013). Telephone versus face-to-face interviews: Mode effect on semistructured interviews with children. *Sociological Methodology*, 43(1), 133-177.

Weir, C., & Wu, J.R.W. (2006). Establishing test form and individual task comparability: A case study of a semi-direct speaking test. *Language Testing*, 23(2), 167–197.

Winke, P., & Gass, S. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, 30(2), 231-252.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Appendixes A-F

Appendix A: Demographic Make-up of Test Takers

First Languages (n = 106)

24 = Chinese

21 = Arabic

14= Japanese

11 = Korean

8 = Spanish

5 = Portuguese

3 = Telugu

2 = Albanian

2 = French

2 = German

2 = Hindi

2 = Hungarian

2 = Vietnamese

1 = Croatian

1 = Czech

1 = Hebrew

1 = Italian

1 = Marathi

1 = Polish

1 = Romanian

1 = Russian

Appendix B – Pre-Test Survey

Participant number: _____

1. I agree to participate in this survey
Yes / No (circle one)
2. I am
male/ female (circle one)
3. My home country is (please write below)
4. My native language is (please write below)
5. My English language proficiency is (please circle one)
beginner / high beginner low intermediate / high intermediate low advanced / advanced
6. When taking a speaking test, I prefer to talk to a person. (please circle one)
Yes / No / I don't know
7. When taking a speaking test, I prefer to talk to a computer. (please circle one)
Yes / No / I don't know
8. Do you have any comments?

Appendix C – Post-Study Survey

Participant number: _____

9. I think I did better on the test when I spoke to a person. (please circle one)
Yes / No / I don't know
10. I think I did better on the test when I spoke to a computer. (please circle one)
Yes / No / I don't know
11. I think my test scores on the two tests will be the same. (please circle one)
Yes / No-My score will be better when I talked to a person / No-My score will be better when I talked to a computer / I don't know
12. If you have taken an English test, which test was it and what was your score?
TOEFL score _____ IELTS score _____ MELAB score _____ TOEIC score _____ ECPE _____ ECCE _____ Other test's name _____ score _____
13. Do you have any other comments about the two tests you took today?

Appendix D1-3 – Interrater Reliability, Intrarater Reliability, Agreement by CEFR Band

The **Interrater Reliability** was good. The scores reflect how well the rater's scores for a TT within each task correlated with the other rater's scores for that TT (see Table D1).

Table D1 – Interrater Reliability

Task	Reliability Between Raters
Task 1	0.826
Task 2	0.887
Task 3	0.860
Task 4	0.865
Task 5	0.841

The **Intrarater Reliability** was better still. These scores reflect how well each rater's scores for a TT for tasks 1-5 correlated with each other (see Table D2).

Table D2 – Intrarater Reliability

Rater	Reliability within Rater
CM	0.933
IPS	0.941

Note that both reliabilities are correlations; they measure how well scores between/within raters track each other. A consistent, unvarying difference between two raters (or within a rater) would still lead to a very high reliability. For that reason, the agreement between raters should be evaluated directly.

The Agreement by CEFR Band reveals the exact and adjacent agreement between the two raters' F2F scores on the CEFR. This level of agreement, while in line with the findings of other studies, reveals an area of potential improvement (see Table D3).

Table D3 – CEFR Bands Assigned Based on Face-to-Face Mode

		CEFR Band Assigned by CM, based on F2F Mode				
		A2	**B1**	**B2**	**C1**	**Total**
CEFR Band Assigned by IPS, based on F2F Mode	**A2**	10	4	1		**15**
	B1		37	9	1	**47**
	B2		5	18	3	**26**
	C1			3	14	**17**
	Total	**10**	**46**	**31**	**18**	**105**
	For 79 TTs (75%), the band assignments were the same.					
	For 24 TTs (23%), the band assignments were adjacent.					
	For 2 TTs (2%), the band assignments were discrepant, neither the same nor adjacent.					
	Note - due to missing information, the responses for one TT that are missing for one or both raters have been excluded from the calculations. Also, the scores used to calculate the CEFR bands in this table only contain the F2F score for each examiner.					

Appendix E: Test Takers' Total Scores and Band on the CaMLA Speaking Test

As mentioned, the CST is aligned with the CEFR (CaMLA, 2015). According to the published cut scores (CaMLA, 2015), the TTs of this study achieved the following proficiency bands on the CEFR: A2-C1 (see Table 13 for complete score breakdown). The total scores were calculated by only considering scores awarded by each TT's examiner. One set of scores comprised the F2F score awarded by the live, synchronous examiner, as this is the current scoring practice of the CST; the other set of scores comprised the audio score awarded by the same rater who was the F2F examiner. Therefore, each examiner's ratings for the TTs that she examined and rated live and from audio recording were considered, and all the ratings assigned by the other rater were omitted for the purpose of calculating the TTs' total scores. It is important to note that the CST was not designed to target A1 or C2 level TTs. Thus, it is possible that TTs who received a total of 5 or 25 points, respectively, could be below the A2 or above the C1 levels.

Table 13: Test Takers' Total Scores and CEFR Band on the CaMLA Speaking Test

	Examiner_F2F							
	CM				IPS			
	Band				Band			
Synchronous Total	A2	B1	B2	C1	A2	B1	B2	C1
5	2				1			
6	1							
7					2			
9	2							
10	1				3			
11		3				3		
12		5				5		
13		5				4		
14		2				2		
15		7				4		
16		2				2		
17			5				5	
18			6				7	
19			5				5	
20			1					
21								2
22				1				3
23				2				2
24				1				4
25				1				
Total	**6**	**24**	**17**	**5**	**6**	**20**	**17**	**11**

Appendix F: Final Score (F2F, Audio) for each TT

Table of final scores for each TT organized by examiner who conducted the F2F administration of the test.

Test Taker ID Number	Examiner (for F2F administration)			
	CM		IPS	
	Total score for F2F administration (5-25)	Total score for Audio administration (5-25), same examiner as for F2F	Total score for F2F administration (5-25)	Total score for Audio administration (5-25), same examiner as for F2F
1			13	13
2	12	17		
3	11	15		
4	12	16		
5	17	18		
6	15	15		
7	15	14		
8			15	13
9			24	19
10			21	18
11			18	13
12	9	10		
13			18	18
14	13	14		
15	19	20		
16			17	14
17			18	13
18	14	15		
19	11	16		
20			10	10
21			16	15
22	12	19		
23			15	14
24			19	12
25			13	12
26			12	10
27			12	11
28	11	14		
29	13	14		
30	12	14		

31			10	11
32			12	10
33			15	11
34			17	15
35			14	14
36	15	17		
37	23	24		
38	25	24		
39	16	15		
40			11	10
41			23	18
42			17	15
43	22	22		
44	18	17		
45	18	18		
46	13	15		
47			14	15
48			11	12
49	17	19		
50	17	21		
51			24	23
52	19	21		
53	9	10		
54			12	10
55	10	10		
56			21	25
57			12	10
58			19	20
59	18	21		
60	23	24		
61	20	20		
62	19	22		
63	17	20		
64	24	25		
65			11	7
66	5	11		
67			7	6
68			5	5

69			7	7
70	19	18		
71			24	23
72	18	19		
73			19	15
74	5	6		
75	6	7		
76			24	24
77			18	21
78			19	22
79	15	17		
80	15	18		
81			18	19
82	16	19		
83			22	21
84	15	17		
85			18	16
86	14	17		
87	18	17		
88	17	16		
89			23	23
90	15	12		
91	12	12		
92			15	14
93	13	13		
94			17	18
95			22	19
96	19	19		
97	13	15		
98	18	19		
99			16	19
100			18	15
101			22	20
102			13	12
104			10	11
106			19	21
107			17	12
108			13	13