COMPARING EXAMINEE ATTITUDES TOWARD COMPUTER-ASSISTED AND OTHER ORAL PROFICIENCY ASSESSMENTS

Dorry M. Kenyon and Valerie Malabonga
Center for Applied Linguistics

ABSTRACT

This article reports the results of a study of examinee attitudinal reactions to taking different formats of oral proficiency assessments across three languages: Spanish, Arabic, and Chinese. All 55 students in the study were administered both the tape-mediated Simulated Oral Proficiency Interview (SOPI) and a new Computerized Oral Proficiency Instrument (COPI). In addition, the 24 students participating in the Spanish study were administered the face-to-face American Council on the Teaching of Foreign Languages (ACTFL) Oral Proficiency Interview (OPI). Participants were graduate and undergraduate students taking language courses at their universities. The order of test administration was based on self- and teacher-assessed proficiency levels and experience in learning the target language. After each test, the learners completed a Likert questionnaire on six aspects of their attitudes towards and perceptions of that test. After finishing all tests and questionnaires, they were requested to compare the two technology-mediated tests directly on those six aspects. This article presents the examinees' responses on those questionnaires, focusing primarily on differences between the two technology-mediated tests. It was found that the adaptive nature of the COPI allowed the difficulty level of the assessment task to be matched more appropriately to the proficiency level of the examinee. This led examinees, particularly those at the lower proficiency levels, to feel the COPI was less difficult than the SOPI. In most other aspects, the two tests were rated similarly. For the Spanish students, the two technology-mediated tests were rated similarly to the OPI, though the OPI appeared to them to be a better measure of real-life speaking skills.

How do examinees react to an adaptive, computer-administered oral proficiency assessment? This article reports on examinee perspectives towards a prototype of the Computerized Oral Proficiency Instrument (COPI). The COPI was developed by the Center for Applied Linguistics (CAL; http://www.cal.org) under a research grant from the International Research and Studies Program of the United States Department of Education. Developed for Spanish, Arabic, and Chinese, the COPI is an adaptation of the tape-mediated Simulated Oral Proficiency Interview (SOPI), originally developed at CAL in 1985 (Clark, 1988). The primary goal of the COPI research project was to examine the effects of using multimedia computer technology to enhance the SOPI by giving examinees more control over various aspects of the testing situation and by increasing raters' efficiency in scoring the test. The project compared examinee performance and affect, and rater efficiency and affect, between the COPI and the American Council on the Teaching of Foreign Languages (ACTFL; http://www.actfl.org) Oral Proficiency Interview (OPI) for Spanish, and the COPI and SOPI for Spanish, Chinese, and Arabic. This article compares examinee reactions to the COPI, the SOPI, and the OPI.

BACKGROUND

In the United States, the major procedure used to assess the oral language proficiency of foreign and second language learners is the *Oral Proficiency Interview* (OPI). The most well-known OPI is the ACTFL OPI, which is described as "a standardized procedure for the global assessment of functional speaking ability" (Swender, 1999, p. 1).

The OPI assessment procedure has been in use since the 1950s. Originally developed by the Foreign Service Institute to assess the speaking proficiency of people serving in the United States Diplomatic Corps, the procedure became widely used in various U.S. government agencies. In the late 1970s and early 1980s, the OPI procedure was introduced to the academic world through the work of the Educational Testing Service and the ACTFL. Through ACTFL's efforts, thousands of educators have been trained in the procedure (Clark & Clifford, 1988).

In the ACTFL OPI, a trained and certified interviewer elicits speech performances through a procedure that has a standardized format. In this procedure, the interviewer adapts the difficulty of his or her questions to the candidate's proficiency level, as defined by the ACTFL *Speaking Proficiency Guidelines* (American Council on the Teaching of Foreign Languages, 1986, 1999). The interviewer also considers the examinee's experience and interests, as revealed in previous answers, in formulating questions. The procedure begins with a *warm-up* phase, during which the interviewer establishes rapport with the candidate and makes an initial evaluation of his or her proficiency level. The interviewer continues the procedure, using questions to elicit demonstration of the candidate's control of functions, contexts and accuracy. *Level checks* help the interviewer ascertain that the examinee can function at a given ACTFL proficiency level. *Probes* are used to determine whether or not the candidate can demonstrate sustained performance of the requirements for the next higher level. The interviewer may use a role-play to elicit a performance more appropriate to language use outside the interview situation. When confident that enough of a speech performance to determine the candidate's proficiency level has been elicited, the interviewer concludes the procedure with a *wind-down* phase that eases the candidate out of the testing situation. Interviewers must do these tasks efficiently, both evaluating candidate speech and determining the next question, to keep the interview within a practical period of time.

Each interview is tape recorded for later verification. Interviewers listen to the tape of the interview to confirm their final proficiency rating, and official interviews are blindly double-rated by another ACTFL-certified rater. While much has been published on the OPI as a testing technique, the most complete and up-to-date information on the ACTFL OPI can be found in Swender (1999).

While standardized in format, the OPI procedure can be flexible and adaptive. The content and level of difficulty of the questions can be tailored and personalized to each individual examinee. The interactive and dynamic process of the OPI allows the interviewer to carry out an extensive assessment of language skills by generating flexible questions and controlling the whole interview process in a manner and at a level appropriate to the examinee.

To be successful, good OPIs require very skilled and well-trained interviewers. In addition, the interviews are very labor-intensive to conduct as they must be done face-to-face. As an alternative to the OPI, CAL developed the self-administering, tape-mediated Simulated Oral Proficiency Interview (SOPI). The original impetus for its development was the need to assess oral proficiency in the Less Commonly Taught Languages (LCTLs). In the United States, LCTL programs are generally very small, geographically dispersed, and have limited resources. As the proficiency movement began to spread in the United States, it seemed that there would be few people in the LCTLs who could be trained and certified to give the ACTFL OPI. It would be impossible for these few individuals to conduct a face-to-face interview at every location where assessment was needed. To address this need, CAL developed SOPI -- a test that can be administered by anyone, even in small language programs that lack access to an OPI-certified interviewer. The recorded examinee response tapes could then be mailed to ACTFL-certified testers. The first SOPI was developed for Mandarin Chinese (Clark, 1988). Development of SOPIs for Portuguese (Stansfield, et al., 1990), Hebrew (Shohamy, Gordon, Kenyon, & Stansfield, 1989), Hausa (Stansfield & Kenyon, 1993), and Indonesian (Stansfield & Kenyon, 1992a) followed. More recently, CAL has developed SOPIs in Japanese, Arabic, Spanish, French, German, and Russian, with accompanying self-instructional rater training kits (Kenyon, 1997; Norris, 1997).

The SOPI procedure emulates the OPI as closely as is practical in a tape-recorded format (Kuo & Jiang, 1997; Stansfield, 1990). The SOPI elicits the candidate's speech performance by means of carefully constructed tasks presented through an audio tape and a printed test booklet. These tasks are constructed by professional language evaluators using the ACTFL *Guidelines* and are specifically designed to elicit speech performances ratable according to the criteria of the ACTFL *Guidelines.* Each task is defined by the proficiency level it seeks to elicit in terms of speech functions (such as "asking questions" or "giving a simple description"), discourse types, content, and contexts appropriate for the target level of the task. A typical full-length SOPI consists of 15 tasks, though for examinees who have had less opportunity to develop their oral proficiency, only the first 7 tasks need to be administered. The typical full-length form lasts about 45 minutes, with candidates speaking for about 25 minutes. The typical short form (i.e., the first seven tasks) lasts about 25 minutes, with candidates speaking for about 10 minutes.

A SOPI may be administered to a group in a language lab or individually using two tape recorders. A master tape plays the test directions, to which the candidates listen while following along in their test booklets. The candidates' responses are recorded on a second, blank cassette, called the *examinee response tape.*

As with performance on the OPI, performance on the SOPI is rated using the ACTFL *Guidelines.* While the SOPI can be administered by anyone, including persons who do not speak the language of the test, it clearly needs to be scored by trained raters. CAL conducts rater-training workshops and also has self-instructional rater-training kits available for SOPIs in seven languages.

While some aspects of the OPI are flexible (e.g., the precise topics covered, which are often brought up by the examinee), there are several features of the SOPI that are fixed for each candidate. For example, the content covered in any SOPI form is the same for all candidates, and candidates have no choice as to which broad topics they are going to talk about. Each task covers a different broad content area, however, so that special knowledge in any area, or lack of it, generally should not unduly influence a candidate's overall performance. On the SOPI, candidates have a fixed amount of time to think about their response, between 15 and 45 seconds, depending on the complexity of the task. They also have a fixed amount of time to respond to the task, usually between 45 seconds and 1 minute and 45 seconds, depending upon the complexity of the task.

Despite these differences between the OPI and SOPI, correlations between proficiency level ratings received on the two tests are very high, averaging .92 (Stansfield & Kenyon, 1992b). In a recent study of lower proficiency German students, 90% received the same rating on the SOPI and the OPI (Kenyon & Tschirner, 2000).

Other technologically mediated modifications to the OPI have also been developed which are similar to the SOPI. The Language Acquisition Research Center at San Diego State University has developed a video-based interview, the *Video Oral Communication Instrument* (VOCI), which is similar to the SOPI. The main difference is that the test tasks are presented via videotape rather than audiotape. The *d-VOCI* has also appeared, in which the video segments have been digitized so that the test may be administered.

THE NEXT GENERATION

With the COPI, CAL is now experimenting with a new generation of technologically enhanced oral proficiency assessments. Compared to the tape-mediated SOPI, the computer makes it possible to give examinees more control over several aspects of the test. For example, tape length is no longer a problem; examinees have control over the time they use to prepare and give their responses.

Prior research (Clark, 1988; Stansfield et al., 1990) indicates that examinees sometimes feel nervous taking the SOPI, often due to a sense of lack of control over the timed aspects of the testing situation. Some examinees report that they feel this nervousness prevented them from giving their best

performance. This lack of control can also be associated with negative affect towards the testing situation. The extent to which a learner feels in control in a given situation is thought to contribute to the learner's affect towards that situation. Schumann (1997) reported on variables that influence a language learner's affective response to a stimuli (e.g., a test). He defined *control* as the ability to be in charge of the outcome of a situation or the ability to direct the effort one can apply to a situation.

Because computers can store a large underlying pool of tasks, they can be adaptive in the sense that they can be programmed to administer tasks that are more appropriate to the examinees' current level of proficiency. This may be expected to facilitate a more positive affect towards the test in a manner analogous to Schumann's (1997) theory about "expectancy" in language learning. Schumann describes *expectancy* as a learner's assessment of the likelihood of success vis-à-vis judgments of task difficulty, effort, and help available. In a testing situation, if examinees perceive that they will be (or are) successful in accomplishing the test task, they may be more likely to have positive feelings towards the test.

With computer adaptive tests, tasks can also be developed to cover a wide range of interests and background experiences from which examinees can choose the ones that most appeal to them. Again, this may produce more positive affect towards the test when we consider Schumann's (1997) description of *relevance* in language learning. If a language learning activity is related to a learner's goals, needs, and values, the learner is more likely to have a positive appraisal of the activity. Likewise, in a testing situation, if examinees can choose assessment tasks that most appeal to them, they may be more likely to have a positive view of the assessment instrument.

Ultimately, the goal of providing greater adaptiveness, examinee control, and examinee choice of tasks in a technology-mediated oral proficiency assessment is to help them give their best performance (Swain, 1985). To the extent that technology produces negative affect in examinees, performance may be negatively affected.

The COPI is built around a large, underlying pool of tasks covering a wide variety of content areas and topics. For each response, the COPI presents examinees three alternatives, written in the form of a simple sentence, from which the examinee must choose one. Examinees can choose those tasks they would prefer to talk about. Tasks are coded by content and an underlying algorithm ensures that examinees are exposed to a wide variety of topics.

The COPI also allows examinees to have control of the time that they take to prepare for and respond to a task. While a maximum time limit needs to be enforced to make sure the testing process continues, that limit is generous enough to ensure that the vast majority of examinees never feel they run out of time.

The COPI also allows examinees some choice in the difficulty level of tasks presented to them. Following the initial directions, the COPI administers and scores a self-assessment of the learner's language proficiency. Based on this information, the COPI then suggests that the examinee begin the test with a task at one specific level of four available. The four levels correspond to the main levels of the ACTFL scale: Novice, Intermediate, Advanced, and Superior. These levels refer to the global tasks or functions that speakers can handle, the contexts in which they can effectively communicate, the content about which they can communicate, and the accuracy with which they communicate. The range covers Novice-level speakers, who are able to produce memorized utterances and respond to formulaic questions, to Superior-level speakers, who can support and defend opinions and speak about abstract topics in formal contexts.

After the computer suggests a starting level, the examinee can then perform a sample task at that level. The examinee then has the option to try a second sample task at a level higher or lower. Once the sample tasks are completed, the actual test begins at a level of the candidate's own choosing. The tasks are chosen alternatively by the candidate and the computer program's underlying algorithm. The algorithm ensures that examinees are pushed to show the full extent of their ability by monitoring that tasks at more difficult

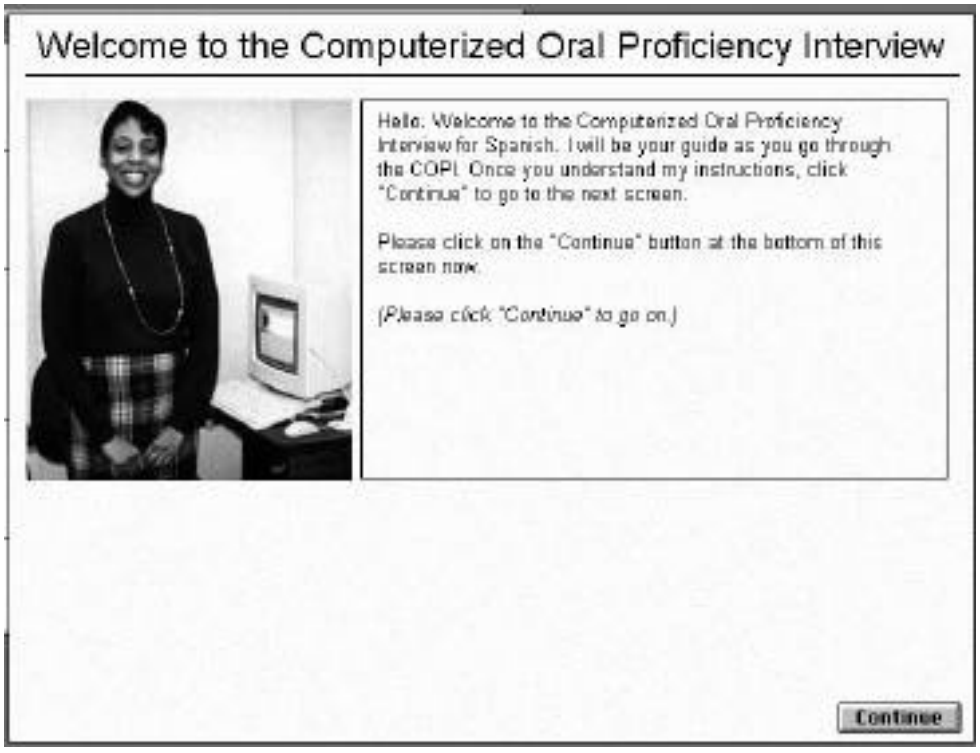
levels are also presented to the examinee. Raters need to hear candidates attempt tasks higher than their proficiency level to determine that they cannot sustain performance on tasks above that level (i.e., they have reached a performance "ceiling"). Nevertheless, the program allows candidates to select many tasks at the level they believe is most appropriate for them.

Because the level of the tasks is targeted to the proficiency level of the examinee, the number of tasks administered on the COPI has been decreased from the 15 tasks on the current long form of the SOPI. In the current COPI, candidates were generally administered seven tasks.

Because the computer can store a large number of tasks, parallel versions of each task, suitable for different populations, can be stored in the task pool. The candidate inputs background information that the computer can then use to select appropriate tasks. This was very useful, for example, in the Arabic COPI, in which the computer selected the task containing the correct form (masculine or feminine) for addressing the examinee. In both the Chinese and Arabic COPI, the ability to create a large number of parallel tasks ensured that some gender-specific task situations could be appropriately administered.

Finally, the COPI allows more of the target language to be exhibited in the assessment instrument than the SOPI can. With the exception of giving the directions for the role-play (which are printed in English on a card the candidate reads), the ACTFL OPI is conducted entirely in the target language. With the exception of giving the directions for the role-play in English, the ACTFL OPI is conducted entirely in the target language. In the SOPI, task directions are always given in English in order to avoid any misunderstanding. In contrast to the SOPI, examinees taking the Spanish COPI can choose to have the task directions in Spanish on Advanced and Superior level tasks. Lower level examinees are given directions in both English and Spanish.

Description of the COPI



The basic COPI format is built around a pool of about 100 assessment tasks. These tasks are adapted from tasks successfully used in SOPIs (for more information on developing SOPIs, refer to Stansfield, 1996; for information on the piloting of SOPI tasks, see Kenyon & Stansfield, 1993). Each task is designed to

elicit speech at one of the four main levels of the ACTFL proficiency scale: Novice, Intermediate, Advanced, and Superior. Each task is coded for its speaking function and content/topic area. For the multimedia COPI test administration program, each task is comprised of several parts. These parts include written files and audio directions in English, written and audio directions in the target language (Spanish, Chinese or Arabic), a graphic file of a picture that accompanies the task (for those tasks that have pictures), and an audio prompt from a native speaker of the target language. Depending on the choices that the examinee makes, the administration of the test takes anywhere from 30-50 minutes.

When taking the COPI, the examinee goes through nine phases: welcome, information on the purpose and structure of the COPI, input and correction of personal information, self-assessment of proficiency level, listening to an adequate response to a sample task(s), practice with the same sample task(s), responding to performance tasks (the actual test), feedback about the levels of the tasks that the examinee took, and closing. A photograph of a friendly virtual female "guide" accompanies the audio and written directions presented on the computer screens.

As with the SOPI, performances on the COPI are assessed following the criteria of the ACTFL *Speaking Proficiency Guidelines.* The multimedia COPI rating program allows raters to hear the examinees' responses for each task and to listen to the examinees' tasks in any order. As raters assess each task, elements of the task, such as its ACTFL level, the picture accompanying the task, the directions, and the target language prompt appear on the screen. These elements give raters background information about each task and facilitate the assessment of the performances. Raters can also listen to any portion of each examinee's response for a particular task as many times as desired, and go back to previously rated tasks. The program also allows raters to write notes to examinees so that, aside from providing a global rating (i.e., the ACTFL proficiency level at which the examinee demonstrated consistent performance), raters are also able to give overall comments and task-specific feedback to each examinee. The rating program calculates the final global rating based on an algorithm (for a more detailed description of the COPI, see Malabonga & Kenyon, 1999).

RESEARCH QUESTIONS

This article focuses on examinee reactions to the COPI format, particularly in comparison to the SOPI. For the Spanish examinees, COPI is compared to both the SOPI and OPI.

How Do Examinees' Views on the COPI Compare to the SOPI?

For the first part of our analysis, we compared learner reactions to the COPI and the SOPI.

1. Did examinees feel they had sufficient opportunity to demonstrate their strengths and weaknesses in speaking the target language on the COPI and the SOPI?

In designing the COPI as the next generation of the SOPI, we sought to ensure that examinees would feel that they had similar opportunities on both tests. We thus envisioned that examinees would have similar perceptions about the adequacy of each test.

2. Did examinees feel that the tests were differentially difficult?

Because we designed the COPI to be adaptive, that is, examinees would be administered tasks more matched to their proficiency, we envisioned that examinees would find the COPI less difficult than the SOPI.

3. Did examinees feel speaking situations in the COPI or SOPI were unfair?

Because the SOPI task format has been extensively piloted and revised through the development of SOPIs in over nine languages, and because the COPI tasks were based on the SOPI tasks, we envisioned that examinees would find both tests to be fair.

4. Did examinees feel differentially nervous taking the tests?

We predicted that, because the COPI provided the examinee more control over the testing situation than did the SOPI, examinees might feel less nervous when taking the COPI than the SOPI.

5. Did examinees feel that the directions for the tests were clear?

The clarity of test directions is especially important, as both tests are essentially self-administered. Examinees have no opportunity to stop the test and ask questions of an administrator once the test has begun. Again, because of the vast experience CAL has had with SOPI tasks and the similarity between SOPI and COPI tasks, we anticipated that examinees would believe that the directions were clear, and that examinees' perceptions about clarity of test directions with regard to both assessments would be similar.

6. Did examinees feel that someone listening to their responses on the test would get an accurate picture of their current ability to speak the target language in real-life situations outside the classroom?

Again, because of the similarities between the SOPI and the COPI, we anticipated that examinees would have a positive attitude towards the two tests; that is, they would believe that the tests are accurate, and their perceptions about test accuracy with regard to the two assessments would be similar.

How Did Examinees' Views on the COPI and SOPI Compare with Their Views on the OPI?

Similar questions were researched regarding the OPI for the examinees in the Spanish study. We predicted that examinees would have similar perceptions about the adequacy of all three tests in demonstrating their strengths and weaknesses in speaking the target language. As to difficulty, because the OPI is also adaptive, we predicted that the Spanish students would view this test as being less difficult than the SOPI and more similar to the COPI. As to the fairness of the test question, we envisioned that examinees would have similarly positive attitudes towards the three tests. Regarding nervousness, we predicted that examinees might feel less nervous on the OPI than either the SOPI or COPI, due to the human interaction effect and its emphasis on the interviewer's putting the examinee at ease. Regarding clarity, we envisioned that examinees' perceptions about clarity of test directions with regard to the three assessments would be similar. Finally, as to the accuracy of the speech sample, we felt that, due to the technology-mediated nature of the COPI and SOPI, examinees might feel that the OPI was more like real-life conversation.

METHODOLOGY

Sample

The examinees were 50 undergraduate students taking Spanish, Chinese or Arabic courses, 4 graduate students taking Arabic courses, and one graduate student taking linguistics courses. The were all taking courses at universities in the Washington, DC metropolitan area. Twenty-four students participated in the Spanish study, 15 in the Arabic study, and 16 in the Chinese study. Students received a small honorarium for their participation in the study, and scores on all the tests were returned to them following the rating.

Instruments

Questionnaire 1. Individual Test -- Examinee Feedback Sheet. After taking each test, examinees were asked to fill out a questionnaire about that test. The questionnaire asked about various aspects of the test that they felt might have influenced their performance. Examinees were presented with statements about the test and asked to indicate the degree to which they agreed with the statement using a four point scale: 4 = *Strongly Agree*, 3 = *Agree*, 2 = *Disagree*, and 1 = *Strongly Disagree.*

The questions in the OPI, the COPI, and the SOPI questionnaires were the same across the three tests and the three languages. Only the name of the test and the name of the language were different. The following is a list of the six questions, followed by a brief descriptive phrase in parentheses:

1. I feel I had the opportunity to adequately demonstrate both my strengths and my weaknesses in speaking (*language*) on the (*test*). (strengths and weaknesses)
2. I feel the (*test*) was difficult. (difficulty)
3. I feel there were questions asked or speaking situations required in the (*test*) that were unfair. (fairness)
4. I felt nervous taking the (*test*). (nervousness)
5. I feel the directions for the (*test*) were clear (i.e., I felt I always knew what I needed to do). (clarity)
6. I feel someone listening to my (*test*) responses would get an accurate picture of my current ability to speak (*language*) in real life situations outside the classroom. (accuracy)

Examinees could write comments on the feedback sheet following their indication of agreement to each statement. Other open-ended questions on the feedback sheet asked about specific aspects of each test.

Questionnaire 2. Test Comparisons -- Examinee Feedback Sheet. After completing all tests and completing Questionnaire 1 for each test, examinees were given a second questionnaire. The questions in Questionnaire 2 directly parallel those in Questionnaire 1, except that examinees were asked to compare the SOPI and the COPI rather than indicate agreement on a 1-4 scale. After each question, examinees circled one of three choices:

SOPI (taped test)
COPI (computerized test)
Both the same

The following is a list of the six questions on Questionnaire 2, followed by a brief descriptive phrase in parentheses. The questions were the same across the three languages. Space was provided for examinees to write optional comments after each question.

1. Which technology-based test did you feel better allowed you to demonstrate both your current strengths and weaknesses in speaking (*language*)?
2. Which technology-based test did you feel was more difficult?
3. Which technology-based test did you feel was fairer?
4. Which technology-based test did you feel more nervous taking?
5. Which technology-based test did you feel had clearer directions?
6. Which technology-based test do you feel would give someone listening to your performance a more accurate picture of your current ability to speak (*language*) in real life situations outside the classroom?

Procedures

Researchers requested chairpersons and professors in the foreign languages, linguistics, and other language-related departments to announce the study to students in their classes. Interested students were asked to submit three forms before they could participate in the study.

- The first form provided background information about the students' experience with the target language (e.g., how many classes they have taken, whether they speak the target language at

home, how much they used the language in daily life, and whether they have lived in a community or country where the target language is spoken).

- The second form asked students to provide a self-assessment of their current speaking proficiency in the target language.
- The third form was a teacher's assessment of the students' speaking proficiency.

In order to ensure that a full range of proficiency levels was represented in the study, information from the examinees' three forms was used to group them *a priori* into four likely proficiency levels approximating the ACTFL levels: *novice, intermediate, advanced,* and *superior.* An attempt was made to assign an equal number of students to each proficiency group. Back-up participants in the study were also identified and notified for each subject chosen. The participants who turned in their forms earliest were assigned as original subjects, whereas those who turned in their forms later were assigned as back-ups. If a subject did not appear at the scheduled testing session, one of the back-ups was notified. The back-up participants belonged to the same proficiency group and were assigned the same order of taking the three tests as the subjects originally chosen for the study.

Based on their proficiency grouping, participants were assigned to the order of taking the tests at the CAL office. This allowed the order of the administration of the tests across proficiency groups to be completely balanced. For example, of the six *novice* subjects in the Spanish study, two were assigned to take the COPI first, two to take the SOPI first, and two to take the OPI first.

There were three testing sessions at CAL, one for each language. Each session was divided into four periods and was conducted over two days. Examinees were assigned to one of the four periods, with six examinees coming to each period for the Spanish study and four to each period for the Arabic and Chinese studies. During each period, examinees were given general instructions as a group before the administration of the tests. Each examinee was then taken to a room with an interviewer (OPI, Spanish students only), tape recorders (SOPI), or a computer (COPI) to take their test. The two OPI interviewers were currently certified ACTFL OPI testers and participated in the study through Language Testing International, ACTFL's official testing arm.

A proctor introduced the examinee to the interviewer (OPI) or assisted the examinee with using the tape recorder (SOPI) or computer (COPI) before and after each SOPI and COPI test. The SOPI and COPI were completed by the examinee alone in a quiet room. Extra COPI and SOPI rooms were set up so that back-up examinees coming in late would still be able to take the tests in the same order as the subjects originally chosen for the study without disturbing the schedule of the other participants.

After taking each test, participants filled out a feedback questionnaire about their attitudes towards that specific test (Questionnaire 1). After completing all tests and the questionnaire for each test, examinees completed a final questionnaire comparing the COPI and the SOPI tests (Questionnaire 2).

The procedure is summarized in Figure 1.

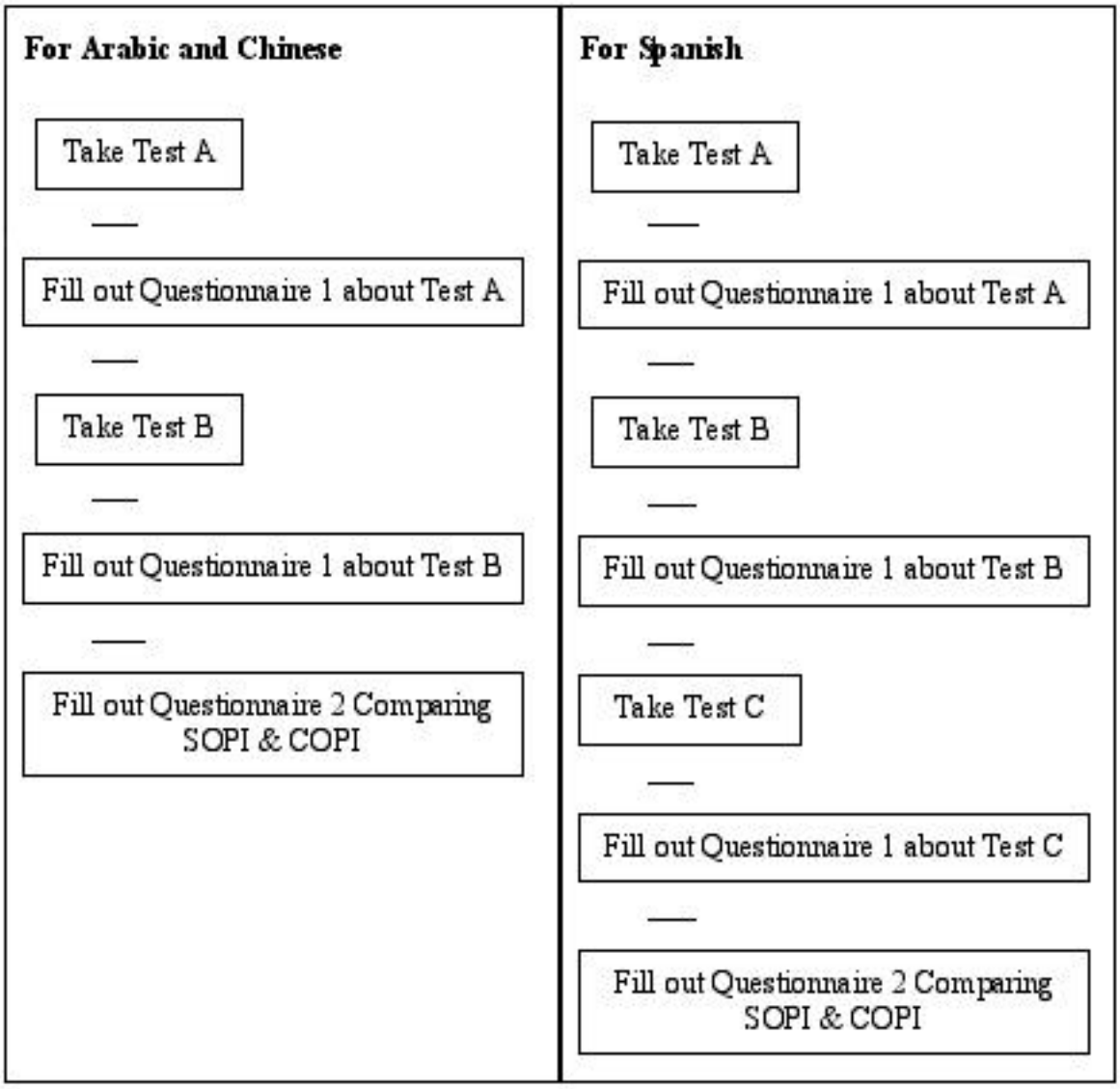


Figure 1. Summary of experimental procedure

Analyses

Our first analysis is of the means and standard deviations for all Likert-scale questions in Questionnaire 1 across the 55 examinees. The paired samples t-test procedure of SPSS was used to compare the 55 examinees' responses to the same questions on the COPI and SOPI questionnaires. This statistic is the appropriate one for comparing the means on two variables for a single group.

For the results of Questionnaire 2, we calculated the percent of examinees selecting each possible response. In analyzing differences in examinees' affect due to language and proficiency level in this questionnaire 2, we used chi-square likelihood ratio tests to compare the actual frequency of examinees' responses to each question to the estimated frequency based on the marginal distributions.

For the Spanish sample of 24 examinees, which compared the OPI, SOPI, and COPI, the non-parametric Friedman test was used in lieu of repeated measures ANOVA. This test was used because the sample was small and the Friedman test is not as strict as the ANOVA on assumptions of normality and homogeneity

of variance. Because this sample group was so small, no investigation of differences due to proficiency was undertaken.

Because the SOPIs and COPIs used the same format in each language but may have had intrinsic differences influencing examinee affect, we also investigated whether there were any differences in the examinee responses due to language.

Finally, we investigated whether there were any differences in the examinee responses to the questions due to the examinees' proficiency levels. We divided the 55 examinees into three proficiency levels (*low, mid,* and *high*) based on their final ratings on the SOPI and COPI. Fifteen examinees who obtained at least one rating of *advanced-high* from any rater on either test were classified as *high.* Twenty-one examinees who never obtained a rating above *intermediate-mid* were classified as *low.* The remaining 18 examinees were classified as *mid.* We used the non-parametric Kruskal-Wallis test to conduct these comparisons, as the number of examinees in each subgroup was small and uneven, and no assumptions about distributions of responses could be made.

All analyses were conducted using SPSS version 10 for the PC.

RESULTS

Performance

While the focus of this paper is on the results of the affective questionnaires that were completed by all 55 students participating in the study, a comparison of the performances of the 46 students whose COPI and SOPI tests were complete and ratable is informative. These results are presented in Table 1. The 10 ACTFL levels were converted into ranks, such that the lowest level (n*ovice-low*) was equal to 1 and the highest level (*superior*) was equal to 10. Using these ranking, there was a rank order correlation of .95 between ratings on the COPI and the SOPI. This result indicates that students scored very similarly on both tests. The average ranking on the COPI was 5.85, which was between an *intermediate-mid* (level 5) and *intermediate-high* (level 6). The mean ranking on the SOPI was 5.57, between those two levels, but somewhat closer to *intermediate-mid.*

Table 1. Performance Results (Mean Ranking and Rank Order Correlations)

Group		**COPI**	**SOPI**	**OPI**
Subjects from all languages with complete test results (*N* = 46)	Mean	5.85	5.57	N/A
	Correlation	COPI – SOPI =.95		
Subjects from Spanish study with complete test results (*N* = 16)	Mean	7.00	6.63	6.25
	Correlations	COPI – SOPI = .95 COPI – OPI = .92 SOPI – OPI = .94		

As the group of examines was deliberately chosen to be very heterogeneous, there was a relatively large standard deviation for both means: 2.38 for the COPI and 2.34 for the SOPI.

For the group of 16 Spanish students who had final ratings on the COPI, SOPI and OPI, the mean rankings were higher for all three tests: 7.00 *(advanced-low)* on the COPI, 6.63 (between *intermediate-high* [level 6] and *advanced-low* [level 7]) on the SOPI, and 6.25 on the OPI (closer to *intermediate-high).* Rank order correlations for this group of 16 students remained high between the COPI and SOPI (.95), the COPI and the OPI (.92), and the SOPI and the OPI (.94). However, when there was a disagreement between ratings, the SOPI or COPI rating had a tendency to be higher than the OPI rating. This accounts for the higher mean rankings on the COPI and SOPI versus the OPI.

Comparison of the COPI and SOPI (Questionnaire 1)

Table 2 gives the descriptive statistics on the six four-point Likert scale questionnaires appearing on the two versions of Questionnaire 1: one for the COPI and one for the SOPI. The first column gives the question and the second column the number of students across each language that responded to the question for both tests. The fourth column presents the average response to that question and its standard deviation for the COPI questionnaire, whereas the fifth column presents the same information for that question on the SOPI questionnaire.

Table 2. Comparative Results on Questionnaire 1 (COPI and SOPI)

Question	N		COPI	SOPI
1. I feel I had the opportunity to adequately demonstrate both my strengths and my weaknesses in speaking (*language*) on the (*test*).	55	Mean	2.98	2.84
		Std. Dev.	.62	.66
2. I feel the (*test*) was difficult.	55	Mean	2.30	2.71
		Std. Dev.	.60	.74
3. I feel there were questions asked or speaking situations required in the (*test*) that were unfair.	55	Mean	1.80	1.89
		Std. Dev.	.56	.53
4. I felt nervous taking the (*test*).	55	Mean	2.24	2.49
		Std. Dev.	.84	.84
5. I feel the directions for the (*test*) were clear (i.e., I felt I always knew what I needed to do).	55	Mean	3.58	3.44
		Std. Dev.	.69	.67
6. I feel someone listening to my (*test*) responses would get an accurate picture of my current ability to speak (*language*) in real life situations outside the classroom.	52	Mean	2.77	2.73
		Std. Dev.	.67	.77

The COPI and SOPI means in Table 2 were compared using a paired samples t-test. Except for the means between examinees' responses to question 2 (difficulty of test, $t = -3.32$, $df = 54$, $p < .05$), none of the comparisons were statistically significant. The group of 55 students agreed more with the statement that "the SOPI was difficult" than with the statement that "the COPI was difficult."

Comparison of the COPI and SOPI (Questionnaire 2)

Table 3 shows the percentage of respondents who chose each option on the six questions on Questionnaire 2, which asked for a direct comparison between the COPI and the SOPI. The first column gives the question, the second the percentage of the 55 respondents who chose *COPI*, the third the percentage who chose *SOPI*, and the fourth the percentage who chose *Both the same*. The final column shows the percentage of the 55 respondents who did not answer the question.

Table 3. Comparative Results on Questionnaire 2 (COPI and SOPI)

Question	COPI	SOPI	BOTH	Missing
1. Which technology-based test did you feel better allowed you to demonstrate both your current strengths and weaknesses in speaking (*language*)?	60.0%	25.5%	14.5%	0.0%
2. Which technology-based test did you feel was more difficult?	29.1%	56.4%	14.5%	0.0%
3. Which technology-based test did you feel was fairer?	47.3%	20.0%	32.7%	0.0%
4. Which technology-based test did you feel more nervous taking?	32.7%	47.3%	18.2%	1.8%

5. Which technology-based test did you feel had clearer directions?	34.5%	10.9%	54.5%	0.0%
6. Which technology-based test do you feel would give someone listening to your performance a more accurate picture of your current ability to speak *(language)* in real life situations outside the classroom?	45.5%	29.1%	21.8%	3.6%

Tables 2 and 3 allow us to examine whether our predicted outcome on examinee affect were accurate. Table 2 shows that for Question 1, there was no statistical difference between the mean ratings on the Likert scale between the COPI and SOPI, and that the average was nearest the scale point 3 *(Agree).* That is, examinees generally agreed that both tests gave them an adequate opportunity to demonstrate both strengths and weaknesses in speaking their target language. However, Table 3 reveals that, in a direct comparison, the majority of examinees (60%) felt the COPI better allowed them to make that demonstration. Only a quarter of the respondents chose the SOPI.

Responses to the comments section on the question reveal examinees' positive reactions to the COPI innovations. Here is a sampling:

- I chose the COPI because, as I wrote on the feedback sheets to the tests, the SOPI left me feeling greater pressure in having to respond in a restricted time. The beeps were somewhat intrusive; seeing the circles on the computer screen was far more helpful and calming.
- I was more comfortable with the COPI because it allowed me to move at my own pace and took proficiency-level [according to the student] into account.
- The COPI was easier to use and less boring.
- COPI - Next best thing to a human!
- COPI - It showed strengths by being able to respond to what you felt you could do best and weaknesses will be shown by any response (tones, vocabulary, grammar).

For Question 2, the mean response regarding whether the test was difficult, the mean for the COPI (2.30) is close to the rating of *Disagree,* whereas the rating for the SOPI (2.71) is closer to the rating of *Agree.* This difference is statistically significant (t =-3.32, df=54, $p < .05$), and Table 3 supports the results in Table 2 by showing that a majority of examinees chose the SOPI as the more difficult test. It appears that examinees felt the COPI was less difficult. Comments, summarized below, again reveal some of the aspects of the SOPI that made it appear more difficult:

- SOPI - Because there were no choices and having time restraints made it a lot harder.
- SOPIs time restraint created difficulty.
- The SOPI was more difficult because you were unable to choose your situation and the times in which to start/finish speaking.
- SOPI - This is because the SOPI tested all levels, whereas the COPI was more realistic and enabled me to show what I knew.

Examinees clearly disagreed that there were tasks on either test that were unfair. The mean response to the third question was 1.80 for the COPI and 1.89 for the SOPI, which was below the scale point of 2, *Disagree.* This result is probably due to the similarity of the COPI tasks to the SOPI tasks, which had been professionally developed and extensively piloted. Question 3 on Questionnaire 2 was a little more general, however. Almost half of the examinees (47.3%) felt the COPI was a fairer test, with the second largest percentage (32.7%) feeling both were the same. Only 20% felt the SOPI was fairer.

Again, examinee comments on Question 3 on Questionnaire 2 reveal that candidates were sensitive to the effects of the innovations in the COPI:

- COPI -- because I got to see how much time was remaining and also I got to choose what I wanted to talk about.
- Conversation doesn't happen in timed sequences and is usually like COPI, geared toward the level of the participant.
- COPI allowed me to choose my level. I never felt inferior.
- COPI allowed choice allowing speaker to demonstrate in areas they feel comfortable with.
- COPI -- I was able to apply more of the vocabulary I know by choosing questions rather than being forced to answer certain ones.

Question 4 was about nervousness. On Questionnaire 1, examinees were midway between agreeing and disagreeing with being nervous taking each test. The mean for the COPI (2.24) was slightly lower than for the SOPI (2.49), which was midway between a rating of 3 *(Agree)* and 2 *(Disagree)*. The standard deviation of the responses to this question (.84 for each) was greater than for any other question, indicating a wider range of responses for this question compared to the others. Although the mean rating for the COPI was lower than for the SOPI (i.e., examinees were less nervous on the COPI), it was not statistically significant at the .05 level. Table 3, however, reveals that almost half of the students (47.3%) indicated they were more nervous taking the SOPI than the COPI. Nevertheless, almost one in three (32.7%) indicated they were more nervous with the COPI.

Examinee responses shed some light on this finding. Again, examinees mentioned aspects of the COPI versus the SOPI that made them less nervous:

- SOPI -- The beeps cutting you off mid-sentence made it nerve-racking.
- COPI had a soothing voice and many practice tasks.

On the other hand, there were aspects of the COPI that unsettled some candidates:

- The SOPI didn't require that I constantly click "Continue." I could just relax and concentrate on Spanish.
- COPI -- Too many choices.

The technological simplicity of the SOPI was an advantage in the eyes of some examinees. A few examinees indicated that they never feel comfortable dealing with computers. Thus, while we see a tendency for examinees to feel less nervous with the COPI, this was not as strong an effect as may have been expected.

Responses to Question 5 indicated that test directions for both tests were perceived as very clear. The average rating was 3.58 for the COPI and 3.44 for the SOPI. This was about midway between the scale points of 3 *(Agree)* and 4 *(Strongly Agree)*. The difference between the means was not statistically significant ($t = 1.112$, $df = 54$, $p < .05$.). These high ratings are important for a self-administering test, as examinees have no proctor to turn to if there are any directions they do not understand. Table 3 indicates that over half the examinees (54.5%) chose *Both the same* in response to which test had clearer directions. Of the remaining, about three times as many (34.5% vs. 10.9%) felt that the directions to the COPI were clearer. Nevertheless, at least one examinee recognized that explaining the directions for using the computer in the COPI came at a price: "Much more time was spent during the COPI exam just learning how to take the exam."

Question 6 asked to what extent examinees felt someone listening to their performances would get an accurate picture of their proficiency. The average response to Question 6 for both the COPI and SOPI was similar (2.77 and 2.73, respectively), indicating a mean rating just under scale point 3 *(Agree)*. Question 6 on Questionnaire 2 indicates that almost half of the examinees (45.5%) indicated that someone listening to the responses on the COPI would get a more accurate picture (despite the fact that most examinees responded to only 7 tasks on the COPI versus 15 on the SOPI). Almost an equal number chose the SOPI (29.1%) as chose both (21.8%).

The optional comments to this question highly favored the COPI, as this sampling shows:

- COPI -- Since it let you choose what you wanted to talk about and at what level, it seemed fairer.
- Probably the COPI because I felt like I had more time to prepare my response.
- COPI -- It asks you more general questions about yourself so it feels like a conversation.
- A combination of clear time countdown, explanations in Arabic, and ability to choose my own level made this test a good judge. Of course, a native speaker conversing is best of all, but in the absence of this, I found the exam a good measure of my level and really enjoyed taking it.
- Although in real life you are unable to choose what unexpected situation you might find yourself in, I was able to speak Chinese much more easier and clearer using the COPI. I feel it was the fairer and better representation of my speaking ability.

Nevertheless, examinees tended to recognize that the technology-mediated tests were not a conversation and that responses were generally one-way, despite the fact that they felt the COPI was more conversational.

- Both -- The answers are given in exactly the same way. It's not like a conversation, it's more like giving a speech. Few people are good a giving speeches in any language. It is a skill that is independent of one's language skills.
- Both -- I don't feel technology based tests can come so close to having interaction with a person. "Speaking" with another person is the essence or key goal you want to capture with a foreign language testing, and with technology you are not "speaking" with a person in either situation.

In comparing the SOPI and COPI, two Spanish students (who had also taken the OPI) commented on the opening "warm-up" simulated conversation as a very positive aspect of the SOPI. This simulated conversation was not replicated in the COPI.

- SOPI -- There was a "mock" two-way conversation at the beginning that may help in establishing a relaxed atmosphere.
- Person speaks, then you have to speak in some parts of SOPI just like in OPI.

Comparison of Language Groups on the COPI (Questionnaire 1)

We next wanted to see whether there were any differences in the responses to Questionnaires 1 and 2 due to the language of the test. The SOPI forms across the three languages were very different in content. The COPI task pool was more similar, as the pool was first developed for Spanish and then adapted for Arabic and Chinese.

Column 1 in Table 4 presents the six questions. The last three columns indicate the means and standard deviations on the four-point Likert scale for each language group.

Table 4. Means on the COPI Questionnaire 1 Across Language Groups

Question		Span.	Arab.	Chin.
COPI	*N*	24	15	16
1. I feel I had the opportunity to adequately demonstrate both my strengths and my weaknesses in speaking (*language*) on the COPI.	Mean	2.88	3.00	3.13
	Std. Dev.	.74	.53	.50
2. I feel the COPI was difficult.	Mean	2.21	2.40	2.38
	Std. Dev.	.72	.51	.50
3. I feel there were questions asked or speaking situations required in the COPI that were unfair.	Mean	2.00	1.60	1.69
	Std. Dev.	.51	.51	.60
4. I felt nervous taking the COPI.	Mean	2.42	2.00	2.19
	Std. Dev.	.88	.76	.83
5. I feel the directions for the COPI were clear (i.e., I felt I always knew what I needed to do).	Mean	3.54	3.47	3.75
	Std. Dev.	.59	.92	.58
6. I feel someone listening to my COPI responses would get an accurate picture of my current ability to speak (*language*) in real life situations outside the classroom.	Mean	2.61	3.00	2.69
	Std. Dev.	.72	.78	.60

The means of the three language groups on the six COPI questions were compared using the Kruskal-Wallis test. None of the differences between the means of the three language groups were statistically significant at the .05 level.

Comparison of Language Groups on the SOPI (Questionnaire 1)

Table 5 uses the same structure as Table 4 to summarize the data from the SOPI Questionnaire 1.

Table 5. Means on the SOPI Questionnaire 1 Across Language Groups

Question		Span.	Arab.	Chin.
SOPI	*N*	24	15	16
1. I feel I had the opportunity to adequately demonstrate both my strengths and my weaknesses in speaking (*language*) on the SOPI.	Mean	2.92	2.80	2.75
	Std. Dev.	.65	.86	.45
2. I feel the SOPI was difficult.	Mean	2.50	2.93	2.81
	Std. Dev.	.78	.70	.66
3. I feel there were questions asked or speaking situations required in the SOPI that were unfair.	Mean	2.00	1.80	1.81
	Std. Dev.	.51	.56	.54
4. I felt nervous taking the SOPI.	Mean	2.42	2.6	2.5
	Std. Dev.	.78	.83	.97
5. I feel the directions for the SOPI were clear (i.e., I felt I always knew what I needed to do).	Mean	3.42	3.27	3.63
	Std. Dev.	.58	.96	.50
6. I feel someone listening to my SOPI responses would get an accurate picture of my current ability to speak (*language*) in real life situations outside the classroom.	Mean	2.71	2.71	2.69
	Std. Dev.	.81	.83	.70

Again, a comparison of the means of the three language groups on the six SOPI questions using the Kruskal-Wallis test revealed that none of the differences between the means of the three language groups were statistically significant at the .05 level.

Comparison of Language Groups for both COPI and SOPI (Questionnaire 2)

Table 6 shows the differences across language groups on Questionnaire 2. The question is printed in the first column. The second column for each question indicates the language group for which the responses were chosen. The final four columns indicate the observed value and expected value (in parentheses) of respondents who chose *COPI, SOPI, Both the same,* or *None.* The expected value is predicted based both on the total number of responses across languages and the total number of responses across responses.

Table 6. Observed and Expected Responses (in parentheses) on Questionnaire 2 Across Languages

Question		COPI	SOPI	BOTH	Missing
1. Which technology-based test did you feel better allowed you to demonstrate both your current strengths and weaknesses in speaking (*language*)?	Spanish	12 (14.4)	8 (6.1)	4 (3.5)	N/A
	Arabic	9 (9.0)	3 (3.8)	3 (2.2)	N/A
	Chinese	12 (9.6)	3 (4.1)	1 (2.3)	N/A
2. Which technology-based test did you feel was more difficult?	Spanish	9 (7.0)	12 (13.5)	3 (3.5)	N/A
	Arabic	3 (4.4)	9 (8.5)	3 (2.2)	N/A
	Chinese	4 (4.7)	10 (9.0)	2 (2.3)	N/A
3. Which technology-based test did you feel was fairer?	Spanish	12 (11.3)	6 (4.8)	6 (7.9)	N/A
	Arabic	8 (7.1)	3 (3.0)	4 (4.9)	N/A
	Chinese	6 (7.6)	2 (3.2)	8 (5.2)	N/A
4. Which technology-based test did you feel more nervous taking?	Spanish	10 (7.9)	9 (11.3)	4 (4.4)	1 (0.4)
	Arabic	3 (4.9)	9 (7.1)	3 (2.7)	0 (0.3)
	Chinese	5 (5.2)	8 (7.6)	3 (2.9)	0 (0.3)
5. Which technology-based test did you feel had clearer directions?	Spanish	9 (8.3)	1 (2.6)	14 (13.1)	N/A
	Arabic	5 (5.2)	3 (1.6)	7 (8.2)	N/A
	Chinese	5 (5.5)	2 (1.7)	9 (8.7)	N/A
6. Which technology-based test do you feel would give someone listening to your performance a more accurate picture of your current ability to speak (*language*) in real life situations outside the classroom?	Spanish	9 (10.9)	10 (7.0)	5 (5.2)	0 (0.9)
	Arabic	10 (6.8)	3 (4.4)	1 (3.3)	1 (0.5)
	Chinese	6 (7.3)	3 (4.7)	6 (3.5)	1 (0.6)

The chi-square test was used to compare the observed values across languages to the statistically expected values. None of the differences between the observed and expected values were significant at the .05 level.

Taken together, Tables 4 through 6 indicate that the language of the test form (COPI or SOPI) did not generate statistically significant differences between ratings on the two questionnaires. This result suggests that we can have additional confidence that the results in Tables 2 and 3 can be interpreted in terms of the format of the different tests (COPI vs. SOPI). The idiosyncratic differences between the SOPI forms and the COPI versions across languages did not seem to influence examinees' ratings.

Comparison of Proficiency Groups on the COPI (Questionnaire 1)

We next wanted to see whether there were any differences in the responses to Questionnaires 1 and 2 due to the proficiency level of the examinee. It is possible that the adaptive nature of the COPI had a stronger influence on the affect of students at one proficiency level than another.

Tables 7 through 9 mimic the structures of Tables 4 through 6, except that in this case the three groups are the proficiency groups rather than the language groups. Thus, the first column in Table 7 presents the six questions, and the last three columns indicate the means and standard deviations on the four-point Likert scale for each proficiency group.

Table 7. Means on the COPI Questionnaire 1 Across Proficiency Groups

Question		**LOW**	**MID**	**HIGH**
COPI	*N*	21	18	15
1. I feel I had the opportunity to adequately demonstrate both my strengths and my weaknesses in speaking (*language*) on the COPI.	Mean	3.05	2.89	3.00
	Std. Dev.	.59	.76	.53
2. I feel the COPI was difficult.	Mean	2.24	2.39	2.27
	Std. Dev.	.62	.61	.59
3. I feel there were questions asked or speaking situations required in the COPI that were unfair.	Mean	1.52	2.00	1.93
	Std. Dev.	.51	.49	.59
4. I felt nervous taking the COPI.	Mean	2.10	2.17	2.53
	Std. Dev.	.77	.92	.83
5. I feel the directions for the COPI were clear (i.e., I felt I always knew what I needed to do).	Mean	3.62	3.72	3.40
	Std. Dev.	.80	.46	.74
6. I feel someone listening to my COPI responses would get an accurate picture of my current ability to speak (*language*) in real life situations outside the classroom.	Mean	2.95	2.44	2.79
	Std. Dev.	.68	.78	.58

The Kruskal-Wallis test was used to compare the means of the three proficiency groups on the six COPI questions. Except for question 3 (fairness of the test, chi square = 8.22, $df = 2$, $p < .05$), none of the differences between the means of the three groups were statistically significant. That is, examinees generally did not have negative attitudes toward the fairness of the COPI as indicated by their responses being 2 or less (i.e., they disagreed that it was unfair). Nevertheless, examinees who scored in the ACTFL Novice and Intermediate range (i.e., at *intermediate-mid* or below) had significantly more positive attitudes regarding the COPI's fairness (mean of 1.53) than candidates at the two higher proficiency groups (means of 2.00 and 1.93, respectively).

Comparison of Proficiency Groups on the SOPI (Questionnaire 1)

Table 8 replicates the structure of Table 7 for Questionnaire 1 responses to the SOPI.

Table 8. Means on the SOPI Questionnaire 1 Across Proficiency Groups

Question		LOW	MID	HIGH
SOPI	*N*	21	18	15
1. I feel I had the opportunity to adequately demonstrate both my strengths and my weaknesses in speaking (*language*) on the SOPI.	Mean	2.76	2.67	3.07
	Std. Dev.	.62	.84	.23
2. I feel the SOPI was difficult.	Mean	3.19	2.55	2.27
	Std. Dev.	.60	.70	.59
3. I feel there were questions asked or speaking situations required in the SOPI that were unfair.	Mean	2.00	1.78	1.87
	Std. Dev.	.63	.55	.35
4. I felt nervous taking the SOPI.	Mean	2.62	2.56	2.27
	Std. Dev.	.97	.86	.59
5. I feel the directions for the SOPI were clear (i.e., I felt I always knew what I needed to do).	Mean	3.48	3.44	3.33
	Std. Dev.	.81	.62	.62
6. I feel someone listening to my SOPI responses would get an accurate picture of my current ability to speak (*language*) in real life situations outside the classroom.	Mean	2.67	2.59	2.87
	Std. Dev.	.73	.94	.64

Again, the Kruskal-Wallis test was used to compare the means of the three proficiency groups on the six SOPI questions. Except for Question 2 (difficulty of the test), none of the differences between the means of the three proficiency groups were statistically significant (chi-square = 15.49, $df = 2$, $p < .05$). The mean rating for the low proficiency students on Question 2 was 3.19, for the *mid* proficiency students, 2.55, and 2.27 for the *high* proficiency students.

This indicates that the *low* proficiency students generally agreed that the SOPI was difficult, while higher proficiency students tended to disagree. It is interesting to contrast these results with those for Question 2 on the COPI, presented in Table 7. There the mean response for the low proficiency students was 2.24. These responses clearly show the differences on examinee affect between the adaptive COPI (which presents the tasks at difficulty levels more targeted to the examinee's proficiency) and the full-length SOPI (which in this study presented Superior-level tasks to Novice- and Intermediate-level examinees).

Comparison of Proficiency Groups on both the COPI and SOPI (Questionnaire 2)

Table 9 replicates the structure of Table 6, with the exception that the three groups are based on proficiency rather than language.

Table 9. Observed and Expected Responses (in parentheses) on Questionnaire 2 Across Proficiency Groups

Question		COPI	SOPI	BOTH	Missing
1. Which technology-based test did you feel better allowed you to demonstrate both your current strengths and weaknesses in speaking (*language*)?	Low	16 (12.8)	4 (5.1)	1 (3.1)	N/A
	Mid	10 (11.0)	6 (4.3)	2 (2.7)	N/A
	High	7 (9.2)	3 (3.6)	5 (2.2)	N/A
2. Which technology-based test did you feel was more difficult?	Low	1 (5.8)	18 (12.1)	2 (3.1)	N/A
	Mid	7 (5.0)	8 (10.3)	3 (2.7)	N/A
	High	7 (4.2)	5 (8.6)	3 (2.2)	N/A

3. Which technology-based test did you feel was fairer?	Low	15 (9.7)	2 (4.2)	4 (7)	N/A
	Mid	6 (8.3)	6 (3.7)	6 (6.0)	N/A
	High	4 (6.9)	3 (3.1)	8 (5.0)	N/A
4. Which technology-based test did you feel more nervous taking?	Low	2 (6.6)	15 (10.1)	4 (3.9)	0 (0.4)
	Mid	6 (5.7)	7 (8.7)	5 (3.3)	0 (0.3)
	High	9 (4.7)	4 (7.2)	1 (2.8)	1 (0.3)
5. Which technology-based test did you feel had clearer directions?	Low	8 (7.4)	1 (2.3)	12 (11/3)	N/A
	Mid	7 (6.3)	2 (2.0)	9 (9.7)	N/A
	High	4 (5.3)	3 (1.7)	8 (8.1)	N/A
6. Which technology-based test do you feel would give someone listening to your performance a more accurate picture of your current ability to speak (*language*) in real life situations outside the classroom?	Low	12 (9.7)	3 (5.8)	5 (4.7)	1 (0.8)
	Mid	6 (8.3)	8 (5.0)	3 (4.0)	1 (0.7)
	High	7 (6.9)	4 (4.2)	4 (3.3)	0 (0.6)

The chi-square test was used to compare the observed values across proficiency groups to the statistically expected values. Results showed that the observed frequencies differed from expected frequencies at a statistically significant level for three of the six questions: Question 2 on test difficulty (chi-square = 14.09, $df = 4$, $p < .05$), Question 3 on fairness (chi-square = 10.46, $df = 4$, $p < .05$), and Question 4 on nervousness (chi-square =16.43, $df = 4$, $p < .05$). Examining Table 9, Question 2, we see that 18 examinees in the *low* proficiency group chose the SOPI as the more difficult test, whereas the total results across proficiency groups and across choices would predict only 12.1 making that selection. Conversely, a lower number (only 1) in the *low* proficiency group chose the COPI as more difficult, while 5 members of that group were expected to make that choice. This result on Questionnaire 2 corroborates the results of Question 2 on Questionnaire 1 for the COPI (i.e., lower proficiency examinees found the SOPI much more difficult than the adaptive COPI).

The results on Question 3 regarding fairness on Questionnaire 2 also corroborate those on Questionnaire 1 for the SOPI. Table 9 shows that a higher number of low proficiency examinees (15) selected the COPI as being fairer than the expected number of 9.7, while a lower number (2) chose the SOPI than expected (4.2). Lower proficiency examinees viewed the adaptive COPI as a fairer test.

On Question 4 regarding nervousness, Table 9 shows that a higher number of *low* proficiency examinees (15) than expected (10.1) chose the SOPI as the test they felt more nervous taking and a lower number (2) than expected (6.6) chose the COPI. Also, a higher number of *high* proficiency students (9) than expected (4.7) chose the COPI as the test they felt more nervous taking, while fewer members of this group (4) than expected (7.2) chose the SOPI. The results to this question indicate that the adaptive nature of the COPI vis-à-vis the full-length SOPI had a more powerful influence on the affect of lower level examinees than on higher proficiency ones.

Comparison of the OPI, COPI and SOPI for the Spanish Group (Questionnaire 1)

Finally, for the Spanish group, we looked at responses across the six questions on Questionnaire 1 for the OPI, COPI, and SOPI. Means and standard deviations for the 24 students in this group are presented in Table 10.

Table 10. Comparative Results on Questionnaire 1 (OPI, COPI and SOPI) for the Spanish Language Group

Question	*N*		OPI	COPI	SOPI
1. I feel I had the opportunity to adequately demonstrate both my strengths and my weaknesses in speaking Spanish on the (*test*).	24	Mean	3.42	2.88	2.92
		Std. Dev.	.72	.74	.65
2. I feel the (*test*) was difficult. (difficult?)	23	Mean	2.26	2.47	2.17
		Std. Dev.	.54	.79	.72
3. I feel there were questions asked or speaking situations required in the (*test*) that were unfair.	24	Mean	1.42	2.00	2.00
		Std. Dev.	.58	.51	.51
4. I felt nervous taking the (*test*).	24	Mean	2.21	2.42	2.42
		Std. Dev.	.83	.88	.78
5. I feel the directions for the (*test*) were clear (i.e., I felt I always knew what I needed to do).	24	Mean	3.75	3.54	3.42
		Std. Dev.	.53	.59	.58
6. I feel someone listening to my (*test*) responses would get an accurate picture of my current ability to speak Spanish in real life situations outside the classroom.	23	Mean	3.39	2.61	2.74
		Std. Dev.	.72	.72	.81

The means of Spanish examinees' responses to the six questions on the three tests were compared using the Friedman test. Results indicate that there were three questions to which the means across the three questionnaires differed: Question 1, perceived opportunity to demonstrate one's strengths and weaknesses (chi square = 15.27, $df = 2$, $p < .05$), Question 3, fairness of the test (chi square = 16.17, $df = 2$, $p < .05$), and Question 6, accuracy of the test (chi square = 16.59, $df = 2$, $p < .05$).

For Question 1, while the means for the COPI and SOPI were similar, (2.88 and 2.92, respectively, near the scale point of 3 *Agree*), the mean rating for the OPI was 3.42, about midway between *Agree* and *Strongly Agree.* While the Spanish students felt that they had some opportunity to adequately demonstrate their abilities in the two technology-mediated tests, they felt more strongly that the face-to-face test provided that opportunity.

The results are very similar for Question 6 (accuracy of the test). There, the means for the COPI and SOPI are again similar and somewhat lower than the scale point of 3 *(Agree),* at 2.61 and 2.74 respectively. The mean for the OPI is 3.39, between 3 *(Agree)* and 4 *(Strongly Agree).* Examinees' comments on the questionnaires focused primarily on the one-sided nature of the technology-mediated assessments as opposed to the face-to-face OPI, which was perceived to be more conversational, interactional, and personal (i.e., focused on topics of interest to the examinee, with the exception of the role-play phase; and more like real-life conversation). In their comments, examinees wondered whether any technologically-mediated speaking test could duplicate these aspects of speaking situations in the real world.

There was a statistically significant difference between the means for Question 3 (fairness of the test). While the means for both technology-based tests were at 2.00 (*Disagree* that there were unfair questions), the mean for the OPI was even lower, at 1.42; that is, between *Disagree* and *Strongly Disagree.*

Although for all 55 students there was a statistically significant difference between the COPI and the SOPI (Tables 2 and 3), with the COPI being perceived as easier, there was no difference between the perceived difficulty of the COPI, SOPI, or OPI for the 24 examinees in the Spanish subgroup. Nor was there a difference in the perception of nervousness taking the test or of the clarity of test directions. This

result is encouraging in the ability of technology-mediated tests to match these aspects of the face-to-face interviews.

Comparison of the OPI, COPI and SOPI for the Spanish Group (Questionnaire 2)

Finally, Questionnaire 2 for the Spanish examinees had one extra question that was not on Questionnaire 2 for the Arabic and Chinese examinees. That question asked, "Which technology-based test did you feel was more similar to the OPI?" Spanish examinees were evenly divided between the COPI (41.7%) and the SOPI (41.7%). *Both the same* was chosen by 16.7% of the examinees. From the ratings and the comments discussed above, it appeared that both technology-mediated tests were unable to replicate the interactive, conversational and personal nature of the face-to-face interview for the Spanish examinees who had experienced all three tests.

DISCUSSION AND CONCLUSIONS

Examinee Views on the COPI Versus the SOPI

When comparing the COPI versus the SOPI, examinees generally reacted positively to the innovations made. Candidates commented that control over choice of tasks, control over difficulty levels, language of directions and thinking and response time were the best features of the COPI. In fact, one examinee wrote that such choices "help people pick what they might be best at." Another examinee remarked that the option of getting the directions in the target language helped them "get in the right frame of mind" and provided them with vocabulary. Other examinees mentioned that they liked having control of when to start and end speaking, and that they felt that they were given adequate time to think. These aspects of the COPI, particularly the one about being able to adjust the difficulty level of the tasks to the proficiency level of the examinee, were especially salient to lower-proficiency examinees. They felt that the COPI was fairer, less difficult, and made them less nervous, than the SOPI.

The ability to match task difficulty to examinee proficiency is one of the advantages of computer-assisted testing. We believe that in the COPI we have demonstrated its successful application to the assessment of oral proficiency and have investigated its influence on examinee attitude and affect towards technology-mediated tests.

Overall, we were pleased to see that the innovations in the COPI produced significant changes in examinee affect towards technology-mediated oral testing. In general, these were in the direction predicted at the outset of the research project. We believe that this research on examinee affect towards the COPI versus the SOPI lends positive impetus for proceeding with further research in applying advances in computer technology to oral proficiency assessment.

The findings in this study also have implications regarding testing conditions and examinee ability that influence examinee affect. In line with expectations based on Schumann (1997), this study provides some evidence that providing more control and choice to the examinee through an adaptive computer test can have a positive impact on examinee affect, particularly by providing lower ability examinees task more appropriate to their proficiency level. Additional research of the data collected in this project will hopefully shed further light on these issues.

Examinee Views on the OPI Versus the COPI/SOPI

Oral communication remains a human phenomenon. Many examinees commented that many aspects of their speech communication could be captured in a technology-mediated assessment (such as pronunciation, vocabulary, fluency, grammatical control, and even the ability "to think on the spot"). However, they strongly felt that other aspects related to the interactional nature of a conversation could not. This raises the question of what is being assessed in the oral assessment. If criteria used to evaluate performances do not differentiate on examinees' interactional abilities, it may not be necessary to use the

labor-intensive face-to-face interview for the assessment, no matter how the examinees feel about the nature of the assessment. On the other hand, in instances where an evaluation of interactional competence is critical, it may be quite a while before it can be replicated with technology.

That being said, there are instances where it is not practical or feasible to provide a one-on-one interview for some examinees. Indeed, a labor-intensive interview may not be necessary for the kinds of decisions that will be made on the basis of the assessment. In other cases, it may be imperative that examinees be given a common assessment and that differences between the skills of different interviewers to conduct a face-to-face interview cannot be tolerated. In such cases, technology-based assessments may be the best option, as in the case of the *Texas Oral Proficiency Test*, a SOPI used as part of the certification process of Spanish, French, and bilingual education teachers in Texas (Stansfield & Kenyon, 1991).

Overall, we were pleased to see that the COPI functioned so well in the area of its influence on examinee affect. From this study, it appears that it was an improvement over the SOPI in technology-mediated oral proficiency testing. From the examinees' perspective, the innovations sought by the introduction of the COPI over the SOPI seem to have made a difference in the areas of examinee affect we had hoped for.

ABOUT THE AUTHORS

Dorry M. Kenyon is director of the Language Testing Division at the Center for Applied Linguistics in Washington, DC. For over 20 years, he has been involved in second and foreign language teaching and assessment. He has an M.A. in Teaching English as a Foreign Language from the American University in Cairo, Egypt, and a Ph.D. in Educational Measurement, Applied Statistics and Evaluation from the University of Maryland. His current research and development interests lie in applying new technology to language proficiency assessment.

Email: dorry@cal.org

Valerie Malabonga is a Research Associate at the Language Testing Division at the Center for Applied Linguistics (CAL) in Washington, DC. Prior to coming to CAL, she worked for the Vocabulary Improvement Project at Harvard's Graduate School of Education and was a Research Associate/Statistician at the Pacific Institute for Research and Evaluation. She has a Ph.D. in Developmental Psychology from George Mason University. Valerie is currently involved in research assessing the acquisition of English literacy by Spanish-English bilingual children.

Email: valerie@cal.org

REFERENCES

American Council on the Teaching of Foreign Languages. (1986). *Proficiency guidelines.* Hastings-on-Hudson, NY: Author.

American Council on the Teaching of Foreign Languages. (1999). *ACTFL proficiency guidelines -- speaking: Revised 1999.* Hastings-on-Hudson, NY: Author.

Clark, J. L. D. (1988). Validation of a tape-mediated ACTFL/ILR-scale based test of Chinese speaking proficiency. *Language Testing, 5*(2), 197-205.

Clark, J. L. D., & Clifford, R. (1988). The FSI/ILR/ACTFL proficiency scales and testing techniques: Development, current status, and needed research. *Studies in Second Language Acquisition, 10*(2), 129 - 147.

Kenyon, D. M. (1997). Further research on the efficacy of rater self-training. In A. Huhta, V. Kohonen, L. Kurki-Suonio, & S. Luoma (Eds.), *Current developments and alternatives in language assessment: Proceedings of LTRC 96* (pp. 257-273). Jyväskylä, Finland: University of Jyväskylä.

Kenyon, D. M., & Stansfield, C. W. (1993). A method for improving tasks on performance-based assessments through field testing. In A. Huhta, K. Sajavaara, & S. Takala (Eds.), *Language Testing: New Openings* (pp. 90-102). Jyväskylä, Finland: University of Jyväskylä.

Kenyon, D. M., & Tschirner, E. (2000). The rating of direct and semi-direct oral proficiency interviews: Comparing performance at lower proficiency levels. *Modern Language Journal, 84*(1), 85-101.

Kuo, J., & Jiang, X. (1997). Assessing the assessments: The OPI and the SOPI. *Foreign Language Annals, 30* (4), 503-512.

Malabonga, V., & Kenyon D. M. (1999). Multimedia computer technology and performance-based language testing: A demonstration of the Computerized Oral Proficiency Instrument (COPI). In M. B. Olsen (Ed.), *Computer mediated language assessment and evaluation in natural language processing: Proceedings of a symposium sponsored by the Association for Computational Linguistics and International Association of Language Learning Technology* (pp. 16-23). New Brunswick, NJ: Association for Computational Linguistics.

Norris, J. M. (1997). The German Speaking Test: Utility and caveats. *Die Unterrichtspraxis/Teaching German, 30*(2), 148-158.

Schumann, J. H. (1997). *The Neurobiology of affect in language.* Malden, MA: Blackwell Publishers.

Shohamy, E., Gordon, C., Kenyon, D. M., & Stansfield, C. W. (1989). The development and validation of a semi-direct test for assessing oral proficiency in Hebrew. *Bulletin of Higher Hebrew Education, 4,* 4-9.

Stansfield, C. W. (1990). An evaluation of simulated oral proficiency interviews as measures of oral proficiency. In J. E. Alatis (Ed.), *Georgetown University Roundtable on Languages and Linguistics 1990,* (pp. 228-234). Washington, DC: Georgetown University Press.

Stansfield, C. W. (1996). *SOPI test development handbook.*. Washington, DC: Center for Applied Linguistics.

Stansfield, C. W., & Kenyon, D. M. (1991). *Development of the Texas Oral Proficiency Test (TOPT). Final Report.* Washington, DC: Center for Applied Linguistics. (ERIC Document Reproduction Service No. ED 332 522).

Stansfield, C. W., & Kenyon, D. M. (1992a). The development and validation of a simulated oral proficiency interview. *The Modern Language Journal, 76,* 129-141.

Stansfield, C. W., & Kenyon, D. M. (1992b). Research on the comparability of the Oral Proficiency Interview and the Simulated Oral Proficiency Interview. *System, 20,* 347-364.

Stansfield, C. W., & Kenyon, D. M. (1993). Development and validation of the Hausa Speaking Test with the ACTFL Proficiency Guidelines. *Issues in Applied Linguistics, 4,* 5-31.

Stansfield, C. W., Kenyon, D. M., Paiva, R., Doyle, F., Ulsh, I., & Cowles, M. A. (1990). The development and validation of the Portuguese Speaking Test. *Hispania, 72,* 641-651.

Swain, M. (1985). Large-scale communicative language testing: A case study. In Y. P. Lee, A. C. Y. Fok, R. Lord, & G. Low (Eds.): *New directions in language testing* (pp. 35-46). Oxford, UK: Pergamon Press.

Swender, E. (Ed.). (1999). *ACTFL oral proficiency interview tester training manual.* Yonkers, NY: American Council on the Teaching of Foreign Languages.