Multimedia Computer Technology and Performance-Based Language Testing: A Demonstration of the Computerized Oral Proficiency Instrument (COPI)

Valerie A. MALABONGA
Center for Applied Linguistics
4646 40th Street, NW
Washington, DC 20016
valerie@cal.org

Dorry M. KENYON
Center for Applied Linguistics
4646 40th Street, NW
Washington, DC 20016
dorry@cal.org

Abstract

The field of language testing has long led the way in integrative, performance-based assessment. However, the use of technology in language testing has often meant limiting assessment options. We believe computer-mediated language assessment can enrich opportunities for language learners to demonstrate what they are able to do with their second language. In this paper, we describe the rationale and operation of the Computerized Oral Proficiency Instrument (COPI), a multimedia, computer-administered oral proficiency test. While at present speech performances on the COPI are evaluated by trained raters using a national standard, the COPI affords an excellent opportunity to investigate the use of Natural Language Processing for computer-assisted evaluation.

Introduction

The Computerized Oral Proficiency Instrument (COPI) is a multi-media, computer-administered adaptation of the tape-mediated Simulated Oral Proficiency Interview (SOPI). Both the SOPI and the COPI are oral proficiency tests based on the *Speaking Proficiency Guidelines* of the American Council on the Teaching of Foreign Languages (ACTFL). Oral proficiency tests like the SOPI and COPI use simulated real life tasks to elicit speech ratable by the ACTFL Guidelines' criteria. The purpose of the COPI is to use the advantages of multi-media computer technology to improve the SOPI by giving examinees more control over various aspects of the testing situation and increasing raters' efficiency in scoring the test.

In this paper we primarily discuss the Spanish version of the COPI, although an Arabic and a Chinese version are also being prepared. This paper provides the context for the COPI, discusses its rationale, its components and its phases, and introduces the scoring program used by raters who assess an examinee's speech performances using the criteria of the ACTFL Guidelines.

1. Computer Technology in Performance-Based Assessments of Speaking Ability

Technology has no doubt been a part of language testing since before the invention of the pencil. Electronic technology, through the phonograph record, reel-to-reel and later casette tape, and the compact disc, has enhanced the assessment of listening skills for decades. Computers allowed for the development of computer-adaptive and computer-administered tests in second languages. Since June of 1998, the Educational Testing Service (ETS) has administered the *Test of English as a Foreign Language* (TOEFL) by computer in many parts of the world. With almost one million test takers a year, the TOEFL is the world's largest language test. The use of computer technology has allowed ETS to introduce a new variety of selected-response type items not easily presented in paper and pencil format. In addition, the computer-based TOEFL allows examinees the option of word-processing a written essay, as opposed to writing it longhand. Of all sections of the current TOEFL, only the essay can be

regarded as performance-based, since examinees provide a demonstration of their linguistic abilities through producing a text.

While some have argued that multiple-choice tests of listening comprehension can provide a proxy measure of speaking ability, speaking skills have traditionally been assessed through some type of performance-based assessment, typically a live face-to-face oral interview procedure. The best known formal procedure is the Oral Proficiency Interview (OPI). The OPI is used by various government agencies involved with language training, including the Foreign Service Institute, where it was originally developed in the 1950s to assess the readiness of US personnel for functioning in oversees diplomatic posts. In US academia, the American Council on the Teaching of Foreign Languages (ACTFL) has promulgated the OPI since the early 1980s through professional development workshops and tester training programs (Stansfield, 1996).

In the mid-1980s, the Center for Applied Linguistics (CAL) began a program of research and development in using technology to elicit speech samples from examinees that can be assessed following the same criteria used in the ACTFL OPI. The impetus for this program was the need to assess speaking skills of students of less-commonly-taught-languages in instructional programs throughout the nation where there was no trained OPI interviewer. Performances elicited by and recorded on tape could then be sent to trained OPI testers for evaluation. The format developed by CAL came to be known as the Simulated Oral Proficiency Interview (SOPI). High correlations (averaging .92) were found between performances on the SOPI and the OPI across a variety of languages (Stansfield and Kenyon, 1992). The testing format was also found to be useful in large-scale testing applications where it was necessary to ensure that all examinees received the same high quality test, and the SOPI format has been used in or adapted for a variety of language testing projects. Other variations of the SOPI appeared, most notably the Video Oral Communication Instrument (VOCI), developed by the Language Acquisition Resource Center at San Diego State University. The VOCI uses a video rather than an audio tape and test booklet to elicit examinee speech performances.

Based on its work with the SOPI, CAL is currently developing a format for a computer-administered assessment of oral proficiency known as the Computerized Oral Proficiency Instrument (COPI).

2. Importance of National Proficiency Standards

Tape-mediated speaking tests existed prior to the development of the SOPI. One example is the original version of the Test of Spoken English (TSE), developed by ETS and used to assess the oral language skills (particularly comprehensibility) of foreign teaching assistants. There are now many tape-mediated speaking tests or portions of larger tests that assess speaking skills through the use of a tape. Tests including tape-mediated speaking portions include the Advanced Placement Exams in modern languages and the PRAXIS examination used by states to certify language teachers. The main difference between these tests and the SOPI and now the COPI, however, is that such tests are assessed using criteria developed specifically for the exam, whereas the SOPI is assessed using the *ACTFL Speaking Proficiency Guidelines* (American Council on the Teaching of Foreign Languages, 1986, 1999), which exist outside the context of the assessment.

The *ACTFL Guidelines* stand in a tradition of oral proficiency testing in the United States that dates to the 1950s, when the then Secretary of State called for the creation of criteria that could be used to identify the foreign language proficiency of U.S. government employees (Stansfield, 1996). The result was a 0-5 scale, ranging from "no knowledge" to "total mastery," with a brief definition of proficiency associated with each point on the scale. Since their original creation, the definitions have undergone a number of revisions, but are still in use and known today as the *Interagency Language Roundtable (ILR) Skill Level Descriptions.*

In the early 1980s, the government's definitions were adapted and disseminated by ACTFL for

use in the nation's secondary schools and colleges. These definitions have come to be known as the *ACTFL Guidelines*. First published in a provisional version in 1982, they were revised and published for large-scale use in 1986 (American Council on the Teaching of Foreign Languages, 1986). While the *Guidelines* cover all four language skills, the *Speaking Proficiency Guidelines* have been recently revised and re-published in 1999 (American Council on the Teaching of Foreign Languages, 1999).

The *ACTFL Guidelines* define proficiency as "the ability to use the language effectively and appropriately in real-life situations" (Buck, Byrnes, and Thompson, 1989, 1.1). The *Guidelines* posit four levels of proficiency: Novice, Intermediate, Advanced and Superior. The first three levels are further broken into three sublevels: Low, Mid, and High. Thus, the *Guidelines* define 10 levels of proficiency: Novice Low, Novice Mid, Novice High, Intermediate Low, Intermediate Mid, Intermediate High, Advanced Low, Advanced Mid, Advanced High, and Superior.

The *Guidelines* define proficiency in terms of the global tasks or functions the speaker can handle, the contexts in which he or she can effectively communicate, the content about which the speaker can communicate, and the accuracy with which he or she communicates. Accuracy is typically considered in terms of how well the speaker is understood by his or her interlocutors.

Thus, unlike other technology-based speaking tests, the overriding goal of the SOPI and the COPI is to use technology to provide a valid surrogate assessment to the face-to-face OPI. In other words, the performance must be ratable using the criteria of the *ACTFL Guidelines* and an examinee should be assessed at the same ACTFL proficiency level using any technique. Because the *ACTFL Guidelines* have had a major national impact and are so widely used in the US in both academia and government, we feel that this is the best way for our current project to have the greatest national impact.

3. A Collaboration between Examinee and Computer in the Production of the Ratable Speech Sample

The goal of general oral proficiency tests such as the OPI, SOPI, VOCI, or COPI is to allow examinees to demonstrate to a trained rater features of oral language proficiency at one of the main global proficiency levels they consistently control. In other words, examinees demonstrate what they *can do* with the language regardless of how they learned it. In order to make appropriate assessments, the test must give the rater evidence that examinees assessed at a particular level do not control the features of the next higher level of proficiency.

In the OPI, examinees have a certain amount of input into the procedure. The interviewer adapts the level of difficulty of the questions to the proficiency level displayed by the examinee. Examinees control the length of their responses to the interviewer's questions. They have some control over the content of the interview in that the interviewer is trained (particularly at lower levels) to follow up on topics nominated by the examinee.

In tape-mediated tests, much of this control is lost. In general, timed pauses prescribe for the examinee how much time he or she has to think about and give a response. All examinees must perform all tasks presented to them on the tape; there is no selection of the tasks.

The main goal of the COPI is to use computer technology to allow the examinee and computer to work together to produce a speech sample ratable using the ACTFL criteria. The program must enable the examinee to show what he or she can do in a second language. Thus, the COPI allows examinee control over several aspects of the test administration. This is made possible by the large amount of electronic data that can be stored in computers and by the random-access nature of data retrieval. Underlying the COPI is a large pool of assessment tasks that cover a wide variety of content areas and topics. The COPI allows examinees to have control of the time 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 long enough to allow for most examinees to experience control.

The COPI allows examinees some choice in the difficulty level of the tasks presented to them. In order to ensure that examinees are pushed to show the full extent of their ability, the difficulty level of all tasks cannot be examinee-selected. Raters need to hear examinees attempt tasks higher than their proficiency level to ascertain their consistent level of performance. By keeping track of the examinee's choices, a program can ensure that this occurs. Limited examinee selection may, however, assure that the tasks administered are as appropriate as possible to each examinee's level of ability.

4. Description of the Test Administration Program for the Spanish COPI

4.1 Technical Requirements

The COPI program works well with a Windows 95 operating system, or higher, with a Pentium processor and 64 MB of RAM. The examinees' responses can be recorded on internal or external zip drives, or the hard drive. We do not recommend the use of Windows NT or laptop computers because the small memory space in these types of computers makes recording responses difficult.

4.2 Assessment Tasks

At the core of the COPI is a pool of about 100 assessment tasks. These tasks are based on tasks successfully used in SOPIs. Each task has a targeted ACTFL level (Novice, Intermediate, Advanced, or Superior) and is coded for its speaking function and content/topic area. Each task is a separate master computer file composed of a single-sentence description of the task, written and audio directions in English, written and audio directions in Spanish (for tasks at the Advanced and Superior levels), a graphic file of a picture that accompanies the task (for those tasks that have pictures) and an audio prompt from a native Spanish-speaker. Depending on the choices that the examinee makes, the test takes anywhere from 30-50 minutes.

The COPI uses an algorithm which allows examinees (within some limits) to choose the following aspects of the test: amount of preparation and response time, speaking function, topic, level of difficulty (i.e., ACTFL level of task), and language of the directions (English or Spanish for Advanced and Superior level tasks) for each performance task.

Figure 1 shows the screen for the sample task at level B (ACTFL Intermediate).

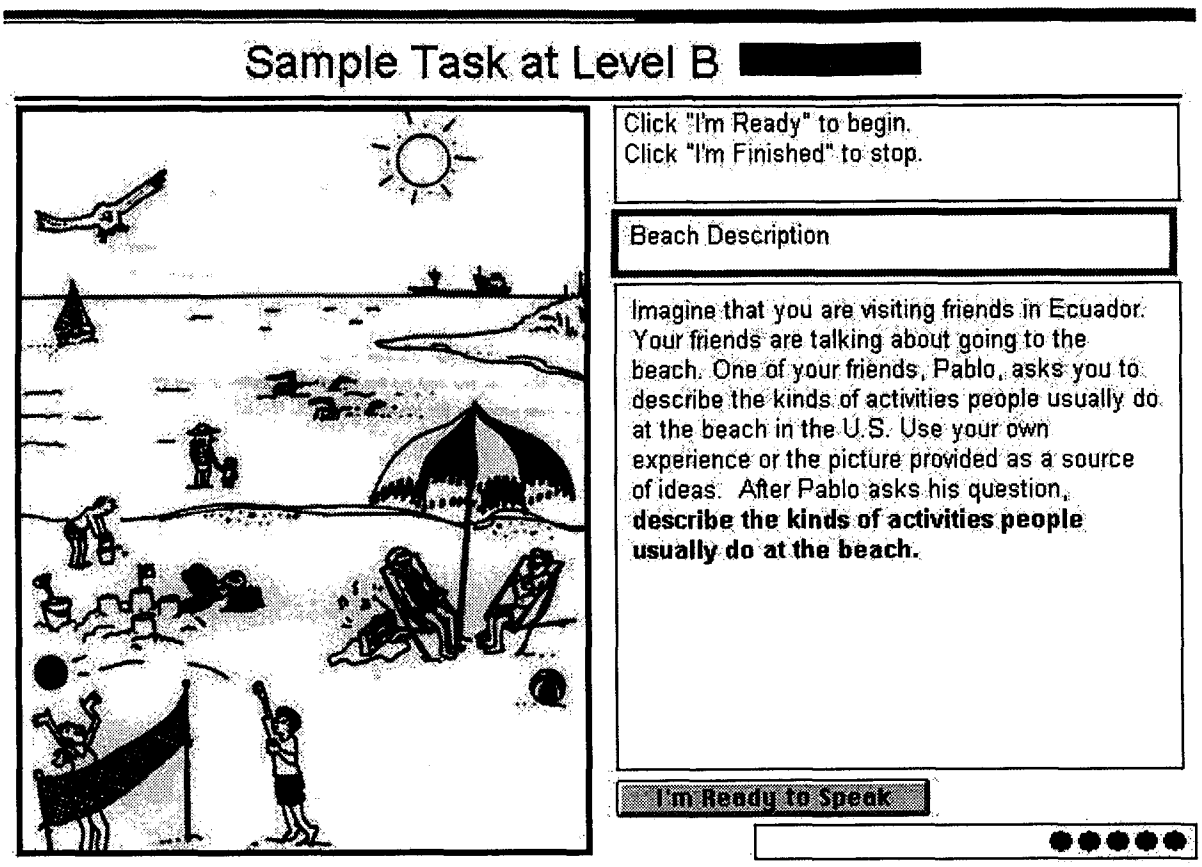

Figure 1

4.3 How Examinee Choice is Operationalized

4.3.1 Time

The COPI provides time for examinees to think about their response and time to give their response. The total amount given for each task is shown by balls in a timer at the bottom right hand of the screen. Each ball represents 15 seconds. More preparation and response time is allotted for higher-level than for lower-lower level tasks, but examinees still have the choice to use all the time allotted or to click on a button when they are ready to speak, or when they have finished speaking. Plenty of time is allotted and in our pilot testing to date, no examinee has indicated feeling pressured by the time factor.

4.3.2 Speaking Functions and Topics

In total, an examinee generally responds to seven tasks on the COPI. Examinees are always given a choice of three tasks, from which they choose one. At the Intermediate level, for example, they may choose from the following task descriptions: "Ask a Spanish exchange student some questions about her family," "Describe your leisure time activities to a visiting Bolivian student." Or "Tell a student from the Dominican Republic about your plans for the weekend." An algorithm in the program ensures that examinees perform each speaking function (e.g., narrating in the past) and talk about each content area/topic (e.g., food) only once during the collaborative development of the speech sample. Examinees are thus exposed to a variety of tasks and can select tasks and topics they feel most comfortable with.

4.3.3 ACTFL Level of the COPI Task

The ACTFL level of the first COPI task administered is determined by the examinees' self-assessment scores and the level of the sample tasks practiced (described more completely in 4.4.3). Following the first task, examinees are given the choice to select tasks at the same level of challenge, a less challenging task, or a more challenging task, after every other task. An algorithm in the program ensures that examinees are offered tasks at a level higher than the one they have generally chosen, to allow the rater to evaluate whether or not they can fulfill the criteria for performance at the next higher level.

4.3.4 Language of Directions

At present, examinees are given the choice to read and hear the directions to the performance task in Spanish or English only for the two highest levels of tasks. Lower level examinees receive all directions only in English. Following piloting testing of the Spanish COPI, however, we will experiment with including both English and Spanish directions for lower-level speakers. This is to provide more target language support for the lower-level speakers.

Results from the pilot test of the COPI showed that the examinees felt more comfortable and less anxious because they were given choices that made the test more flexible. This, we feel, is an improvement from the SOPI, where such choices were unavailable.

4.4 Phases of the COPI

When taking the COPI, the examinee goes through the following 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, female "guide" accompanies the screens. The guide's photograph is present in all the screens that welcome, give instructions, and close the program. Notes on these phases follow.

4.4.1 Welcome/Information on the Purpose and Structure of the COPI

The purpose of these two phases is to introduce examinees to the COPI and help them feel at ease.

4.4.2 Input and Correction of Personal Information

Examinees enter their personal data and are given an opportunity to correct any wrong information. The information is used to identify the examinees and to allow the program to select tasks appropriate to the examinees' profiles. For instance, in Arabic culture (for the Arabic COPI), it is inappropriate for unmarried persons of the opposite sex to do certain activities together (e.g., share an apartment). Therefore, an algorithm in the Arabic COPI ensures that a female or male version of these tasks is presented to the examinee depending on whether the examinee is identified as a female or male, respectively.

4.4.3 Self-Assessment of One's Proficiency Level

Examinees answer 18 questions about their abilities to communicate in the test language; for example, give directions, ask questions, hypothesize, and so on. Kenyon (1996) showed that the correlation between examinees' answers to these 18 self-assessment questions and their actual ACTFL assessments was .78. The COPI program uses examinees' score on the self-assessment to determine at which level they receive the first sample task.

4.4.4 Listening to and Practice on Sample Tasks

Examinees are given an opportunity to listen to an adequate response to a sample task. They are then asked to respond to the same sample task for practice. This is the point in the program at which the directions for navigating the tasks are explained. After giving their performance on the sample task, examinees are asked if they want to practice with a more challenging or a less challenging task before going on to the actual test.

4.4.5 Responding to Performance Tasks

Examinees can select the level of their first performance task based on their self-assessment results and experience with the sample task(s).

The program algorithm is set to ensure that examinees respond to a minimum of four tasks at the level of the first task selected and three at the next higher level (or next lower level if their self-assessment level is already at Superior) for a minimum of seven tasks. Depending on the choices examinees make, however, they can be administered a maximum of 11 tasks, though this is very rare.

4.4.6 Feedback on the Levels of the Tasks that Examinees Took/Closing

After completing the last performance task, examinees receive feedback about the levels of tasks they have taken and are thanked for their participation.

5. Description of the Current Scoring Program

As with the SOPI, performances on the COPI are assessed following the criteria of the ACTFL *Speaking Proficiency Guidelines.* The scoring 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 Spanish prompt appear on the screen. These elements give raters background information about each task and facilitate the assessment of the performances. Raters can also rewind each examinee's response for a particular task and they can likewise 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. In addition, the COPI allows raters to listen to performances on only those tasks that are necessary to give an accurate assessment of the examinees' ACTFL proficiency level, thereby increasing raters' efficiency. For example, if an examinee responded to four Superior tasks and three Advanced tasks, we suggest that the rater start assessing the highest level (Superior) tasks first. If the examinee is clearly a Superior speaker based on his or her performance on the four Superior-level tasks, then it is not necessary for the rater to listen to his or her performances on the three Advanced-level tasks.

6. Opportunities for Interfacing with Natural Language Processing

Performances on the current version of the COPI are assessed by trained human raters. While the COPI scoring program is designed to improve efficiency in rating, assessing speech performances elicited by the COPI using the criteria of the ACTFL Guidelines remains a labor-intensive effort. The COPI harnesses technology to provide examinees an opportunity to demonstrate their oral proficiency without the labor intensity involved on the part of a test administrator (as compared to the individually administered face-to-face OPI). In a similar manner, we feel that this program provides opportunities for interfacing with natural language processing to provide technological assistance in assessing examinee speech performances. While that discussion is outside the scope of this paper, we feel implementation of oral proficiency assessment, particularly for lower-level learners, would increase were it possible for technology to assist in the evaluation of speech performances. An increase in the practicability of large-scale technologically mediated oral assessments has the potential for a great washback effect in our nation's classrooms to promote the development of oral proficiency in second languages. Educational practitioners have long understood that ultimately what gets assessed is what gets taught and practiced.

Conclusion

We believe the COPI offers significant improvements in terms of administration of technologically mediated oral proficiency assessments over tape- or video-mediated assessments. Pilot testing to date indicates that examinees are comfortable with the administration format and understand what is required of them. A validation study is planned for the near future to compare performances on the COPI with those on the SOPI and OPI. Other refinements suggested by the piloting testing are being incorporated into the Arabic and Chinese versions. If such improvements are found to be

helpful through pilot testing, they will be brought into the Spanish version.

Acknowledgments

Support for the development of the COPI is provided by grant P017A70019-98, International Research and Studies Program, International Education and Graduate Program Service, U.S. Department of Education. The work presented in this paper builds on continuing research and development projects in technologically-mediated oral proficiency testing over the past 15 years at the Center for Applied Linguistics. It is impossible to properly acknowledge all who have contributed directly or indirectly to the current efforts. However, we do acknowledge our colleagues at CAL contributing to the COPI project, including Weiping Wu, Kate Jerris, Jane Herlihy, Deanne Marein-Efron, Helen Carpenter, David MacGregor, Charles Stansfield, and Suzy Meyer, COPI programmer.

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.

Buck, K., Byrnes, H. and Thompson, I. (Eds.) (1989) *The ACTFL oral proficiency interview tester training manual.* Yonkers, NY: ACTFL.

Kenyon D. M. (1996, November) *Self-assessment and speaking tasks: Research and application.* Workshop, Annual Meeting of the American Council on the Teaching of Foreign Languages, Philadelphia, PA.

Stansfield, C. W. (1996) *Test development handbook: Simulated oral proficiency interview.* Washington, DC: Center for Applied Linguistics.

Stansfield, C. W. and Kenyon, D. M. (1992) *Research on the comparability of the Oral Proficiency Interview and the Simulated Oral Proficiency Interview.* System, 20, 347-64.