



AMERICAN INSTITUTES FOR RESEARCH

---

**Assessing Students' Conversational Spanish Proficiency using Telephone-Mediated Interaction: A Cognitive Laboratory Evaluation**

Paper presented in "Unique Measurement Features and Challenges in NAEP Foreign Language Instrument Development," J. Brown (Chair), symposium at the annual meeting of the *American Educational Research Association*: New Orleans, LA

**April 4, 2002**

Christine Andrews Paulsen  
Sarah Mushlin  
Steve Ferrara  
*American Institutes for Research*

Margaret E. Malone  
Dorry Kenyon  
*Center for Applied Linguistics*

Elvira Swender  
*American Council on the Teaching of Foreign Languages*

---

## Introduction and objectives

---

In 1999, the National Center for Education Statistics (NCES) awarded a contract to the Educational Testing Service (ETS), the American Institutes for Research (AIR), the Center for Applied Linguistics (CAL), and National Computer Systems (NCS) to develop and field test the first National Assessment of Educational Progress foreign language assessment (NAEP FL). The first administration of the assessment is planned for the year 2004.

The NAEP FL assessment is based on a general framework approved for assessing communicative ability in languages other than English. In this framework, listening, speaking, reading, and writing skills are assessed within three modes of communication: the *interpersonal* mode, which involves two-way, interactive communication; the *interpretive* mode, which relates to the understanding of spoken or written language; and the *presentational* mode, which involves creating spoken or written communication (NAEP Framework for the 2003 Foreign Language Assessment).

The framework states that students' communicative abilities will be assessed through authentic communication tasks that reflect skills required in daily life, school, and work. Assessment tasks will reflect four interrelated goals that provide the basis for communication. These goals include the following:

- Gaining knowledge of other cultures;
- Connecting with other academic subject areas to acquire knowledge;
- Developing insights into the nature of language and culture through comparisons;  
and
- Participating in multilingual communities at home and around the world.

Student performances will be evaluated on how well the students understand (comprehension) and can be understood (comprehensibility). The criterion of comprehension/comprehensibility subsumes language knowledge, the appropriate use of communication strategies, and the application of cultural knowledge.

The interpersonal communication task is being developed collaboratively by staff from the American Institutes for Research (AIR), the Center for Applied Linguistics (CAL) and the American Council on the Teaching of Foreign Languages (ACTFL). These three organizations bring a variety of strengths in developing standardized tests and tests of speaking ability, as well as experience in operating small and large-scale speaking test programs. Specifically, the team is charged with developing a standardized, adaptable test that will determine to what extent students can participate in an interpersonal conversation in Spanish. Such a standardized approach to testing the interpersonal mode in speaking and listening has never been developed before for such a large number of students. As such, AIR and CAL, with extensive input from ACTFL, are collectively responsible for conducting research to aid in the development of the test items and ancillary materials.

A number of approaches currently exist that assess students' oral proficiency in foreign languages. The most common of these is the ACTFL Oral Proficiency Interview (ACTFL OPI), a face-to-face or telephonic interview conducted by a trained interviewer to determine an examinee's language proficiency.<sup>1</sup> This approach is used in academic, governmental, and business contexts. ACTFL OPI interviewers undertake a rigorous training program resulting in high inter-rater agreement. However, while the testing protocol is standardized, the context and content areas are not standardized. In other words, the specific questions asked of examinees differ from interview to interview. Therefore, in developing a standardized test of interpersonal conversation, the OPI was a good starting point. However, the team knew that the interviewer would need a "script," adaptable to different student language abilities, to ensure standardization.

Developing this innovative approach to testing requires a great deal of input from leaders in the fields of foreign language assessment and standardized testing. Therefore, the NAEP-FL Standing Committee has reviewed different versions of approaches to the interpersonal conversation task. Based on the Standing Committee's

---

<sup>1</sup> Swender, E. 1999, ACTFL Oral Proficiency Interview Tester Training Manual. Yonkers, NY.

input, the team changed the initial approach to this test from a face-to-face test to a telephone delivered test. This test provides context to students by explaining to them, through a letter in Spanish and English, that they will be participating in a conversation with a teacher from Santiago, Chile who is interested in setting up a two-way exchange program between the teacher's school in Chile and the students' respective school in the United States. Students are also asked to choose two topics, from a list of five topics, upon which they would like to ask the interviewee to elaborate, again in Spanish and English. Therefore, the test and ancillary materials now consist of a script adaptable to different student language abilities, a letter from the teachers in Santiago de Chile (English and Spanish) and a sheet describing the kinds of tasks students can choose to discuss in the latter part of the test.

To develop the test and materials, between July 2001 and April 2002, AIR and CAL conducted a series of three small-scale tryout/cognitive laboratory studies with Spanish conversational tasks. The objectives of each component of the study are summarized below:

- **Small-scale tryout #1** was designed to determine whether or not it is possible to administer Spanish conversational tasks by telephone, and whether or not these tasks are performing as intended for both high and low performing Spanish speakers.
- **Small-scale tryout #2** was designed to determine whether or not it is possible to train, in a standardized way, several Spanish-speaking interviewers to administer conversational tasks by telephone, and to further validate the use of the conversational tasks with high and low performing Spanish speakers.
- **Small-scale tryout #3** was designed to evaluate, and make refinements to, the scoring procedures, and to further refine and validate the interviewer training and conversational tasks.

The purpose of this paper is to discuss the development of the NAEP FL conversational assessment tasks, including a discussion of the cognitive evaluation methods the team used to assess the effectiveness and feasibility of the materials and procedures.

## Methods and Procedures

---

### Student Participants

Students were screened, prior to participation in the study, using a conversational proficiency screener developed by AIR, in order to categorize students as either medium to low proficiency or high proficiency. For example, high proficiency students were able to discuss what they plan to be doing in five years using appropriate future tenses, state and support a position on a complex topic with a person who has an opposing view (i.e., debate), and understand subtitles in movies. Medium to low proficiency students, however, were able to handle only simple tasks, such as counting to ten in Spanish.

We recruited 18 participants for small-scale tryout #1 in August 2001. Twelve students were classified as high proficiency, while six were classified as medium to low proficiency. For small-scale tryout #2, we recruited 23 students—10 high proficiency and 13 medium to low proficiency (November 2001). Thirty-six students participated in small-scale tryout #3—18 high proficiency and 18 medium to low proficiency (January 2002).

### Language Proficiency Interviewers

Since the objective for small-scale tryout #1 was to focus on the validity of the tasks and not on standardizing the interview, we employed one highly qualified Spanish-language interviewer, who is ACTFL-certified in Spanish and has fifteen years of experience in oral proficiency testing, to conduct all the interviews. Because this interviewer was part of the interpersonal communication task team, she adapted readily to changes made to the script throughout the process. In addition, her participation enabled us to control for variation in interviewer quality.

For small-scale tryout #2, the following criteria were used to hire interviewers:

- Spanish proficiency (e.g., at least ACTFL-Advanced in Spanish);
- Some experience with oral proficiency rubrics;

- Availability for training and interviewing times;
- Availability to conduct telephonic interviews in Spanish (training provided) at AIR's Georgetown office; and
- Pleasant conversational manner to put student interviewees at ease.

Potential interviewer candidates were screened by the Project Coordinator to determine how well each met the criteria listed above. An attempt was made to balance male and female interviewers; however, there were very few male applicants. Ultimately, the four interviewers chosen at least minimally met the criteria. All were women. Three were former classroom Spanish teachers currently working as Spanish tutors, and one was a Ph.D. student in Spanish who has taught Spanish at the high school and university levels.

Although all interviews collected during small-scale tryout #2 appeared ratable, team members and the NAEP FL Standing Committee felt that the qualifications for interviewers should be changed. Therefore, all interviewers for small-scale tryout #3 were required to have previous experience in oral testing, and preference was given to those with ACTFL OPI training. The Project Coordinator screened potential interviewer candidates according to the qualifications. Again, the team made an attempt to balance male and female interviewers; however, no qualified male applicants were available for the dates of the training and interviewing. So, again, all four interviewers were women. Two interviewers were native speakers of Spanish. Three currently teach Spanish in public schools, and one is the foreign language supervisor for a metropolitan, DC county.

For both small-scale tryouts #2 and #3, all interviewers participated in a two-day training session. The training reviewed the purpose of the NAEP and included some sample interviews from previous small-scale tryouts to help interviewers understand how to conduct the interview. In small-scale tryout #2, interviewers conducted two practice interviews with AIR or CAL staff prior to operational interviewing. In small-scale tryout #3, interviewers conducted four or more practice interviews with examinees from different proficiency levels via telephone.

## **Interviewing and Cognitive Laboratory Procedures**

The studies were conducted at AIR's Washington, DC facility. After signing an informed consent form, students sat in an office with a trained observer, while a trained language interviewer sat in a separate office. The students were told that they would be participating in a half-hour telephone conversation, in Spanish, with a teacher at the Colegio Andres Bello in Santiago, Chile. Students were provided with letters of introduction in both English and Spanish that summarized the objective for the telephonic interview. The letter stated that the teacher was interested in exploring the possibility of a two-way student exchange program with a school in the U.S. Students were instructed to have a conversation with the teacher about various topics related to student life at their own school, and to learn more about student life at the Colegio Andres Bello. Students were asked to choose two topics for discussion (from a list provided to them in Spanish and English).

The telephonic interviews began with a brief warm-up interchange between the Spanish-language interviewers and the students. This was followed by three social conversation tasks in which interviewers discussed specific facets of student life with the students (e.g., social activities, food, sports, working). Most of the discussion during the social conversation tasks was initiated by the interviewers in an attempt to elicit information and discussion from the students (i.e., the interviewers asked questions and the students responded). These discussions were not scripted, but were standardized.

The next major segment of the interview included two role play tasks, in which students were instructed to initiate discussion with respect to the two topics they had chosen at the beginning of the interview session. During this segment, interviewers were expected to take less of an active role in the interchange in an attempt to help students demonstrate their ability to accomplish a specific task through the negotiation of meaning. Part of interpersonal communication is the negotiation of meaning on the part of both the examinee and the interviewer. Simply responding to questions posed by the interviewer does not constitute two-way communication. Therefore, this part of the interview was designed to provide a more natural opportunity for examinees to ask

questions of the interviewer. Because of the context of the interview—that the examinee needs to learn about the school in Santiago and the teacher needs to learn about the examinee’s school—this part of the interview was intended to elicit questions from the students about the school in Santiago and Santiago in general.

Trained observers stayed with the students throughout the interviews to:

- track the timing of each interview segment;
- note any difficulties conducting the assessment by telephone;
- note interviewer behavior that appeared to deviate from the script and expectations for the task;
- note interviewer behavior that appeared to impact (positively or negatively) the student’s performance;
- record student behavior that appeared to impact negatively the interviewer’s performance; and
- debrief students with a set of guided recall probes (described below).

### **Debriefing Students with Guided Recall (Cognitive Laboratory)**

The cognitive laboratory method utilizes procedures intended to assist testers in understanding respondents’ thought processes as they respond to questions. Trained observers watched each interview session and, immediately following the interviews, “walked” participants back through each segment of the interview and asked specific questions and probes to understand the cognitive processes students used in responding to questions. The cognitive labs provide a technique to examine how well individual students understand the instructions and test questions for the NAEP FL Conversational Tasks.

Over the past several years, AIR has utilized the think aloud procedure and cognitive laboratory methods to evaluate the construct validity of assessment items and survey questions, and to improve them.<sup>2</sup> In asking students to think aloud as they respond to assessment items or tasks, one gains insight into the cognitive processes they employ as they work toward a final response. One may then compare the paths students take and the cognitive strategies they employ to the constructs the assessment items were intended to elicit. Through directed probing after students provide a complete response, one can gather even more data about characteristics of the items that may be affecting students' comprehension, recall, or their ability to synthesize a response.

To gather these evaluation data, interview protocols were designed to help cognitive laboratory observers identify problems (and promising practices) related to:

- the wording of the letter or instructions;
- interviewer questions;
- specific words or phrases used;
- managing the telephone;
- student cognition problems;
- timing problems; and
- any behavior that appeared to inhibit or help the student or interviewer.

Sample prompts and probes (directed to students) included the following:

1. *What did you think about taking an "assessment" over the phone? (PROBES: How did it compare to having an in-person conversation in Spanish? Was there anything about Senora Gonzales that made it easy or difficult to understand her*

---

<sup>2</sup> Paulsen, C.A., Best, C., Levine, R., Milne, A., & Ferrara, S. (1999). *Lessons learned: Results from try-outs of items in cognitive labs*. Paper presented in Analyses Used to Guide Development of the Voluntary National Tests, M. Feuer (Moderator), invited symposium at the annual meeting of the National Council on Measurement in Education: Montreal (April 21, 1999).

Paulsen, C.A. & Levine, R. (1999). *The applicability of the cognitive laboratory method to the development of achievement test items*. Paper presented in Research in the Development of Tests and Test Items, C. Welch (Chair), B. Zumbo (Discussant), at the annual meeting of the American Educational Research Association: Montreal (April 23, 1999).

*over the phone? Do you think it would have been any different for you if she had talked with you in person?)*

2. *When you were reading this letter for the first time, what did you think the letter was asking of you? (PROBE: Did you understand what information Senora Gonzales was requesting? Did you understand what was going to be expected of you?)*
3. *The first topic you chose was (FILL IN). How did you choose that topic? (PROBE: Is that something you know a lot about? Did you think about whether you would be able to use a lot of Spanish words to discuss this topic?)*
4. *Were you satisfied with the number of questions you were allowed to ask to get the information you wanted about this topic? Why or why not? (PROBE: Did you feel like you asked all the questions you wanted to? Do you feel like you answered her questions the way you wanted? How did you know when to stop the discussion?)*

For small-scale tryout #2, we designed the protocols to examine how interviewers use the training and materials to administer the assessment—especially how they make decisions about which paths to follow during an interview. The protocols were also designed to explore:

- difficulties conducting assessment by telephone;
- timing of the individual parts;
- interviewer behavior that appeared to deviate from the script and expectations for the task;
- interviewer behavior that appeared to impact (positively or negatively) the student's performance; and
- student behavior that appeared to impact negatively the interviewer's performance.

Sample prompts and probes (directed to interviewers) included the following:

1. *Do you feel that the student was ready to begin the conversational tasks after the warm-up? How did you decide when the student was ready to proceed?  
(PROBE if the interviewer had difficulty making a judgement: What training would help you decide whether students are ready after the warm-up?)*
2. *Based on the script as it is written, did you have any trouble deciding which follow-up questions to ask?*
3. *What additional training would you like to have to be an effective interviewer?*

For small-scale tryout #3, we asked students to complete a survey about their experiences, rather than engaging them in guided recall since the third study will focus on scoring procedures. Interviewers also completed an extensive survey about their interviewing experience.

## Selected Findings and Implications for Telephone-mediated Assessment

---

### Telephone-mediated assessment is feasible

Our study demonstrates that it is feasible to conduct a conversational assessment by telephone. Despite some occasional, minor technical difficulties, all of the interviews were administered successfully. That is, interviewers and students were able to complete the full range of tasks within the expected amount of time.

Some students may perceive telephone-mediated assessment as more difficult than in-person interviews. This perceived increase in difficulty is due to the lack of non-verbal cues and the reality that most language-learners have never before been assessed via telephone. For example, in small-scale tryout #3, only six students indicated they had ever participated in a tape-recorded or mediated speaking test. Most students had participated in such assessments only rarely (1-2 times per year) or occasionally (3-4 times per year). However, we should note that some students in our study actually *preferred* the telephone method because it relieved them of the perceived pressure of a face-to-face interview.

In small-scale tryout #3, students were asked a variety of questions regarding their opinions on how well the test went for them, including how comfortable they felt with the test, and how well they felt the interviewer conducted the test. The majority of students (over 90%) responded that they either “strongly agreed” or “agreed” that the conversation went well. Fifty-eight percent of students (n = 21) forgot, at times, that they were participating in an assessment. Eighty-three percent of all students (n = 30) felt comfortable throughout the test. Only five students (14%) indicated disagreement with the statement “I felt comfortable throughout the test.”

All students who completed the questionnaire either “strongly agreed” or “agreed” that the interviewer conducted the interview with ease and confidence. All students who

completed the feedback form also “agreed” or “strongly agreed” that the teacher gave sufficient feedback to indicate that she was listening.

All students “agreed” or “strongly agreed” that the letter (provided to students at the beginning of the conversational task) was easy to read/understand. Thirty-four students “agreed” or “strongly agreed” that the letter helped them to understand the purpose of the conversation. Only one student disagreed with this statement.

### **Background, training and practice help develop confident interviewers**

As mentioned earlier, the interviewers in small-scale tryout #2 had less experience in conducting Spanish language interviews with students than interviewers in small-scale tryout #3. Three of the four interviewers in small-scale tryout #3 had participated in some kind of ACTFL OPI training, and they expressed more confidence overall than interviewers in small scale tryout #2.

Training and practice also emerged as important issues. Small-scale tryouts #2 and #3 both included two days of training. However, in small-scale tryout #2, interviewers conducted only two practice interviews, neither of which occurred on the telephone. Interviewers in small-scale tryout #3, however, conducted at least four practice interviews, all on the telephone. In addition to giving interviewers more experience in conducting interviews, training and practice acquaints interviewers with the telephonic equipment and tape recorder they will be using.

### **To be successful, tasks and interviewers must be as authentic as possible**

The NAEP FL conversational assessment tasks need a defined context to give the conversation more authenticity. To make the experience more authentic, CAL developed an informational dossier. This provided some realistic contextual information for the interview and helped to standardize the information interviewers’ provided to the students. As such, this authenticity provided purpose and defined goals for the conversation.

With respect to authenticity, interviewers should be native speakers of Spanish or they must be highly proficient. In this study, three out of the four interviewers were non-native speakers of Spanish. Observers noted that a few interviewers often made slight Spanish mistakes or did not have authentic accents. As a result, both the interviewers' perceived credibility and the students' engagement in the made-up scenario were negatively impacted.

### **Conversational assessment tasks can be standardized**

It is possible to standardize the conversational task (with some flexibility built-in), but training materials must be detailed and explicit about performance expectations and provide guides for decision-making throughout the interview tasks (e.g., when to use elaboration versus expansion questions). During language proficiency interviews, one can expect slight deviations from the interview script. These can be minimized, however, by providing specific options to help interviewers handle special situations. In our study, such deviations were demonstrated most frequently with students who were performing at zero-level and not comprehending most of the interview (called simplification questions in this project). Due to the specific nature of the simple questions used for following-up with zero-level students, the interviewer must be trained to be deliberate when choosing a question so that it fits within the context of the conversation and can be handled by the student. Developers must also ensure the questions are appropriate to each specific student's demonstrated level of proficiency.

The small-scale tryouts helped us to improve substantially the interviewer training for small-scale tryout #3. Revisions were based on feedback from interviewer training evaluations, the guided recall protocol conducted with students and interviewers after each interview, and feedback from the Standing Committee.

Revisions included the following:

- The development of a training manual for interviewers;

- The design of directed practice exercises based on audio segments from sample interviews from small-scale tryout #2;
- Additional opportunities for practice interviewing (e.g., all interviewers conducted four or five practice interviews, three or four over the telephone and one in front of the group, prior to beginning with students);
- Improvements to the “dossier” of information on the interview so as to reflect actual questions posed by students during the small-scale tryout; and
- Improvements to the Spanish letter given to the students in order to provide a better context for the interview situation.

Overall, the four interviewers usually agreed that the interview went well, that they developed a good relationship with the student, and that they felt comfortable using the telephone. Only two interviews were reported to have technical problems. Feedback from interviewers on the effectiveness of the dossier and other materials allowed CAL staff to make iterative changes to the materials.

All interviewers “agreed” or “strongly agreed” that the training was sufficient.

The focus of the cognitive laboratory evaluation of the scoring procedures will be to examine the extent to which the tools and training provide scorers with the information they need to make reliable decisions about student performance.

## Summary

---

This study has demonstrated that it is possible to successfully develop and administer a telephone-mediated language proficiency interview, and, specifically, that it is feasible for conversational tasks of the NAEP FL to be administered telephonically. To date, telephonic mediation is a common methodology used by OPI for language proficiency testing. However, OPI assesses adults, specifically individuals at the higher levels of proficiency. Moreover, telephone-mediated conversational assessment has never been utilized for a standardized assessment. The telephone-mediated NAEP FL conversational task represents the first time that a foreign language assessment has been designed to test younger populations, at all levels of proficiency, in a standardized manner.

The NAEP FL conversational assessment also represents a more cost-effective and efficient assessment mode than in-person administration. In addition to saving resources and avoiding tester fatigue, telephone-mediated assessment increases the possibility of standardization across test administrations. For example, fewer numbers of highly trained interviewers may conduct more interviews in a shorter period of time using this method and it is simply easier for interviewers to follow the script closely because students cannot see the interviewer in person.

It is important to recognize that telephone-mediated conversational assessment is measuring a slightly different set of theoretical constructs than face-to-face conversational assessment. It is likely that telephone-mediated assessment is actually more authentic because no nonverbal communication is being measured, resulting in less interviewer bias. This authenticity will remain essential throughout the assessment's developmental process and the reporting of results.

Overall, these results show that many students were comfortable participating in the interview, and that it resembled a two-way interaction rather than a test. Moreover, the

results demonstrate that the script and training program that the team developed are successful in training interviewers to conduct assessments of students' oral language.