

*Third National Research Symposium on Limited English Proficient Student Issues:  
Focus on Middle and High School Issues*

# **An Approach to Performance Testing of Oral Language Proficiency for Bilingual Education Teacher Certification <sup>1</sup>**

**Charles W. Stansfield**  
*Center for Applied Linguistics*

## **Abstract**

This paper describes the development and implementation of the Texas Oral Proficiency Test (TOPT) for use throughout the state of Texas. The TOPT follows the format of the Simulated Oral Proficiency Interview (SOPI), which has demonstrated good reliability and validity in several research studies. Effective November 1991, the TOPT replaced the OPI which had been used in Texas since 1979.

The paper is divided into sections that describe: (a) the TOPT test format; (b) a research study that determined the speaking tasks that teachers felt were appropriate for a test for teacher certification candidates; (c) the test development and field testing process; (d) a content validation study; (e) a study that determined what proficiency level should be considered passing performance on the test; and (f) information concerning the operational program.

## **Introduction**

## **Background**

Bilingual education is often defined as "the teaching of a substantial part of the regular school curriculum in a non-English language." This definition implies that the teacher in the bilingual classroom will be proficient in a language other than English. Yet inadequate attention has been directed to ensuring that bilingual teachers are indeed bilingual. The matter is important because if the teacher's second language proficiency is inadequate, the teaching of content in the second language will be impaired. In addition, if the teacher is not proficient in the second language, nearly all instruction will be carried out in the teacher's first language.

Bilingual educators in Texas have been required to demonstrate oral proficiency in a second language (Spanish) for many years. In 1978, the bilingual education unit of the Texas Education Agency (TEA) included the training of oral proficiency interviewers in its Title VII proposal to the U.S. Department of Education. Utilizing these funds, the TEA contracted the Educational Testing Service (ETS) to set up and administer the Language Proficiency Interview (LPI) program locally. Staff from regional educational service centers in Texas were trained to administer and score the Language Proficiency Interviews as part of

their regular duties (Stansfield, 1980). Eventually, the task of interviewing was passed on to Texas university faculty members. Each interview was taped and then sent to the ETS Austin office, which had the tape rated by another interviewer or rater located in a different part of the state.

While the LPI program operated for over a decade, it suffered from a number of problems. First, matching interviewers with examinees throughout the state presented some difficulties because the process was inconvenient for both interviewer and examinee. Problems of reliability and validity were also reflected in both the interviews and ratings. For example, LPI raters would often indicate that an interview had not been competently accomplished; in other words, it had not produced a ratable sample of speech. In addition, interviewers sometimes disagreed with the rating assigned by raters in another part of the state. Similarly, instructors familiar with the LPI scale and the proficiency level of their students often disagreed with assigned ratings, frequently complaining that students were given a passing level when their true proficiency level was less than that required to pass.

The Division of Teacher Assessment of the TEA was aware of the problems with the LPI. In an effort to improve the current situation for bilingual educators, in January 1990, it issued a request for proposals (RFP) to develop a new testing program. The RFP called for the development of Spanish and French oral proficiency tests to certify Spanish, French, and bilingual education teachers.

The Division of Foreign Language Education and Testing of the Center for Applied Linguistics proposed to utilize the Simulated Oral Proficiency Interview (SOPI) format which it had designed and had already applied very successfully to the development of semi-direct tests in five languages. The SOPI format has been shown to correlate as highly with the OPI as the OPI correlates with itself (Stansfield, 1989). In addition, it offers greater standardization and control, which is important in a large scale testing program, and especially important in high-stakes tests such as those used to certify teachers. Parallel forms of the SOPI can be developed to alleviate the concern about security that occurs when only one form is available. Yet, parallel forms of the SOPI, unlike different interviewers, can be developed under strict guidelines with subsequent pilot tests and revisions to ensure that the forms are comparable.

The rating of a SOPI is facilitated by the fact that all examinees take the same test. Under these circumstances, it is easier to place examinees on the ACTFL scale. To illustrate how a SOPI facilitates reliable rating, a parallel can be drawn with the scoring of essays. Using any given scale, a packet of essays on the same topic will normally be rated more reliably than a packet of essays on different topics. The SOPI is invariant across examinees at the same administration, while the OPI is not.

Because the SOPI seemed to offer significant controls over reliability and validity, in April 1990, CAL was awarded a contract to develop three forms of a SOPI in Spanish and three forms in French for Texas educators. Because these particular SOPIs would become the property of the TEA, and because they were designed for a particular population (educators), it was decided to name them Texas Oral Proficiency Test (TOPT) to distinguish them from CAL's several SOPI Speaking Tests (e.g., Hebrew Speaking Test).

When implemented in the fall of 1991, the TOPT replaced the Language Proficiency Interview (LPI). As of that date, all persons seeking either an endorsement or a certificate in bilingual education in the state of Texas must pass the TOPT. Although occasional references will be made to the TOPT forms which exist in both French and Spanish, the focus of our discussion will be on the Spanish TOPT. Similarly, while the TOPT is used in the certification of Spanish, French, and bilingual education teachers, this paper will only present data relevant to its use in the certification of the latter group.

## Description of the TOPT

The TOPT is a semi-direct, tape-mediated test of oral language proficiency which may be taken by groups in a language laboratory or by individuals using three cassette tape recorders. The examinee hears the directions and items for all parts of the test from a master test tape. Directions and items are also written in the test booklets. In addition, in three of the four parts of the test, the examinee uses pictures in the test booklet to answer items. All examinee responses are recorded on a separate examinee response tape.

Because the TOPT is a test of speaking ability, the general directions to the test and the directions for each item are in English. However, each item concludes with a target language question or statement heard on the master tape. Following the English directions and in response to this target language prompt, all examinee responses are spoken in Spanish into the microphone and recorded on the response tape.

The master test tape sets the pace of the test, which lasts approximately 45 minutes. The examinee speaks Spanish for approximately 20 minutes during timed pauses throughout the test. The examinee response tape is subsequently evaluated by trained raters within two weeks following the examination.

The TOPT consists of a warm-up section followed by fifteen items designed to allow the examinee to demonstrate his or her ability to perform a variety of speaking tasks covering a variety of topics and situations. All directions are given in English. After the examinee hears the directions for each item, he or she is given between 15 and 30 seconds to prepare a response. Then, after hearing a statement or question in Spanish, the examinee responds in the time allowed.<sup>2</sup> The following paragraphs provide some additional information on the TOPT.

### Warm-Up

A warm-up follows the reading of the general directions. This section is designed to put the examinee at ease, to allow the examinee to make the transition to speaking in the target language, and to become accustomed to the test situation. Therefore, the warm-up is not scored by the rater. In the warm-up, a native speaker of Spanish asks the examinee several personal background questions, involving his or her educational background, interest in teaching and experience with the language.

### Picture-Based Items

A set of five picture-based items follows the warm-up. The picture-based items are designed to permit the examinee to demonstrate the ability to organize discourse in a way that would permit him or her to describe a place, to give directions, and to narrate events in present, past, and future time.

### Topic Items

The next set of five items allows the examinee to demonstrate the ability to speak about a variety of topics. The examinee is asked to discuss advantages and disadvantages of a certain proposition, such as using public transportation, to give someone step by step directions on how to do something, to present a brief factual summary on a familiar topic such as current events, and to present and support an opinion on a topic related to society or education.

Topic items are psychometrically appropriate for examinees at the Advanced and Superior levels on the ACTFL scale because they require the examinee to perform speaking tasks that are indicative of the kinds of language skills that examinees at these levels are expected to have. Topics involving formal language, such as a speech to a group or a professional discussion, are appropriate for the Superior level examinee. While examinees at lower levels are able to respond to each item, the linguistic and rhetorical characteristics of their performance illustrate the limitations in their speaking ability.

## Situation Items

The final set of five items allows the examinee to demonstrate the ability to respond to real-life situations. The examinee may be asked to give advice to a friend, apologize to someone, lodge a complaint, resolve a problem, or attempt to convince someone to take a different course of action. All situations require the ability to tailor one's speech in a sociolinguistically appropriate manner to the individuals and circumstances presented in the item. Like topic items, situation items on the TOPT are designed to assess the speaking ability of examinees at the Advanced and Superior levels, although some items designed for the Intermediate level examinee may be included.

## Job-Relatedness Survey

### Background

In 1978, the Equal Opportunity Commission published the *Uniform guidelines on employee selection procedures*. These guidelines emphasize the need for evidence that the content of a test is related or relevant to the job for which the individual is being considered. In the subsequent years, test developers, first Educational Testing Service which publishes the National Teachers Examination (NTE), and then other testing organizations as well, have collected evidence of job relevancy by putting together a panel consisting of 10-20 individuals (NTE Programs, 1989). These individuals are selected to represent classroom teachers, program administrators, and teacher education faculty. The members of the panel then judge each item on the proposed test and rate whether the item tests something that is important for beginning teachers to know or be able to do (Rudner, 1987).

This approach exhibits several problems. First, the panel is not truly representative of practitioner viewpoints. That is, this approach employs a panel whose members have been previously selected, usually by staff at the state education agency. Thus, the panel members are not randomly selected. Because they are not randomly selected, the members may use different standards than the average teacher in judging items. Panel members may be selected because they are professionally active, in which case they may be inclined to believe that the beginning teacher needs to be very competent. Or, panel members may be selected because they are "cooperative." In this case they would also be inclined to validate test items. On the other hand, if individuals participating in the process are randomly selected, then their ratings as a group are more likely to represent those of the average practitioner.

Another problem with this currently popular approach is that the size of the panel is inadequate for it to be representative of teachers at large. A sample of 20 does not represent a population of thousands. Thus, in order to have a truly representative validation of the content relevance of a test, it is necessary to sample a much larger number of teachers. As a basic guideline, the sample should include at least five percent of the population.

A third problem with the current approach is that it analyzes a test after it has been written, rather than asking teachers to provide data that can be used in the test development process. If teachers are surveyed to find out what knowledge or skills are important for a beginning teacher to have prior to the writing of test items, then this information can be used to design the test, and the finished product will be more appropriate. The knowledge and skills that are validated by teachers can then become part of the content specifications for the test and parallel forms can be created to match those specifications.

We implemented these ideas about appropriate procedures for establishing job-relatedness in our development of the TOPT.

## Preparing the Survey

In order to ensure that the speaking tasks to be included on the TOPT were appropriate for the population of examinees for which the TOPT was intended (prospective Texas classroom teachers), it was necessary to conduct a job-relatedness survey. To do this, CAL staff, with assistance from a local test advisory committee, developed a list of 38 speaking tasks based on the ACTFL Guidelines to be distributed in survey form. The speaking tasks, which were presented in random order, ranged in ability level from Intermediate Low on the ACTFL scale (e.g., "Introduce yourself") to Superior (e.g., "Explain a complex process in detail"). Examples of the speaking tasks can be found in the copy of the machine-readable survey response sheet included as Appendix A.

The task of the respondents was to indicate, on a scale of 1 (E) to 5 (A), whether the level of ability required to perform each speaking task is needed by bilingual education teachers in Texas. In other words, respondents were asked if they believed bilingual educators should possess the level of ability to perform each specified task. Respondents indicated their answers by marking the appropriate column on a machine-readable response sheet. Their choices were:

**A = Definitely Yes**

**B = Probably Yes**

**C = Maybe**

**D = Probably No**

**E = Definitely No**

Respondents were also requested to provide certain basic demographic data. The questionnaires sent to the three groups of teachers (Spanish, French, and bilingual education) were identical except for direct references to group membership.

## Survey Distribution and Results

A random sample of 400 certified Bilingual Education classroom teachers was selected to receive the survey. This number represents approximately 6 percent of the certified Bilingual Education teachers in Texas.

Of the 400 surveys sent to bilingual education teachers, 240 were returned to CAL for a response rate of 60 percent. Of these, nine were incomplete and two were late. Table 1 gives the demographic statistics of the 229 bilingual education teachers whose responses could be tallied.

**Table 1**  
**Demographic Statistics of the Respondents to the Job-Relatedness Survey (Bilingual Education)**

**A. Current Level of Assignment**

Level of Assignment	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Elementary	211	96.3	211	96.3
Junior High/ Mid School	3	1.4	214	97.7
High School	3	1.4	217	99.1
Other	2	0.9	219	100.0

(Frequency Missing = 10)

**B. Certificate or endorsement in bilingual education held?**

Certificate	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Yes	172	78.5	172	78.5
No	46	21.0	218	99.5

(Frequency Missing = 10)

**C. Years of Experience**

Experience	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1-2 years	48	22.2	48	22.2
3-5 years	41	19.0	89	41.2
6-10 years	60	27.8	149	69.0
11-15 years	44	20.4	193	89.4
16-19 years	16	7.4	209	96.8
20 or more	7	3.2	216	100.0

(Frequency Missing = 13)

**D. Level of Class Taught**

Class Level	Frequency	Percent	Cumulative	Cumulative
-------------	-----------	---------	------------	------------

			<b>Frequency</b>	<b>Percent</b>
<b>Early Childhood</b>	39	18.4	39	18.4
<b>Grades 1-3</b>	137	64.6	176	83.0
<b>Grades 4-6</b>	35	16.5	211	99.5
<b>Invalid Response</b>	1	0.5	212	100.0

(Frequency Missing = 17)

#### **E. Highest Degree Held**

<b>Highest Degree</b>	<b>Frequency</b>	<b>Percent</b>	<b>Cumulative Frequency</b>	<b>Cumulative Percent</b>
<b>No Degree</b>	22	10.0	22	10.0
<b>Bachelor's</b>	140	63.9	162	74.0
<b>Master's</b>	57	26.0	219	100.0

(Frequency Missing = 10)

#### **F. Ethnicity**

<b>Ethnic Group</b>	<b>Frequency</b>	<b>Percent</b>	<b>Cumulative Frequency</b>	<b>Cumulative Percent</b>
<b>Hispanic</b>	190	87.2	190	87.2
<b>Black</b>	3	1.4	193	88.5
<b>White</b>	24	11.0	217	99.5
<b>Other</b>	1	0.5	218	100.0

(Frequency Missing = 11)

#### **G. Sex**

<b>Sex</b>	<b>Frequency</b>	<b>Percent</b>	<b>Cumulative Frequency</b>	<b>Cumulative Percent</b>
<b>Male</b>	23	10.5	23	10.5
<b>Female</b>	196	89.5	219	100.0

(Frequency Missing = 10)

Table 1 indicates that the typical respondent was a Hispanic female with a bachelor's degree and 6-10 years of teaching experience. She holds a certificate or endorsement in bilingual education and teaches in an

elementary school (grades 1-3). Only 12.8 percent of the respondents were not Hispanic.

Table 2 presents the results of the job-relatedness survey for the bilingual education teachers. The tasks are ordered by average mean ranking; that is, the tasks which a high number respondents thought were most important for bilingual educators are ranked highest. The standard deviation is presented in the second column as an indication of the agreement or disagreement of the group on the mean ranking; the smaller the standard deviation number, the greater the amount of agreement on the ranking of that particular task. The third column presents the approximate ACTFL level (I=Intermediate, A=Advanced, S=Superior) of the speaking task. The final column shows the speaking task.

**Table 2**  
**Results of the Job-Relatedness Survey for Bilingual Education**

Mean	SD	ACTFL Level	Speaking Task
4.87	0.47	A	Give Instructions
4.82	0.59	I	Introduce Yourself
4.67	0.70	I	Describe Typical Routines
4.66	0.72	I	Give Directions
4.64	0.63	A	Describe a Sequence of Events in the Past
4.62	0.73	I	Explain a Familiar, Simple Process
4.52	0.83	I	Describe a Place
4.48	0.83	A	Express Personal Apologies
4.47	0.86	I	Describe Your Daily Routine
4.47	0.83	A	Describe Expected Future Events
4.41	0.79	A	Give Advice
4.35	0.90	A	Change Someone's Behavior Through Persuasion
4.31	0.97	A	Compare and Contrast Two Objects or Places
4.28	0.95	A	State Advantages and Disadvantages
4.25	0.97	A	Give a Brief, Organized, Factual Summary
4.25	1.03	I	Talk About Family Members
4.19	0.95	S	Propose & Defend a Course of Action with Persuasion
4.16	1.04	S	Support Opinions
4.16	1.02	I	Make Arrangements for Future Activities
4.14	0.96	I	Give a Brief Personal History
4.09	1.00	I	Talk About Personal Activities
4.07	1.07	I	Describe Health Problems

4.05	1.02	A	Hypothesize About Probable Outcomes
4.02	1.10	S	State Personal Point of View (Controversial Subject)
4.00	1.16	I	Order a Meal
3.99	0.98	A	Hypothesize About a Personal Situation
3.99	1.02	A	Hypothesize About an Impersonal Topic
3.96	1.08	I	Make Purchases
3.92	1.07	A	Lodge a Complaint
3.91	1.10	A	Correct an Unexpected Situation
3.89	1.15	A	Describe Habitual Actions in the Past
3.86	1.11	S	Evaluate Issues Surrounding a Conflict
3.85	1.16	S	Discuss a Professional Topic
3.83	1.19	A	Talk About Your Future Plans
3.79	1.23	S	Explain a Complex Process in Detail
3.72	1.25	S	Give a Professional Talk
3.69	1.22	S	Explain a Complex Process of a Personal Nature
3.62	1.29	S	Describe a Complex Object in Detail

The results of the survey of certified bilingual education teachers indicated that TOPT items could be based on any of the speaking tasks in the survey. Table 2 reveals that the bilingual education teachers validated all of the speaking tasks (i.e., the mean rating was above 3.50).

## Development of the Trial Form of the TOPT<sup>3</sup>

### Introducing the Project to the Test Advisory Committee

In order to ensure the quality of the test, the TEA was asked to select members of an ethnically diverse Test Advisory Committee (TAC) based in Texas. The membership of the committee was intended to reflect interests of both teacher trainers and classroom teachers as well as the different geographic areas of the state. The members of the TAC were:

- Dr. George M. Blanco, University of Texas, Austin
- Ms. Mary Diehl, Round Rock ISD
- Dr. Ellen de Kanter, University of St. Thomas
- Dr. George Gonzalez, University of Texas, Pan American
- Dr. Barbara Gonzalez Piño, University of Texas, San Antonio
- Ms. Claudina Hernandez, Alice ISD
- Ms. Carmen Muñoz, Pharr-San Juan-Alamo ISD

- Ms. Luz Elena Nieto, El Paso ISD
- Ms. Annette Ortega, Amarillo ISD
- Ms. Maggie Stovall, Alamo Heights ISD
- Dr. Marion R. Webb, Houston Baptist University

The TOPT-Spanish Test Advisory Committee (TAC) met in April 1990 in Austin, Texas and was introduced to the project and speaking tasks to be included in the job-relatedness survey discussed earlier.

## **Development of Draft Test Items**

An item-writing team composed of CAL staff experienced in writing items for SOPs (known as the Local Test Development Team or LTDT) worked to develop items for the four initial forms of the TOPT. The LTDT focused on the characteristics of the examinees who would be taking the TOPT in order to construct items appropriate and accessible to them. The LTDT assumed the typical TOPT examinee would: (1) have an interest in teaching; (2) be familiar with school and college life; (3) have some interest in language; and (4) have some familiarity with the state of Texas. These assumptions are reflected in TOPT items that are of either a personal nature or require some factual knowledge. The LTDT made every effort to avoid items that were too personal (and thus might be uncomfortable for some examinees) or too specific (and thus be unknown to some examinees).

The LTDT worked intensively between April and June 1990 to develop the items for the four forms. After each item was written, it was reviewed, revised and rewritten until the Project Director (Charles W. Stansfield) and Project Coordinator (Dorothy M. Kenyon<sup>4</sup>) were satisfied with its quality. Once all items were completed, they were carefully selected for placement into forms that would be parallel in terms of speaking tasks covered, number of education/noneducation related items, difficulty of item prompts, and variety of topics covered. Special care was taken that a variety of topic areas were covered on each form and that no form contained more than one item in any topic area (e.g., computers). Since the TOPT is an assessment of general speaking ability, the context of the items on the TOPT could not be restricted to only school-related settings and language usage. However, in light of the population of bilingual educators who would be taking the test and for whom Spanish language usage in the context of the classroom is primary, an effort was made to ensure that approximately 50 percent of the items on the Spanish TOPT were directly school or education related.

At the end of May, multiple forms of the TOPT were assembled in Spanish and French. The TEA reviewed the assembled TOPT forms and their suggested revisions were incorporated into the tests.

## **Review of Test Forms by the Committees**

The TEA nominated a Bias Review Committee (BRC) composed of two teachers who were unfamiliar with the test. The BRC reviewed the TOPT forms for a variety of potentially bias-related problems in the items. Their comments and suggestions for revisions were brought to the attention of the TAC at a meeting held during the following two days. At that meeting, TAC members were presented with the results of the job-relatedness survey and collectively reviewed the forms item by item, commenting on each item's appropriateness, accessibility to all candidates, clarity, and potential for eliciting responses displaying use of and ability in the targeted speaking tasks. Parallel items across the forms were reviewed together so that TAC members could comment on their comparability in terms of their wording and level of difficulty.

## **Preparation of the TOPT for Trialing**

TOPT trial forms were revised according to all revisions accepted during these meetings. Once the TEA approved the trial forms of the TOPT, the tapescript (containing the test directions, items, and native language prompts) was recorded at a professional recording studio. Test booklets were prepared together with the forms that would be used to collect data during the trialing. These forms are described in a later section.

## **Trialing the TOPT**

### **The Purpose of Trialing**

For a performance test such as the TOPT, a careful examination of its ability to elicit a ratable speech sample is necessary. Thus, trialing was used to study the TOPT's ability to elicit ratable speech. Trialing may be described as an intensive qualitative approach to test development (as opposed to an extensive quantitative approach based on the piloting of the test and calculation of item statistics). Trialing produces feedback from examinees, observers, and raters which allows the study of important characteristics of a performance-based test, such as the ability of each item to allow examinees to demonstrate their skill, the adequacy of the time allotted for the performances (in the case of the TOPT, the length of the pauses between items), the clarity of the instructions for each item, the perceived appropriateness and fairness of each item, the interpretability of drawings or pictures used, and the usefulness of the performance (the speech elicited) in determining a rating. Feedback from examinees, observers, and raters further helps ensure that the forms are comparable in difficulty.

### **Recruiting Examinees for the Trialing**

The TEA and CAL adopted various methods to recruit examinees. With the input of the members of the two TACs, trialing sites throughout Texas were chosen:

- El Paso: University of Texas at El Paso
- Austin: University of Texas at Austin
- Arlington: University of Texas at Arlington
- Hurst: Tarrant County Community College
- Edinburg: Pan American University
- San Antonio: University of Texas at San Antonio
- Houston: University of St. Thomas

Potential examinees were identified by TAC members. The examinees were typically groups of students enrolled in advanced undergraduate courses for bilingual education teachers at the colleges located at these sites. The professors in these courses announced the trialing and encouraged students to participate. They then compiled a list of those students who were willing to participate in the trialing. These individuals were sent information about the trialing and a return postcard on which they could indicate their willingness to participate.

### **Participation in the Trialing**

Table 3 presents background information on the 119 examinees who took the trailing version of the Spanish TOPT.

**Table 3**  
***TOPT Spanish Examinees: Descriptive Data***

**A. TOPT Form**

<b>FORM</b>	<b>Frequency</b>	<b>Percent</b>	<b>Cumulative Frequency</b>	<b>Cumulative Percent</b>
<b>Form A</b>	26	21.8	26	21.8
<b>Form B</b>	28	23.5	54	45.4
<b>Form C</b>	38	31.9	92	77.3
<b>Form D</b>	27	22.7	119	100.0

**B. Trialing Site**

<b>CITY</b>	<b>Frequency</b>	<b>Percent</b>	<b>Cumulative Frequency</b>	<b>Cumulative Percent</b>
<b>El Paso</b>	11	9.2	11	9.2
<b>Austin</b>	25	21.0	36	30.3
<b>Hurst</b>	9	7.6	45	37.8
<b>Edinburg</b>	53	44.5	98	82.4
<b>San Antonio</b>	15	12.6	113	95.0
<b>Houston</b>	6	5.0	119	100.0

**C. Current Status in Respect to Teaching**

<b>STATUS</b>	<b>Frequency</b>	<b>Percent</b>	<b>Cumulative Frequency</b>	<b>Cumulative Percent</b>
<b>Preservice</b>	72	63.2	72	63.2
<b>In-Service</b>	19	16.7	91	79.8
<b>Other</b>	23	20.2	114	100.0

**D. Ethnicity**

<b>ETHNIC GROUP</b>	<b>Frequency</b>	<b>Percent</b>	<b>Cumulative Frequency</b>	<b>Cumulative Percent</b>
<b>Black</b>	1	0.8	1	0.8
<b>Hispanic</b>	95	79.8	96	80.7
<b>White</b>	23	19.3	119	100.0

**E. Sex**

<b>SEX</b>	<b>Frequency</b>	<b>Percent</b>	<b>Cumulative Frequency</b>	<b>Cumulative Percent</b>
<b>Female</b>	88	73.9	88	73.9
<b>Male</b>	31	26.1	119	100.0

**F. Self Rating on the ACTFL Scale**

<b>SELF RATE</b>	<b>Frequency</b>	<b>Percent</b>	<b>Cumulative Frequency</b>	<b>Cumulative Percent</b>
<b>High-Sup</b>	13	11.5	13	11.5
<b>Sup</b>	20	17.7	33	29.2
<b>Adv. +</b>	40	35.4	73	64.6
<b>Adv.</b>	18	15.9	91	80.5
<b>Int.-H</b>	15	13.3	106	93.8
<b>Int.-L/M</b>	6	5.3	112	99.1
<b>Novice</b>	1	0.9	113	100.0

*(Frequency Missing = 6)*

Table 3 reveals that although an unequal number of examinees took each form of the Spanish TOPT, no form was taken by fewer than 26 examinees. Of the total number of examinees, 44.5 percent took the TOPT at Edinburg, due to the interest of local professors. Over half (63.2 percent) were preservice teachers, while only 16.7 percent were in-service. Self-ratings reveal an uneven distribution of ability with over half (51.3 percent) at one level (Advanced), with 0.9 percent at the Novice level, 18.6 percent at the Intermediate level, and 29.2 percent at the Superior level. 80.5 percent rated themselves above the passing score of Advanced, with only 19.5 percent under the passing score. Hispanics made up approximately 79.8 percent of the total sample, and males 26.1 percent.

**Data Collection During Trialing**

During the first part of the testing session, which lasted approximately 50 minutes, examinees took the TOPT. Feedback on the test was subsequently collected from two sources. The first source was the examinee. After the test was administered, each examinee completed a two-part evaluation form eliciting quantitative ratings and qualitative written comments on both specific TOPT items and on the test in general.

The first part of the form collected quantitative data. On a machine-readable answer sheet (Appendix B), examinees gave background and demographic information and assigned themselves a rating on a simplified ACTFL scale. These self-ratings served as a "ballpark" estimate of the examinee's TOPT score and were used later to determine which Spanish tapes would be listened to in their entirety by raters. The results of the background, demographic, and self-rating responses were presented earlier in Table 3. The machine-readable response sheet also included two statements for every item: the first dealt with the adequacy of the time allowed for the item, the second with its overall quality. In addition, for the five picture items, one statement per item dealt with the perceived clarity of the picture. To each of these 40 statements, examinees had to indicate, on a scale of one to five (with five being highest), their degree of agreement.

The second part of the examinee response form collected qualitative data. For any statement to which the examinee gave a lower rating on the first part of the form, this part requested the examinee to explain why via written comments. Examinees were requested to address any concerns about the test that had not been addressed elsewhere.

The second main source of trialing feedback came from six judges, individuals familiar with the ACTFL scale who listened to examinee tapes focusing on the quality of the speech elicited by each item on the test to make a rating for that examinee. The judges recorded their comments on a form where they marked on a scale of one to three the usefulness of the speech sample elicited by each item in determining that examinee's proficiency level. For each item and each examinee, they also indicated the appropriateness of the time allowed for the response. Judges also noted any potential problems with specific items based on the examinee's performance. Each tape was listened to by one judge. A total of 80 (20 for each form) of the 119 Spanish tapes were listened to. All tapes made by non-Hispanics and all tapes made by examinees who gave themselves a rating lower than Advanced were included in the 80 tapes.

## **Results of the Trialing**

The feedback collected during the trialing guided the post-trialing test revision process. Quantitative data was reviewed and written comments for all items were read before determining whether any revision should be made.

### **Results of the Examinee Data Form, Part 1**

A data file was created from the machine-readable examinee response forms. Each test form was analyzed separately. Since five was the highest rating, any statement with an average score of 3.50 or below identified a problematic TOPT item. Those with a score above 3.50 but below 3.75 identified TOPT items marked for careful analysis to determine the presence of any potential problems. Statements with a mean rating above 3.75 were considered to identify TOPT items not needing serious revision. Because each form was taken by a different group of examinees, this data was not used to make general inferences about the quality of each trialing test form.

## Results of the Examinee Data Form, Part 2

All written comments were coded for test form, item referred to, and the degree of negativity expressed in them, and were entered into a database. The written comments served primarily to inform the specific revisions that needed to be made. In analyzing the printout of comments, we considered the classifications into which comments generally fell. Although examinees were requested to comment only if there were a problem with the item, a substantial number of positive comments were received. Concurring comments pointed out some flaw requiring revision, usually corroborating lower mean ratings of statements. Other, more unique positive comments suggested helpful revisions, even when not corroborated by the comments of others or by the ratings. Some negative comments indicated that the examinee's problem with the item was not due to any attribute of the item per se. The most common suggestion was to increase the amount of time allowed to prepare and give the response. In general, the examinees' written comments were very helpful in making revisions.

## Results of the Judge's Response Sheet (Quantitative)

The judge's response sheet contained quantitative data on the quality of the speech sample elicited by the item and on any perceived time problems with the item.

Table 4 contains the average quality ratings by TOPT form as rated by the judges who scored the trialing tapes.

**Table 4**  
**Average Item Quality Rating for Each Form**

<b>Form</b>	<b>Average Rating</b>	<b>Range</b>
A	2.73	2.57 - 2.80
B	2.59	2.35 - 2.78
C	2.57	2.38 - 2.67
D	2.71	2.55 - 2.85

Table 4 indicates that across items, the mean item quality was perceived by raters as being quite high, since one indicated poor quality, two meant average, and three indicated excellent. Judges generally perceived none of the items to be particularly problematic.

## Revisions of the Trial Forms

After collecting and analyzing all data from the trialing, results were reviewed item by item and revisions were made as appropriate. A presentation of the trialing results and the revisions that were made was given during final meetings of the BRCs and TACs in August 1990. The BRCs reviewed the revised items for bias

and made comments and suggestions for revisions. The TACs then discussed all revised items and acted on comments made by the BRCs. Final wording and revisions were the basis of group decisions made at these meetings, the outcome of which became the final version of each TOPT form.

## **Content Validation**

At the foundation of the TOPT's content lie the speaking tasks, based on the ACTFL guidelines, which were validated during the job-relatedness survey. Each item on the TOPT was written to elicit language in response to one of these tasks. In order to investigate whether the items on the final forms of the TOPT did match the speaking tasks which served as the specifications for the items, a content validation study was undertaken. Data was collected at the Fall 1990 Texas Association of Bilingual Educators meeting, on Thursday, November 1, in Lubbock, Texas. The following individuals served as judges in the content validation study:

- Ms. Rosa M. Chahin, Houston ISD
- Ms. Virginia Moore, Midland ISD
- Ms. Elizabeth Martin, Grand Prairie ISD
- Dr. Juan Lira, Laredo State University

The task of the judges was to examine each item on the three operational forms of the TOPT and determine whether it elicited the speaking task (e.g., support an opinion, give directions, state advantages and disadvantages, etc.) specified for it.

All TOPT items were validated by the content validation studies as matching the targeted speaking task for the item. In only two cases did a judge give a negative mark, and in each of those cases the judge's comment revealed that mark was unrelated to the match between the task and the item (i.e., it was assigned because the judge had some other concern about the item).

## **Standard Setting Study**

In order to provide additional data to assist the TEA and the Texas State Board of Education in setting passing scores for the TOPT, a standard setting study, following the model described in Livingston (1978) and adapted by Powers and Stansfield (1982), was carried out in the fall of 1990, concurrently with the content validation studies described above. These studies required a sampling of examinee performances and a panel of judges to rate the performances as acceptable or unacceptable.

## **Preparation of the Standard Setting Master Tape**

Before the study could be conducted, it was necessary to prepare a master tape containing TOPT performances of examinees at different levels of speaking proficiency. This was accomplished through the following process.

## **Initial Selection of Representatives of Various Levels of Speaking Proficiency**

The judges who listened to the examinee response tapes following the trialing gave a preliminary rating to each examinee they listened to. Additionally, each examinee provided a self-rating during trialing. These

ratings were used to identify prospective tapes for inclusion on the master tape. It was hoped that a master tape could be constructed that would contain 21 examinees, three for each of the seven ACTFL levels between Intermediate Low and High Superior. Because the preliminary ratings provided by the examinees and the trialing judges suggested there were virtually no Intermediate Low level examinees, the number of tapes selected for the first step for Spanish was 31 instead of the projected 21.

## **Ratings by Texas ACTFL-Certified Raters**

The TEA and CAL jointly chose two prominent Texas ACTFL-certified raters to score the selected trialing tapes within a period of three weeks. Dr. George Blanco of the University of Texas at Austin and Dr. Vickie Contreras of the University of Texas, Pan American, served as raters.

The order of the 31 tapes to be rated was prescribed. The raters gave each examinee a rating on each item and the opening conversation as a whole, and also rated each one holistically. The raters had few holistic and item level score disagreements.

## **Construction of the First Master Tape**

In constructing the master tape for the standard setting study, the goal was for each individual examinee to be represented by three performances indicative of a certain ACTFL level. The examinees and segments were chosen as follows. Data from the ratings of the Texas ACTFL raters were entered into a database. Then, items were identified where the two raters agreed on the item-level rating. Once these item-level matches were identified for any individual examinee, the overall holistic rating of that individual was consulted. Finally, three items whose responses coincided with the overall holistic rating for the examinee were chosen. By applying this procedure, we constructed a tentative master tape containing 29 examinees.

## **Confirming the Master Tape Ratings**

In order to confirm that each examinee's global rating based on the three performances was correct, we conducted a confirmatory study of the tentative ratings. To do this, CAL obtained from ACTFL the names of outstanding ACTFL raters and trainers for Spanish from across the country. Five were contacted and agreed to participate in the study. Each was sent a copy of the master tape, a rating sheet, and instructions. These raters were told to work independently, listening to each person and estimating his or her ACTFL rating based on the three performances contained on the tape.

## **Assigning "True Scores" to the Master Tape**

Once the raters' responses were received by CAL, their scoring data was entered into a computer. Their ratings for each examinee were examined together with the examinee's tentative rating (i.e., the agreed upon level(s) of the two Texas raters). When possible, a "true score" was assigned to each examinee. This "true score" was the level the examinee's responses on the first master tape was intended to typify. Examinees for whom ratings were discrepant were eliminated from the pool of potential examinees to be included in the final master tape. Eliminating examinees for whom true scores could not be assigned reduced the number of examinees on the master tape from 29 to 22. Ultimately, there was not an equal number of examples from each ACTFL level on the final master tape for Spanish. Below are the numbers of examples for each level appearing on the master tapes:

<b>ACTFL Level</b>	<b>French</b>	<b>Spanish</b>
Intermediate Low	0	0
Intermediate Mid	6	3
Intermediate High	3	5
Advanced	4	5
Advanced Plus	3	2
Superior	3	6
High Superior	0	1

## Setting the Passing Standards

The TEA submitted to CAL a list of teachers and teacher trainers from throughout Texas whom they deemed qualified to serve as judges on the standard setting committees. These individuals were sent invitations to the standard setting sessions.

The standard setting session was held in conjunction with the annual Texas Association of Bilingual Educators fall conference. Below are the names and affiliations of the committee members:

- Ms. Carmen A. Dominguez, Houston ISD
- Ms. Lucia E. Elizarde, Harlingen ISD
- Ms. Yolanda Espinoza, San Marcos ISD
- Dr. Maria Loida Galvez, Pasadena ISD
- Ms. Susana Gomez, Lubbock ISD
- Ms. Joyce Hancock, Lufkin ISD
- Dr. Roy Howard, Texas Tech University
- Mr. Manuel A. Martinez, Austin ISD
- Ms. Isabell McLeod, Amarillo ISD
- Ms. Elba-Maria Stell, El Paso ISD
- Ms. Juanita Villegas, Lubbock ISD
- Dr. Judith Walker de Felix, University of Houston
- Ms. Elsa Meza Zaragosa, Corpus Christi ISD

The task given to the members of the standard setting committee was to listen to each individual on the master tape and indicate whether or not they felt that person demonstrated enough speaking ability to be certified to teach in Texas.

## Results of the Standard Setting Study

Table 5 presents the make-up of the 13 members of the Bilingual Education Standard Setting Committee.

**Table 5*****Descriptive Statistics on the Bilingual Education Standard Setting Committee Members*****A. Position**

<b>Position</b>	<b>Frequency</b>	<b>Percent</b>	<b>Cumulative Frequency</b>	<b>Cumulative Percent</b>
<b>Classroom Teacher</b>	10	76.9	10	76.9
<b>District Supervisor</b>	1	7.7	11	84.6
<b>Teacher Trainer</b>	2	15.4	13	100.0

**B. Sex**

<b>Sex</b>	<b>Frequency</b>	<b>Percent</b>	<b>Cumulative Frequency</b>	<b>Cumulative Percent</b>
<b>Male</b>	2	15.4	2	15.4
<b>Female</b>	11	84.6	13	100.0

**C. Ethnicity**

<b>Ethnic</b>	<b>Frequency</b>	<b>Percent</b>	<b>Cumulative Frequency</b>	<b>Cumulative Percent</b>
<b>Hispanic</b>	10	76.9	10	76.9
<b>White-Non Hispanic</b>	3	23.1	13	100.0

**D. Region of Texas by First Two Digits of Zip Code**

<b>REGION</b>	<b>Frequency</b>	<b>Percent</b>	<b>Cumulative Frequency</b>	<b>Cumulative Percent</b>
<b>75--Northeast</b>	1	7.7	1	7.7
<b>77--East</b>	3	23.1	4	30.8
<b>78--South Central</b>	4	30.8	8	61.5
<b>79--West</b>	5	38.5	13	100.0

Table 6 shows the percent of bilingual education committee members rating examinees as acceptable at each ACTFL level.

**Table 6**  
***Mean Percent of Judges Rating Examinees Acceptable at Each Level (Bilingual Education)***

<b>ACTFL Level</b>	<b>Number of Examinees</b>	<b>Mean Percentage</b>
Intermediate Mid	3	12.8
Intermediate High	5	21.5
Advanced	5	83.1
Advanced Plus	2	88.5
Superior	6	96.2
High Superior	1	100.0

Table 6 shows a clear dividing line between Intermediate High and Advanced. For the bilingual education committee members, an Intermediate High level performance was clearly not adequate, while the Advanced level performance was deemed adequate on average by over 80 percent of the committee members.

This data was submitted to the Texas State Board of Education, which established a passing score of Advanced on the TOPT for bilingual education teacher certification in February 1991.

## **Description of the TOPT Operational Program**

After setting the passing score, the State Board of Education asked National Evaluation Systems (NES), which administers all teacher certification tests in Texas, to administer the TOPT beginning in the fall of 1991. The first three TOPT administrations were held on November 2, 1991, February 29, 1992, and June 27, 1992 (National Evaluation Systems, 1991).

NES contracted CAL to train raters to score the examinee response tapes resulting from these administrations. Raters are trained to score the TOPT in one and a half day workshops. Following training, raters are given a calibration tape consisting of 10 examinees. After listening to performance on three tasks, the rater must assign a global rating to each examinee. The global ratings assigned to the 10 examinees by each rater are then examined to determine if the rater is ready to begin scoring operational tapes. Raters who misclassify no more than two examinees may begin operational scoring. Those who do not pass this test of their rating ability, undergo further training. Additional recalibrations are taken by raters twice a day each successive day of scoring. Those who do not pass these recalibrations are given further training. At the most recent training session, held in July 1992, 55 raters were trained. Of these, only three did not pass the initial test of their rating ability following training. This demonstrates that raters can be trained to score the TOPT

consistently in a relatively short period of time.

Six hundred and nine candidates for teacher certification completed the TOPT at the most recent administration (July 1992). The TOPT is offered three times per year and it is anticipated that annual program volume will soon be about 2,000.

The TOPT has been well received by examinees, by IHEs involved in teacher preparation, and by those who become TOPT raters. Initial results suggest that the program provides a valid and reliable approach to large scale oral proficiency assessment for bilingual education teachers. We believe the TOPT solves the problems that plagued the LPI which it replaced. These problems were (a) inadequate interview; and (b) inconsistent ratings. The TOPT provides every examinee with the same high quality of test. Ratings appear to be both consistent and accurate. Further research on the validity of the TOPT should be conducted and the reliability of the program should be monitored with care. At this point, however, it appears that the TOPT is an appropriate test for the purpose and the population for which it was intended. Its implementation helps ensure that only those bilingual education teacher certification applicants who are competent in spoken Spanish will receive a credential. Thus, bilingual education teachers will be competent to instruct using the child's native language.

## Endnotes

<sup>1</sup> This article is based on a monograph describing the development of the TOPT (Stansfield and Kenyon, 1991). For a more complete description of the development of the TOPT, the reader may refer to the monograph.

<sup>2</sup> For a complete description of the TOPT, the ACTFL scale on which it is scored, a complete practice test, and recorded examples of examinee performance at five different proficiency levels, it is recommended that the reader obtain a copy of the TOPT Test Preparation Kit, which is available for \$30 from the Center for Applied Linguistics.

<sup>3</sup> I wish to express my appreciation to Nolan Wood, Director of the Division of Teacher Assessment at the Texas Education Agency. He and his staff provided advice and helpful comments throughout the project.

<sup>4</sup> It is appropriate to acknowledge the many contributions of Dorry Mann Kenyon to this project. Mr. Kenyon also provided helpful comments on an earlier version of this manuscript.

## References

ACTFL. (1986). *ACTFL Proficiency Guidelines*. Yonkers, NY American Council on the Teaching of Foreign Languages.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and Department of Justice. (1978, August). Uniform guidelines on employee selection procedures. *Federal Register*, 43(166), 38290-38315.

Livingston, S. A. (1978). Setting standards of speaking proficiency. In J. L. D. Clark, (Ed.), *Direct testing of speaking proficiency: Theory and application*. Princeton, NJ: Educational Testing Service.

NTE Programs. (1989). *Validity using the NTE tests*. Princeton, NJ: Educational Testing Service.

National Evaluation Systems. (1991). *Texas Oral Proficiency Test Registration Bulletin*. Amherst, MA: National Evaluation Systems.

Powers, D. and C. W. Stansfield. (1985). Testing the oral English proficiency of foreign nursing graduates. *English for Specific Purposes Journal*, 4,1, 21-36.

Rudner, L. M. (1987). Content and difficulty of a teacher certification exam. In L. M. Rudner, editor, *What's happening in teacher testing: an analysis of state teacher testing practices* (33-38). Washington, D.C.: US Department of Education, Office of Educational Research and Improvement.

Stansfield, C. W. (1980). Testing the language proficiency of bilingual teachers. In *Problems and issues in bilingual education*, Selected papers from the 1979 Denver Conference "The Lau Task Force Remedies: Comprehensive Bilingual Education Program Planning." (pp. 191-203). Denver: Coalition of Indian Controlled School Boards.

Stansfield, C. W. (1989). *Simulated oral proficiency interviews*. ERIC Digest. Washington, D.C.: Center for Applied Linguistics.

Stansfield, C. W. & Kenyon, D. M. (1991). *Development of the Texas Oral Proficiency Test*. Final report to the Texas Education Agency. Washington, DC: Center for Applied Linguistics. Springfield, VA: ERIC Document Reproduction Service. ED 332 552.

## **Appendices A & B**

Appendices A & B were not included in the electronic version of this publication.