DOCUMENT RESUME

ED 334 226 TM 016 716

AUTHOR Kenyon, Dorry Mann; Stansfield, Charles W.
TITLE A Method for Improving Tasks on Performance Assessments through Field Testing.
PUB DATE Apr 91
NOTE 44p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, April 4-6, 1991).
AVAILABLE FROM Center for Applied Linguistics, 3520 Prospect St., NW, Washington, DC 20007 (price varies).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150) -- Tests/Evaluation Instruments (160)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Bilingual Teachers; *College Students; *Data Collection; *Educational Assessment; *Field Tests; French; Higher Education; Language Tests; Licensing Examinations (Professions); *Performance; Qualitative Research; Second Language Learning; Spanish; Standardized Tests; Teacher Certification; *Test Construction
IDENTIFIERS Texas Oral Proficiency Test

ABSTRACT

One aspect of the development of a performance assessment is addressed--the field testing of the tasks. In a performance-based assessment, the tasks must allow each examinee a fair and equal opportunity to give the best possible demonstration of his or her ability, and must elicit a performance sample that enables scorers to evaluate each examinee adequately. A method of field testing is proposed that collects information via two supplemental instruments administered during the field test: (1) one completed by examinees; and (2) one completed by the field test raters. Each instrument produces quantitative data via machine-readable response forms and qualitative data via written comments. The method is illustrated by field testing four parallel forms in French and Spanish of the Texas Oral Proficiency Test, a performance-based assessment for certification of Spanish, French, and bilingual education teachers in Texas. The field test involved 100 examinees and eight raters. The use of the two instruments during field testing allows both examinees and raters to identify poorly functioning items on a performance-based assessment. This method, which encourages examinees and raters to provide focused written feedback, enables both to be involved in test development and improvement. Three tables present data, and an appendix provides the supplemental data collection forms. (SLD)

A Method for Improving Tasks on Performance Assessments through Field Testing

Dorry Mann Kenyon

and

Charles W. Stansfield

Center for Applied Linguistics

1118 22nd Street, NW

Washington, DC 20037

(202) 429-9292

Paper presented at

the Annual Meeting of the

National Council on Measurement in Education

Chicago, IL

April 4, 1991

Introduction

For more than a decade, interest in and support for the use of performance tasks in standardized testing has grown steadily among the public and even in the measurement community. Direct writing assessments are now common in most states and many occupational licensure and certification programs are now moving toward incorporating either work samples or simulations of real life professional tasks into their testing programs. In spite of the interest in performance assessment, there is still a paucity of published information on the development of quality simulation tasks for performance assessment.

The development of a performance assessment involves a number of stages. These include determining the test specifications, writing the item specifications, training item writers, reviewing and revising items and preliminary drafts of the test, selecting the criteria for scoring and developing scoring scales, and selecting and training raters. The purpose of this paper is to focus on one aspect of the development of a performance assessment, the field testing of the tasks.

Problem

On a performance-based assessment, particularly one that will have important repercussions for the examinee and society, the tasks (also referred to as exercises, prompts or items) used to elicit examinee performance must have two characteristics that are critical to validity and accuracy of measurement. First, in

order to ensure that the test is fair and unbiased to the fullest extent possible, the items must allow each and every examinee a fair and equal opportunity to give the best possible demonstration of his or her ability. Second, in the interest of obtaining the most accurate measurement of ability possible, these tasks must elicit a performance sample that enables scorers to adequately evaluate the examinee. Test developers are faced with the problem of ensuring that the items or tasks on their performance-based measurements exhibit these characteristics. Although the specification of clear item writing guidelines, the thorough training of item writers, and the review of items by experts help in the development of appropriate items, those steps alone cannot guarantee that the items exhibit these two critical characteristics. Just as in the development of objective paper and pencil measures, field testing is required for performance assessments as well. To our knowledge, no one systematic technique has been adopted by test developers to ensure that these goals are being reached through field testing. In fact, Stiggins (1987), in his excellent step-by-step approach to the design and development of performance assessments, does not mention anything about field testing the exercises used in the assessment. In this paper, we propose a method of obtaining quantitative and qualitative data during the field testing of a performance assessment that can either aid in or confirm the development of high quality performance tasks. Additionally, this method provides for the involvement of both examinees and

raters, those who play the most direct role in the assessment, in the test development process.

To illustrate the method, this paper uses examples from our experience in developing four parallel forms in both French and Spanish of the Texas Oral Proficiency Test (TOPT), a performance-based assessment of proficiency in spoken language used for the certification of Spanish, French, and Bilingual Education teachers in Texas. The TOPT is one of seven tape-mediated, simulated oral proficiency interview (SOPI) tests (Stansfield, 1989) that we have developed in different languages under federal grants and state contracts during the past several years. It elicits examinee speech performance on 24 items via a master tape and a test booklet. The first 9 items are relatively easy, short-answer items linked to a single speaking context. They are intended as a warm-up for the examinee. The remaining 15 items function as prompts to elicit multiple samples of examinee speech. Each prompt focuses on a different language function or task. Examinee responses to all items are recorded on a second tape (the examinee response tape). That tape is scored at a later session by trained raters using a holistic scale for rating speaking ability in a second language. The scale was developed by the American Council for the Teaching of Foreign Languages (ACTFL) and is called the ACTFL Oral Proficiency Guidelines. The TOPT field testing used to illustrate our method involved 160 examinees and eight raters.

Method

Once items or tasks for the performance-based assessment have been developed, reviewed, revised, and the test form(s) have been assembled, then the preliminary version(s) of the performance-based test is field tested and scored under conditions reflective of the operational program. To obtain additional information, two supplemental data collection instruments are developed and administered during the field testing: one is completed by the field test examinees and the other by the field test raters. Each instrument consists of two parts; one gathers quantitative data and can be formatted on a machine-readable response form, while the other collects qualitative data, providing spaces for written comments.

1. Examinee Feedback Forms

The two-part instrument for the field test examinees is used to detect anything in the test (as a whole and for each individual item) that may be hindering examinees from giving their best possible performance. The quantitative part is similar to an attitude questionnaire. It presents examinees with a series of statements about issues salient to the test and each of its items. After taking the performance-based test, examinees respond to these statements, indicating their degree of agreement or disagreement to each on a five point scale, with 5 reflecting the most positive response to the statement.

An example from the TOPT will illustrate the quantitative

section of this feedback instrument. For the TOPT, the two most salient issues for every item were whether examinees felt that the item allowed them to produce speech indicative of their current level of ability, and if they felt that the time allowed for their preparation and their response was adequate. These concerns were presented to the test-takers in statements such as the following: "As a whole, I felt picture item #1 allowed me to give a response that reflects my current ability to speak the language," and 'In picture item #1, the time allowed for preparing my answer and making my response was appropriate."

Five of the TOPT items involved the use of pictures. Statements about the clarity and interpretabilty of these pictures were developed for this instrument as well. For example, for picture item #1, which involved the use of a map of a section of a town, based on a bird's-eye view, the examinee responded to the statement "The map for picture item #1 was clear and understandable."

Finally, three general statements about the TOPT as a whole were included. These statements addressed the issues of examinee nervousness, the use of a target language question or statement as a signal to begin one's answer (as opposed to a "beep" signal), and the degree to which the examinee is satisfied that a rater listening to the examinee response tape would get an accurate picture of his or her ability to speak French or Spanish.

The entire machine-readable questionnaire contained 40

statements. Appendix A contains a copy of the machine-readable questionnaire and the instructions for completing it.

Calculating mean examinee responses to these statements provides quantitative data to inform the item and test revision process. For example, for Form D of the Spanish TOPT, the 27 field test examinees awarded their highest rating (mean = 4.22) to the statement about the clarity of the picture used for picture item #4, while the lowest rating (mean = 3.19) was awarded for the time allotted to answer picture item #3. In developing the TOPT, all statements receiving a mean rating of 3.75 or less identified for the test developer potential concerns requiring revisions or refinement. Table 1 gives as an example all the mean examinee responses for Form D of the Spanish TOPT.

Insert Table 1 here

The second part of this instrument provides examinees with a form containing spaces for written comments. The form asks examinees to explain why they awarded any item a lower rating (3, 2, or 1) and to suggest revisions that would, in their opinion, improve the item. Examinees are also encouraged to comment on any general statements about the test and to write about any additional concerns not addressed elsewhere. Appendix A contains a copy of this form as used in the trialing of the TOPT.

In our experience with the TOPT, these written comments shed light on lower rated items and contained many useful suggestions

for revising them. For analysis purposes, these comments were typed into a word-processing database and coded with examinee and test form and item information. In addition, the person typing in the comments assigned a rating to each as to the degree of positiveness or negativeness reflected in the comment. These comments, ordered from most positive to most negative, were then output for use in the revision process. An example of one page of output for Picture #4 from Form B of the TOPT-Spanish is found in Appendix B.

2. Rater Feedback Forms

The second instrument is for the field test raters. They use it to indicate whether each performance actually elicited by the test's tasks or items provides adequate and helpful information to evaluate the examinee using the designated scale or criteria. To do this, each field test rater, for each item and for each examinee, is asked in the first part of the instrument to rate the adequacy of the performance for assigning the examinee a score. If appropriate, the rater's instrument can also ask raters to provide quantitative information about other salient aspects of the test format. Qualitative data is collected by encouraging raters to make written comments on the second part of this instrument on problems in the performances of examinees. For example, they can mention if anything in the item prompt seems to hinder the examinee from giving an appropriate performance, if the task elicits the desired type of performance,

or if any type of examinee seems to have an advantage of disadvantage in responding to the prompt.

For the TOPT, both quantitative and qualitative data were collected on one form, which appears in Appendix C. The TOPT field test raters listened to a subset (67%) of the field test tapes and evaluated the quality of each item by indicating on a scale of 1 to 3 the item's usefulness in determining the appropriate rating. A rating of 1 represented a contribution that was "not useful" to the rater in assigning the examinee an overall score, 2 represented a "useful" contribution, and 3 a "very useful" contribution. For Spanish Form D, for example, the mean of the ratings by the four field test raters ranged from a high of 2.85 for Picture #4 to a low of 2.55 for Situation #3. Mean ratings of 2.50 or below identified items to the test developers that were in need of revision. As an example, Table 2 below shows all the mean ratings on the opening conversation and the 15 subsequent items for Spanish Form D. There were no items receiving a rating below 2.50 on the TOPT-Spanish, Form D.

Insert Table 2 here

Like the amount of time allotted to an examinee on an essay test, the timed pauses on the TOPT play an important role. Because the amount of time allotted to a task on a performance test can limit the examinee's ability to complete the task, it is necessary to confirm the adequacy of timing. Thus, the raters

were asked to provide quantitative information on whether, for each item, the time limit posed a problem for the examinee. If the examinee had more time than necessary to make a complete and adequate response to an item, the rater indicated the amount of extra time, in number of seconds, that the examinee had. If the examinee needed more time to carry out the language function presented in the item, the rater estimated the number of seconds needed. A frequency distribution of seconds extra (positive numbers) or seconds needed (negative numbers) revealed whether patterns of too long or too short time limits existed in the group of examinees field tested and informed revisions to the times allotted for each speaking task. Table 3 shows an example of these tables for two items. The data for Opening Conversation question #10 (question #9 on the final form) indicates that the response time for that item should be lengthened. Only 50% of the examinees had an adequate amount of time. In contrast, the data for Picture Item #1 indicates that slightly under 50% had much more than enough time to give an adequate response for this item. About 36% had 5 or more seconds to spare.

Insert Table 3 here

The TOPT raters' written comments on the examinees' performances on the items were coded and typed into a word-processing database, and then printed. An example of raters comments from Spanish Form B is given in Appendix D. These

comments helped identify causes of problems and gave suggested solutions to them. In many cases, both the quantitative and qualitative data collected from the raters supported the data collected from examinees, clearly identifying problems with the prompts and in almost all cases suggesting remedies to those problems.

Conclusions

From our experience in developing the TOPT, we have seen how the use of these two instruments during field testing efficiently collects data that identify, from both the examinees' and raters' perspectives, poorly functioning items on a performance-based assessment. In addition, this method, by encouraging examinees and raters to provide focused written feedback, enables both to be involved in the test development process. These quantitative and qualitative data alert the test developers to specific problematic aspects of the test items or the test as a whole that need to be addressed. We found with the TOPT that in many cases the suggestions of examinees and raters could be used to improve the test. Finally, this method of collecting field test data gives the test developers confidence in the quality of those items receiving high ratings by examinees and raters. Similarly, items which receive poor ratings and are beyond repair can be deleted from the test.

The information on item functioning collected through this process can be useful to test score users as well as test

developers. Should the validity or fairness of the test be challenged, the information provides proof that examinees' and raters' reactions to the items were sought out during the test development process. The information also provides proof that problematical items were either deleted or revised according to examinees' and raters' suggestions. The process also allows the test developer to demonstrate that items not revised or deleted were perceived to be fair, appropriate and useful by pretest examinees and raters. Ultimately, when an examinee indicates that an item "allowed me to give a response that reflects my current ability," that item accrues a good deal of validity. Thus, the mean rating assigned by examinees to an iter can serve as an index of validity as perceived by the group most directly affected, the examinees.

Additional benefits may accrue from using this method. For example, four forms of the TOPT were developed simultaneously, each parallel in terms of the language functions addressed by its items. The use of the method of field testing described in this paper allowed for the comparison of the functioning of parallel items across forms. In some cases, when only one of a set of four parallel items received a lower rating, it was possible for us to identify and correct that aspect of the lower rated item that distinguished it from similar items. This increases the test developer's knowledge about item characteristics that may affect performance. Additionally, comparing mean ratings of parallel items across the four forms identified generic task or

prompt types that in general were more problematic to examinees than other task or prompt types. For example, for the 119 examinees who took the field test Spanish TOPT, the picture-based item requiring narration in the past tense received the highest mean overall rating across forms (3.90). The item requiring a summary of factual information received the lowest mean overall rating across forms (3.33). This information can be added to the test developer's storehouse of knowledge with the result that he or she will know with some degree of confidence how both examinees and raters will perceive future items of each type.

We believe this method is efficient even when the performance-based assessment consists of a single prompt, as in an essay test. A likert-type questionnaire addressing salient points can be developed and administered to field test examinees. After concluding the test, examinees can then be given time to reflect on their test taking experience and directly describe or document problems they had with understanding the prompt, organizing the essay, or with time. This is more efficient than having readers try to infer what the deficiencies of a prompt may have been on the basis of a reading of the examinee's essay.

Indeed, not all of the pretest papers need to be read in order to obtain data on the adequacy of the test or the prompt with our method. Since paying raters is often the largest cost in performance-based assessment, more data on the field test version can be collected for less cost by having examinees describe and document problems with the test. Field test raters

can rate a random subset of the essay papers, commenting on the adequacy of the prompt to elicit a ratable writing sample, without losing the input into the field testing process of those examinees whose papers were not read.

We employed this procedure (rating a random subset of the tests) with noteworthy success. While raters' comments were helpful, examinees' comments were even more helpful, and could be collected without incurring the cost of rating the tape. Thus, while we believe that both examinees' and raters' comments should be sought during the field testing of a performance based item, examinees' comments are most useful and can be obtained more efficiently. Currently, few performance-based tests systematically seek out examinees' comments during the test development phase. Our experience with the TOPT suggests that paying attention to examinee comments significantly improves the test product.

References

Stansfield, C.W. (1989). Simulated oral proficiency interviews. ERIC Digest. Washington DC: ERIC Clearinghouse on Languages and Linguistics. (ERIC Document Reproduction Service No. ED 317 036)

Stiggins, R.J. (1987). NCME instructional module on design and development of performance assessments. Educational Measurement: Issues and Practice 6, 33-42.

Table 1
Trialing Examinees' Mean Ratings
on Quantitative Questionnaire
for TOPT-Spanish Form D

N Obs	Variable	Label	N	Mean	
27	ITEM1	Opening Conversation--General	27	3.963	
	ITEM2	Opening Conversation--Time	26	4.077	
	ITEM3	Picture #1--General	27	4.000	
	ITEM4	Picture #1--Map	27	4.037	
	ITEM5	Picture #1--Time	26	4.038	
	ITEM6	Picture #2--General	26	3.423	**
	ITEM7	Picture #2--Picture	27	3.815	
	ITEM8	Picture #2--Time	27	3.185	**
	ITEM9	Picture #3--General	26	3.885	
	ITEM10	Picture #3--Pictures	27	4.185	
	ITEM11	Picture #3--Time	27	3.852	
	ITEM12	Picture #4--General	26	4.115	
	ITEM13	Picture #4--Pictures	27	4.222	
	ITEM14	Picture #4--Time	27	3.778	
	ITEM15	Picture #5--General	26	3.692	*
	ITEM16	Picture #5--Pictures	27	3.593	*
	ITEM17	Picture #5--Time	27	3.519	*
	ITEM18	Topic #1--General	25	3.720	
	ITEM19	Topic #1--Time	27	3.815	
	ITEM20	Topic #2--General	27	3.889	
	ITEM21	Topic #2--Time	27	3.778	
	ITEM22	Topic #3--General	26	3.577	*
	ITEM23	Topic #3--Time	27	3.630	*
	ITEM24	Topic #4--General	26	3.577	*
	ITEM25	Topic #4--Time	26	3.577	*
	ITEM26	Topic #5--General	27	3.926	
	ITEM27	Topic #5--Time	27	3.704	*
	ITEM28	Situation #1--General	26	4.038	
	ITEM29	Situation #1--Time	26	3.846	
	ITEM30	Situation #2--General	27	3.778	
	ITEM31	Situation #2--Time	27	3.704	*
	ITEM32	Situation #3--General	26	3.654	*
	ITEM33	Situation #3--Time	26	3.577	*
	ITEM34	Situation #4--General	27	3.815	
	ITEM35	Situation #4--Time	27	3.556	*
	ITEM36	Situation #5--General	26	3.923	
	ITEM37	Situation #5--Time	26	3.769	
	ITEM38	Unduly Nervous?	24	3.292	
	ITEM39	Replace Target Language Prompt?	27	2.704	
	ITEM40	An Accurate Picture?	26	3.154	

* between 3.51 and 3.75
** 3.50 or below

Table 2
Example of TOPT Field Test Raters'
Mean Item Quality Ratings
TOPT-Spanish, Form D

N Obs	Variable	N	Mean
21	CQUAL	19	2.68421
	P1QUAL	20	2.70000
	P2QUAL	19	2.68421
	P3QUAL	19	2.78947
	P4QUAL	20	2.85000
	P5QUAL	20	2.70000
	T1QUAL	21	2.71429
	T2QUAL	20	2.70000
	T3QUAL	21	2.61905
	T4QUAL	20	2.65000
	T5QUAL	21	2.61905
	S1QUAL	21	2.76190
	S2QUAL	21	2.71429
	S3QUAL	20	2.55000
	S4QUAL	21	2.76190
	S5QUAL	19	2.78947

Table 3
Example of Quanitative Time Data
TOPT-Spanish, Form A

(Opening Conversation #10)

C10TIME	Frequency	Percent	Cumulative Frequency	Cumulative Percent
-15	1	4.5	1	4.5
-7.5	1	4.5	2	9.1
-5	6	27.3	8	36.4
-2.5	3	13.6	11	50.0
0	11	50.0	22	100.0

Frequency Missing = 4

(Picture Item #1)

P1TIME	Frequency	Percent	Cumulative Frequency	Cumulative Percent
-7.5	1	4.5	1	4.5
0	11	50.0	12	54.5
5	2	9.1	14	63.6
7.5	3	13.6	17	77.3
10	1	4.5	18	81.8
12.5	4	18.2	22	100.0

Frequency Missing = 4

Appendix A

Instructions and Forms used in the TOPT Trialing to Collect Feedback Data from Examinees

Texas Oral Proficiency Test (TOPT)
Trialing Feedback

Please do not write in this booklet!

INSTRUCTIONS

Thank you very much for participating in this trialing of the TOPT. Your comments on the test are valued and will be given full consideration in the test revision process before the TOPT is finalized. Your frank input on the test will help ensure that the final version of the TOPT is a fair test that allows all examinees to demonstrate their current ability to speak French or Spanish.

Your feedback on the test is being collected in two formats. The first (Part I) is through your responses to statements in this booklet. You will record your responses on the blue, machine-readable response sheet. The second (Part II) allows you to write your own comments in response to the issues raised in Part I. It also allows you to describe any concerns you have about the test. You will record your responses to this part in the white booklet.

IMPORTANT: In giving your feedback on the test, please remember that the purpose of the TOPT is to **provide each candidate with the opportunity to demonstrate** his or her current ability to speak French or Spanish. In other words, its purpose is to capture on tape a "snapshot" of one's ability to speak French or Spanish. Think about your own performance on the TOPT you just completed. Our goal in developing the TOPT is to ensure that a person listening to the tape containing your responses gets an accurate picture of your current ability to speak French or Spanish.

PART I. MACHINE-READABLE RESPONSE SHEET. (Use a **No. 2 pencil** to mark your responses.)

STEP 1 IDENTIFICATION

Please fill out the information requested in the upper right-hand corner of the blue machine-readable response sheet. Be sure to circle the language in which you took the TOPT (French or Spanish), the form of the TOPT you took (A, B, C, or D), and the subject area you are preparing to teach or are already teaching (French, Spanish, or bilingual education).

STEP 2 ID Number

Please write your social security number in the boxes in the area entitled ID NUMBER on the top left-hand corner of the machine-readable response sheet. Then fill in the circle corresponding to the number in each box.

Please do not write in this booklet!

STEP 3 BACKGROUND INFORMATION

For demographic purposes, please answer each lettered question presented on the next page in the box labeled BACKGROUND INFORMATION. Write your answer in the area entitled SPECIAL CODES on the top left-hand corner of the response sheet. For each lettered question (A through G), write the number of your answer in the block on the answer sheet. Then fill in the circle corresponding to the number of your answer.

Please do not write in this booklet!

BACKGROUND INFORMATION

A. For which language did you take the TOPT?

(0) French
(1) Spanish

B. Which form of the TOPT did you take?

(0) Form A
(1) Form B
(2) Form C
(3) Form D

C. In which city did you take the TOPT?

(0) El Paso
(1) Austin
(2) Arlington
(3) Hurst
(4) Edinburg
(5) San Antonio
(6) Houston

D. What is your current status in respect to teaching?

(0) Pre-service (not yet certified)
(1) In-service (certified in Texas and teaching in the classroom)
(2) Other

E. Which type of certification do you have or will you be seeking?

(0) Elementary certificate with French specialization
(1) Secondary French certificate
(2) Elementary certificate with Spanish specialization
(3) Secondary Spanish certificate
(4) Certificate or endorsement in bilingual education
(5) (1) and (3)
(6) Other

F. What is your ethnic group?

(0) Hispanic
(1) Black
(2) White
(3) Other

G. What is your sex?

(0) Male
(1) Female

Please do not write in this booklet!

STEP 4 SELF RATING

We would like you to describe your ability to speak French or Spanish. Below are seven descriptions of different levels of ability, ordered from high to low. In the box labeled "J" in the area entitled SPECIAL CODES on the machine-readable response sheet, please write in the number of the description below that most accurately represents your ability to speak French or Spanish. After you have written in your answer, fill in the circle corresponding to the number of your answer.

(0) I can speak the language about as well as a well-educated native speaker and can handle sophisticated language tasks such as public speaking, formal interpreting, etc.

(1) Using a standard or international form of the language, I can participate effectively and with ease in both formal and informal conversations on abstract and professional topics, as well as on practical and social topics. I can discuss my particular interests and fields of competence with ease.

(2) I can handle a broad variety of everyday, school, and work situations relating to my particular interests and fields of competence. I am usually, though not always, effective in supporting my opinions and explaining or describing things in detail.

(3) I can handle informal conversations successfully. That is, I can begin, continue, and bring to completion a wide variety of conversational tasks, including those involving a complication or those generated by an unforseen turn of events. Using general vocabulary, I can communicate facts and talk casually about topics of current public interest and of personal interest.

(4) I can handle most uncomplicated communication tasks and social situations. For example, I can discuss my background, interests, and leisure time activities. I have some ability, although limited, to converse on impersonal topics such as current events.

(5) I can handle very simple face-to-face conversations on familiar topics such as my family, the weather, food, clothing, etc. I can ask and answer simple questions, usually without difficulty.

(6) My ability to ask and answer questions is limited to the use of memorized utterances, although I occasionally speak in sentences.

Please do not write in this booklet!

STEP 5 RESPONSES TO STATEMENTS ABOUT THE TOPT

Listed below (and abbreviated on the blue machine-readable response sheet) are a number of statements about the individual items on the TOPT and on the test in general. For each statement, fill in the letter under the column that best reflects the degree to which you agree with the statement. The columns are as follows:

A	=	Strongly Agree
B	=	Agree
C	=	Agree and Disagree
D	=	Disagree
E	=	Strongly Disagree

Please feel free to use your test booklet to refresh your memory about the items as you respond to the statements.

Statements about individual items.

1. The **opening conversation** with the native speaker helped ease me into the testing situation.

2. In the **opening conversation**, the time allowed for making my responses was, in general, appropriate.

3. As a whole, I felt **picture item #1** allowed me to give a response that reflects my current ability to speak the language.

4. The map for **picture item #1** was clear and understandable.

5. In **picture item #1**, the time allowed for preparing my answer and making my response was appropriate.

6. As a whole, I felt **picture item #2** allowed me to give a response that reflects my current ability to speak the language.

7. The drawing for **picture item #2** was clear and understandable.

8. In **picture item #2**, the time allowed for preparing my answer and making my response was appropriate.

9. As a whole, I felt **picture item #3** allowed me to give a response that reflects my current ability to speak the language.

ID NUMBER	SPECIAL CODES
	A B C D E F G [redacted] J
0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0
1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1
2 2 2 2 2 2 2 2 2	2 2 2 2 2 2 2 2
3 3 3 3 3 3 3 3 3	3 3 3 3 3 3 3 3
4 4 4 4 4 4 4 4 4	4 4 4 4 4 4 4 4
5 5 5 5 5 5 5 5 5	5 5 5 5 5 5 5 5
6 6 6 6 6 6 6 6 6	6 6 6 6 6 6 6 6
7 7 7 7 7 7 7 7 7	7 7 7 7 7 7 7 7
8 8 8 8 8 8 8 8 8	8 8 8 8 8 8 8 8
9 9 9 9 9 9 9 9 9	9 9 9 9 9 9 9 9

GENERAL PURPOSE DATA SHEET II
form no. 70921

USE NO. 2 PENCIL ONLY

Name ______________________________

Date ______________________________

Please circle your answers

TOPT Language: FRENCH SPANISH

TOPT Form: A B C D

Your Subject: French Spanish bil ed

Strongly Disagree

Disagree

Agree and Disagree

Agree

Strongly Agree

1. The opening conversation helped ease me...----------> A B C D E
2. Opening conversation time allowed was appropriate--> A B C D E
3. Picture item #1 allowed a response that reflects---> A B C D E
4. The map for picture item #1 was clear--------------> A B C D E
5. Picture item #1 time allowed for preparing answer--> A B C D E
6. Item #2 allowed response that reflects my ability--> A B C D E
7. The drawing for picture item #2 was clear----------> A B C D E
8. Picture item #2 time allowed for preparing answer--> A B C D E
9. Picture item #3 allowed a response that reflects---> A B C D E
10. The pictures for picture item #3 were clear--------> A B C D E
11. Picture item #3 time allowed for preparing answer--> A B C D E
12. Picture item #4 allowed a response that reflects---> A B C D E
13. The pictures for picture item #4 were clear--------> A B C D E
14. Picture item #4 time allowed for preparing answer--> A B C D E
15. Picture item #5 a response that reflects ability---> A B C D E
16. The pictures for picture item #5 were clear--------> A B C D E
17. Picture item #5 time allowed for preparing answer--> A B C D E
18. Topic item #1 allowed a response that reflects-----> A B C D E
19. Topic item #1 time allowed for preparing answer----> A B C D E
20. Topic item #2 allowed a response that reflects-----> A B C D E
21. Topic item #2 time allowed for preparing answer----> A B C D E
22. Topic item #3 allowed a response that reflects-----> A B C D E
23. Topic item #3 time allowed for preparing answer----> A B C D E
24. Topic item #4 allowed a response that reflects-----> A B C D E
25. Topic item #4 time allowed for preparing answer----> A B C D E
26. Topic item #5 allowed a response that reflects-----> A B C D E
27. Topic item #5 time allowed for preparing my answer-> A B C D E
28. Situation item #1 allowed a response that reflects-> A B C D E
29. Situation item #1 time allowed for preparing answer> A B C D E
30. Situation item #2 allowed a response that reflects-> A B C D E
31. Situation item #2 time allowed for preparing answer> A B C D E
32. Situation item #3 allowed a response that reflects-> A B C D E
33. Situation item #3 time allowed for preparing answer> A B C D E
34. Situation item #4 allowed a response that reflects-> A B C D E
35. Situation item #4 allowed for preparing my answer--> A B C D E
36. Situation item #5 allowed a response that reflects-> A B C D E
37. Situation item #5 time allowed for preparing my----> A B C D E
38. I was not unduly nervous during the test-----------> A B C D E
39. I would prefer a beep signal as a signal to begin--> A B C D E
40. A person listening will get an accurate picture----> A B C D E

WRITE-IN AREA 1

WRITE-IN AREA 2

WRITE-IN AREA 3

FOR OFFICE USE ONLY

1	(A)	(B)	(C)	(D)	(E)
2	(A)	(B)	(C)	(D)	(E)
3	(A)	(B)	(C)	(D)	(E)

NATIONAL COMPUTER SYSTEMS

TOPT Trialing Feedback Form
Part II

NAME _____ DATE _____
SOCIAL SECURITY NUMBER _____

Please circle your answers

TOPT Language: FRENCH SPANISH

TOPT Form: A B C D

Your Teaching Area: French Spanish bilingual education

Test Site: El Paso Austin Arlington Hurst
Edinburg San Antonio Houston

Thank you very much for participating in this trialing of the TOPT. Your comments on the test are valued and will be given full consideration during the revision process before the final version is printed. Therefore, please answer the questions below as clearly and as frankly as you can.

Part A. In the outline below, **comment on any items to which you awarded a C, D or E in Part I.** These may be items you felt were unclear, unfair or otherwise problematical for you. Especially comment on anything that you feel interfered with your ability to answer the question to the best of your present ability to speak French or Spanish. Such things might be unclear directions, unclear pictures, unrealistic situations, too little time, etc... Feel free to suggest revisions.

Item	Comments
"Opening Conversation"	
Picture item #1	
Picture item #2	
Picture item #3	

Picture item #4	
Picture item #5	
Topic #1	
Topic #2	
Topic #3	
Topic #4	
Topic #5	

Situation #1

Situation #2

Situation #3

Situation #4

Situation #5

Part B. Use the spaces below to make any comments on the statements about the test in general.

1. I was not unduly nervous during the test.

Comments?

2. I would prefer that a "beep" signal be used in place of the French or Spanish speaker as a signal to begin my response after preparing my answer. (In other words, I would prefer that the only place French or Spanish be heard is in the opening conversation.)

Comments?

3. A person listening to the tape containing my responses will get an accurate picture of my current ability to speak French or Spanish.

Comments?

Part C. Please use the rest of this page to comment on any aspect of the test that is not covered in any of the preceding questions. We would especially appreciate any suggestions as to how this test might be improved. Thank you very much!

Appendix B

Example of Trial Examinee Comments for the TOPT

4520843 Spanish El Paso SPANISH B p4

Humorous. Unclear at first but clears up as you go on to the rest of the pictures

---

4523716 bilingual education Edinburg SPANISH B p4

accurate picture; response a bit slow

4553315 bilingual education Edinburg SPANISH B p4

This example was quite clear but I seemed to get tongue tied with my vocabulary. I also felt there should be a little more time allowed.

4554973 Spanish Edinburg SPANISH B p4

Just a little bit confusing because if you see pictures 3 and 4 on that page, the owner and the second guy yet mixed up because they have the same color shirt.

Appendix C

Instructions and Form Used to Record Quantiative and Qualitative Data From Field Test Raters (Observers)

TOPT Trialing
Observer's Evaluation Sheet

Instructions

INTRODUCTION. The TOPT is intended to elicit from each examinee a speech sample suitable for rating on the ACTFL scale. Alternatively phrased, the goal of the TOPT is to provide each candidate with the opportunity to demonstrate his or her current ability to speak French or Spanish. The tape of recorded responses should present a "snapshot" of the individual's ability to speak French or Spanish and should convey an accurate picture of the candidate's strengths and weaknesses.

First, before you begin listening to the candidate, fill out all information requested at the top of the form (except for the examinee level estimation).

Remember that the two purposes for observing the candidate's performance are 1) to judge whether items are doing their job of allowing candidates to show what they can do and 2) to inform the test revision process. Thus, you should make recommendations about how to improve the test and its items if they are not functioning as intended. There are four main areas on which you need to comment.

1. TIME. It is important that candidates have an appropriate amount of time for their responses. The majority of candidates should have time to give a complete response without having to wait during a long silence for the next item to begin. Waiting can create nervousness. On the other hand, if candidates are interrupted too often by the next question, they can also get nervous. If a candidate is cut off because time is too short, there should be enough of a sample of the type of speech elicited by that item on the tape to give the rater a good idea of the candidate's ability to deal with that language function and of where the candidate would have gone if he or she had had more time.

On the observer response sheet, your feedback on time problems is requested in the area for each item marked "T".

No Problem with Time — Circle "NP" if there was there was *No Problem* with time for this candidate on that item. (This includes being cut off but still giving an appropriate sample.)

Too Much Time Allowed — If *too much* time was a problem, mark the timeline with an "X" to show the approximate number of seconds the candidate had to wait for the next item to begin (under the + + + area). Example for 7-8 seconds too much time:

```
+ + +           - - -
15--10--5--0--5--10--15
      X
```

Too Little Time Allowed If *too little* time was a problem, mark the approximate number of additional seconds the candidate could have used in order to demonstrate his or her ability with this task (under the - - - area). Example for 5 seconds too litle time:

```
+ + +          - - -
15--10--5--0--X--10--15
```

Note: Marks in between the five second intervals printed are allowed and encouraged; i.e, a mark between a "5" and a "0" indicates that 2-3 seconds are intended.

2. **CONTRIBUTION OF RESPONSE TOWARDS MAKING AN ACTFL RATING.** It is important to know something about the quality of the speech sample elicited from the candidate by the item. For each picture, situation, and topic item, and for the opening conversation as a whole, you are asked to make a judgment on the quality of the speech elicited in terms of the goal of getting a ratable speech sample. Make your judgment in the area marked "C". Circle either a 1, 2, or 3 to show your judgement, where:

1 = Speech elicited by item not useful in making a rating
2 = Speech elicited by item useful in making a rating
3 = Speech elicited by item very useful in making a rating

3. **COMMENTS.** The rest of the area is for your comments. You should consider the following:

(a) **BLOCKING:** Comment on anything that appeared to block the examinee's response; i.e., did not allow the examinee to give as complete a response as he or she may have been able. It could be unusual vocabulary items. in which case write the offending word. It could also be an unclear understanding of the directions to the item. It could also be a question of time, i.e., not enough time to think about an answer. It could also be a problem with the French or Spanish prompt following the English task description.

(b) **OTHER PROBLEMS:** Comment on other problems you notice with the item on the basis of the candidate's performance on it.

(c) **RECOMMENDATIONS:** Be sure to write down any ideas that come to mind to remedy problems you have noticed during observation.

NOTE: If you are observing the candidate live, be sure to write down any questions you may have for him or her about any unexpected performance or behavior you observe during the taking of the test.

4. **EXAMINEE LEVEL ESTIMATION.** After you listen to the candidate, please estimate, to the best of your ability, the examinee's ACTFL level. Write this estimate in the space provided on the top of the first page of the form. This is **NOT** intended as an official rating. Your estimation will solely be used to help select tapes for rating in the next phase of this project.

TOPT Trialing
Observer's Evaluation Sheet

Name of Examinee _____
Social Security Number _____

TOPT Language: FRENCH SPANISH

TOPT Form: A B C D

Examinee's Teaching Area: French Spanish bilingual education

Observer's Name _____

Date of Observation _____

Examinee Level Estimation _____

Conversation

1.
T: NP
+ + + - - -
15--10--5--0--5--10--15

2.
T: NP
+ + + - - -
15--10--5--0--5--10--15

3.
T: NP
+ + + - - -
15--10--5--0--5--10--15

4.
T: NP
+ + + - - -
15--10--5--0--5--10--15

5.
T: NP
+ + + - - -
15--10--5--0--5--10--15

6.
T: NP
+ + + - - -
15--10--5--0--5--10--15

7.
T: NP
+ + + - - -
15--10--5--0--5--10--15

8.
T: NP
+ + + - - -
15--10--5--0--5--10--15

9.
T: NP
+ + + - - -
15--10--5--0--5--10--15

10.
T: NP
+ + + - - -
15--10--5--0--5--10--15

(for total opening conversation)
C: 1 2 3

Picture #1
T: NP
+ + + - - -
15--10--5--0--5--10--15

C: 1 2 3

Picture #2
T: NP
+ + + - - -
15--10--5--0--5--10--15

C: 1 2 3

Picture #3
T: NP
+ + + - - -
15--10--5--0--5--10--15

C: 1 2 3

Picture #4
T: NP
+ + + - - -
15--10--5--0--5--10--15

C: 1 2 3

Picture #5
T: NP
+ + + - - -
15--10--5--0--5--10--15

C: 1 2 3

Topic #1
T: NP
+ + + - - -
15--10--5--0--5--10--15

C: 1 2 3

Topic #2
T: NP
+ + + - - -
15--10--5--0--5--10--15

C: 1 2 3

Topic #3
T: NP
+ + + - - -
15--10--5--0--5--10--15

C: 1 2 3

Topic #4
T: NP
+ + + - - -
15--10--5--0--5--10--15

C: 1 2 3

Topic #5
T: NP
+ + + - - -
15--10--5--0--5--10--15

C: 1 2 3

Situation #1
T: NP
+ + + - - -
15--10--5--0--5--10--15

C: 1 2 3

Situation #2
T: NP
+ + + - - -
15--10--5--0--5--10--15

C: 1 2 3

Situation #3
T: NP
+ + + - - -
15--10--5--0--5--10--15

C: 1 2 3

Situation #4
T: NP
+ + + - - -
15--10--5--0--5--10--15

C: 1 2 3

Situation #5
T: NP
+ + + - - -
15--10--5--0--5--10--15

C: 1 2 3

PLEASE MAKE ANY ADDITIONAL COMMENTS HERE:

Appendix D

Example of Field Test Rater Comments for TOPT-Spanish, Form B (Picture 2)

SPANISH B

2824699

Stansfield
2.0

p1 +10 3

SPANISH B

4507701

Tisnado
2.0

p1 +7.5 2

SPANISH B

3716290

Downey
1.90

p1 0 1 Gave directions as if the man who asked the question were looking at the map also. Used words like "right here" as if he were pointing to something on the map.

SPANISH B

4553315

Tisnado
1.90

p1 +7.5 2

SPANISH B

4576537

Marc Ferrara
1.80

p1 0 . Addresses executives as "tu". Distinction between "derecha" and "derecho" unclear.

SPANISH B

4661702

Downey
2.0

p2 5 1 Didn't give any details and had lots of extra time. Maybe should include "Include as many details as possible" in the instructions.

SPANISH B

2824699

Stansfield
2.0

p2 . 3

SPANISH B

3716290

Downey
1.90

p2 0 3

SPANISH B

4553315

Tisnado
1.90

p2 2 She would have needed much more time since she did not follow instructions properly. She started describing "houses" more than activities at American homes.

SPANISH B

4576537

Marc Ferrara
1.80

p2 0 . la "estova" for stove, "lampas" for lamp

SPANISH B

4503904

Tisnado
.

s3 0 2

SPANISH B

4649411

Downey
(missing)

s3 -5 2

SPANISH A

4634975

Stansfield
4

s3 +10 3 Student did not name a place. Instead, she cited advantages of school trips.

SPANISH A

4576868

Stansfield
3.0

s3 0 3 Proposes to take group to Danals Supermarket in Dallas, where they will learn about Mexican food & customs.

SPANISH A

4635590

Bass
3

s3 -10 3

SPANISH A

4590691

Stansfield
3.0

s3 0 . E invited Isabel to go with him to San Antonio