DOCUMENT RESUME

ED 360 844 FL 021 421

AUTHOR Kenyon, Dorry; Stansfield, Charles W.
TITLE Evaluating the Efficacy of Rater Self-Training.
INSTITUTION Center for Applied Linguistics, Washington, D.C.
PUB DATE 3 Aug 93
NOTE 27p.; Paper presented at the Annual Meeting of the Language Testing Research Colloquium (15th, Cambridge, England, August 3, 1993).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Bilingual Education; Comparative Analysis; Evaluators; Individual Characteristics; *Interrater Reliability; *Language Proficiency; *Language Tests; Models; Motivation; Performance Tests; *Rating Scales; Testing
IDENTIFIERS ACTFL Proficiency Guidelines; Rasch Model; *Texas Oral Proficiency Test

ABSTRACT

This paper examines whether individuals who train themselves to score a performance assessment will rate acceptably when compared to known standards. Research on the efficacy of rater self-training materials developed by the Center for Applied Linguistics for the Texas Oral Proficiency Test (TOPT) is examined. Rater self-materials are described and analysis of data is reported from a study conducted during their development. Eight individuals worked through the materials on their own and submitted qualitative feedback on the materials and on their experience with them. They also participated in a calibration study in which they independently rated recorded segments from a TOPT administration. These ratings were analyzed by traditional approaches and by a multifaceted Rasch approach. Findings indicate that the raters as a group scored the TOPT consistently, although differences in rater severity led to some important disagreements with the rating key. The study illustrates the role of background characteristics and motivation in the success of rater self-training. (Contains 8 references.) (JL)

Evaluating the Efficacy of Rater Self-Training

Dorry Kenyon and Charles W. Stansfield, Center for Applied Linguistics

Paper presented at the 15th annual *Language Testing Research Colloquium*, University of Cambridge, Cambridge, England, August 3, 1993

Abstract

Can individuals who train themselves to score a performance assessment rate acceptably when compared to known standards? This paper examines the efficacy of self-training materials developed to train individuals to rate the *Texas Oral Proficiency Test* (TOPT), a performance assessment of oral proficiency.

This paper describes the rater self-training materials and reports on the analysis of data from a study conducted during their development. Eight individuals worked through the materials on their own and submitted qualitative feedback on the materials and on their experience with them. They also participated in a calibration study in which they independently rated recorded segments from a TOPT administration. These ratings were analyzed by traditional approaches and by a multi-faceted Rasch approach.

The study reveals that the raters as a group scored the TOPT consistently, although differences in rater severity led to some important disagreements with the rating key. The study illustrates the role of background variables in the sucess of rater self-training.

Keywords

Rater Training, Performance Assessments, multi-faceted Rasch analysis

Problem

Can individuals train themselves to score a performance assessment holistically? How acceptable will their ratings be when compared to those of professionally-trained raters and rater trainers?

As the use of performance assessments and performance-based testing in large-scale assessments grows, rater training will become an increasingly important topic (for examples, see Baker, 1992; Braungart-Bloom, 1986). The Educational Testing Service (ETS) has several large-scale programs involving extensive rater training such as the Test of Written English and the Test of

Spoken English, which are associated with the TOEFL, and the various direct writing assessments that are associated with other ETS tests, such as the NTE, the Graduate Record Examination, and the Scholastic Aptitude Test. Many local school districts as well as state school systems now mandate performance-based writing assessments, producing materials and outlining methods to train raters (for examples, see Haven and Anderson, 1987; Mueller *et al.*, 1984).

The most common model for training raters for large-scale assessments has been group-based, under the direction of one or more qualified trainers. A common model for group training is to have a trainer (called a chief reader in the writing assessment area) who trains new raters for an extended period of time. This training can range from a couple of hours in the case of some writing tests to one full week, in the case of the Oral Proficiency Interview of the American Council on the Teaching of Foreign Languages (ACTFL). Sometimes novice raters must demonstrate a certain minimum level of proficiency in rating in order to take part in the operational rating program. This is often the case in high stakes testing programs where there is a demand for reliability at the level of the score of the individual examinee.

An alternative to group-based training is rater self-training. In this model, individuals are provided specially-developed materials that they can work through on their own, at their own pace and using their own strategies for learning how to rate.

This paper presents research on the efficacy of rater self-training materials developed by the Center for Applied Linguistics (CAL) for the *Texas Oral Proficiency Test* (TOPT).[1] The TOPT is a simulated oral proficiency interview (SOPI) which elicits speech from examinees through a master tape and a test booklet. Examinee speech is recorded on a second tape that is then scored by two raters using a slightly modified version of the ACTFL scale. The Spanish version of the TOPT is used by Texas in the certification of Spanish and bilingual education teachers. A score of Advanced on the test is required for certification.

---

[1] One of the authors, Stansfield, directed the development of the *TOPT Rater Training Kit*. He and Judy Liskin-Gasparro, who jointly developed the operational rater training program for the TOPT, served as the chief raters in scoring and selecting examples of examinee speech used in the materials. Assisted by Meg Malone and Sylvia Rasi of CAL, they also prepared the written explanations explaining the ratings awarded to each speech segment.

Background

The authors know of only one example of self-training materials for a tape-mediated assessment of oral proficiency. This is the SPEAK (*Speaking Proficiency English Assessment Kit*) (ETS, 1982).[2] The SPEAK is an off-the-shelf product for administering and scoring the *Test of Spoken English* (TSE) by institutions. The SPEAK includes materials for both rater-training and for administering the test. Materials provided for rater training include (1) a copy of the *Guide to SPEAK*, a manual containing instructions for both working through the rater training program and for administering the test, (2) eight training cassettes, (3) six testing cassettes, (4) a SPEAK test book, (5) two copies of the SPEAK scoring rubric, and (6) one hundred SPEAK rating and score summary sheets.

Scoring for the SPEAK is analytical. Raters score the six scoreable sections of the SPEAK on four-point scales of pronunciation, grammar, fluency, and comprehensibility. Scores in each category are averaged across items in each section. These section scores are then averaged to get a final score in each category. The diagnostic scores (pronunciation, grammar and fluency) are rounded to the nearest tenth. The comprehensibility score is multiplied by 100 and rounded to the nearest ten unit.

Rater trainees can evaluate their performance on the six testing tapes. The *Guide to SPEAK* informs trainees that if the difference between what they award as the average score in each of the four categories and the answers given in the manual for that testing tape is less then .95, then they are within the accepted range on that testing cassette. If the difference is greater, then they should review the training tapes again.

In contrast, the TOPT is rated holistically according to the ACTFL *Guidelines* (1986). The TOPT consists of fifteen items. Each item is designed to elicit examinee speech using speaking tasks described by the *Guidelines*. These tasks are grouped into three sections as outlined in Figure 1.

---

[2] One of the authors, Stansfield, conceptualized the SPEAK and authored the *Guide to SPEAK* while serving as director of the *Test of Spoken English* program at ETS in 1982.

Figure 1
Speaking Tasks on the TOPT

Picture-based Items

- Give Directions
- Describe a Place/Activities
- Narrate in Present Time
- Narrate in Past Time
- Narrate in Future Time

Topic Items

- Give Instructions
- State Advantages/Disadvantages
- Give a Brief Factual Summary
- Support an Opinion
- Hypothesize on an Impersonal Topic

Situation Items

- Speak with Tact (e.g., Apologize, Lodge a Complaint)
- Speak to Persuade Someone
- Propose and Defend a Course of Action
- Give a Professional Talk
- Give Advice

Raters can award one of five ACTFL levels (Intermediate-Mid, Intermediate-High, Advanced, Advanced-High, or Superior) to each of the 15 speaking performances. Each speaking task has a prescribed ACTFL level associated with it, which corresponds to the level of proficiency required to perform that task adequately. Tasks at lower levels (for example, giving directions) generally limit the range of scores that can be awarded to it, since successful completion of the task does not demand evidence of higher level proficiency.

A single final holistic ACTFL level is awarded to the examinee. This is not a simple average of the item-level scores. Instead, the final score is the highest level on the scale at which the examinee displays consistent performance.

The *TOPT Rater Training Kit* was developed to provide interested persons at institutes of higher education in Texas with a better understanding of the TOPT, and to enable them to test and score their own students on a disclosed form of the test. In this way, they will be able to give guidance to their students as to their readiness to take the operational test. To achieve this, the materials would have to familiarize trainees with the TOPT and the ACTFL scale, to train them to apply it in evaluating examinee performances on each of the 15 TOPT items, and to instruct them on how to award the single global score based on performances on the TOPT items. Upon completing the

materials, trainees should be able to advise candidates reliably on the likelihood of their receiving a passing score on the operational TOPT.

Description of the Materials

The *TOPT Rater Training Kit* consists of the several components. The first is the *Rater Training Manual*. This handbook describes the TOPT and provides instructions on (1) scoring each item, (2) assigning a global rating, (3) completing the practice scoring exercises, and (4) administering the *Test Kit Version* of the TOPT. The second component is a set of three cassette tapes. Tapes A and B contain excerpts of examinee performance to accompany descriptions and exercises in the *Rater Training Manual*. Tape C contains practice calibration sets used to determine the trainee's ability to rate individual items accurately (Side 1) and the complete responses of one examinee to the TOPT (Side 2). The third component of the Kit is the *TOPT Rater Training Kit Workbook*. This workbook is used by trainees to record their notes and ratings of the practice samples as they work through the materials. The fourth component is the *Reference Guide for Scoring the TOPT Rater Training Kit Materials*. This booklet contains a copy of the *ACTFL Proficiency Guidelines for Speaking* as well as the key to and explanations for all exercises contained in the Kit.

The *Rater Training Kit* also contains materials for administering the *Test Kit Version* of the TOPT. These materials are the TOPT Test Booklet (five copies), the Test Master Tape (one copy), and the Rater's Evaluation Sheet (five copies, which may be photocopied).

Chapter 1 of the manual instructs trainees on how to proceed through the kit materials. They are encouraged to begin by reading Chapters 2 and 3, which present background information on the TOPT. They are then advised to administer the TOPT to themselves in order to understand the test better. Following this, they are to go on to Chapter 4, which provides an introduction to the ACTFL Guidelines and how the TOPT is scored. Then, trainees work through Chapters 5 and 6 very carefully. These chapters provide instruction on how to score each TOPT item. For each item, trainees are instructed to follow this sequence.

1. In the *Manual*, read the description of the item and the general scoring guidelines for scoring the item.
2. Study the profiles of typical performance at each score level for that item.
3. Review the tips for rating that item.
4. Listen to the sample performances for that item on the tape and try to assign a rating to each.

5. Read the justifications for the ratings of these sample performances.
6. Listen to the sample performances again whenever it is not clear why the ratings were assigned.
7. Listen to the two exercises (i.e., two examinees responding to the item) for the item, and assign a rating to each in the *Workbook*.
8. Compare the ratings given and notes made in the *Workbook* with the ratings and notes printed in the separate *Reference Guide*.

Chapter 5 of the *Manual* provides instruction on five of the fifteen items. The tape presents the recorded performances and the manual provides detailed analyses of ratings for each possible level for each of the five items. Chapter 6 is similar, but provides detailed analyses for only two possible ratings. Usually this is for a performance below Advanced and one at Advanced or above.

Once trainees have completed studying all the items and have practiced rating each one, they go on to Chapter 7. That chapter instructs them in assigning global ratings. An exercise at the end of the chapter gives trainees the opportunity to check their understanding of this process. A short chapter follows containing information on using the "warm-up" section of the TOPT. The last chapter gives instructions on using Tape C. Side One of Tape C provides three calibration exercises. In these exercises, trainees listen to three sets of three examinees each responding to three TOPT items. Trainees assign a rating to each segment. After rating all three segments, they assign a single global rating to the examinee based on the highest level at which the examinee showed consistent performance. This follows the same logic as assigning a rating based on all segments. Trainees check their own ratings and notes against those provided in the separate reference guide. Side Two of Tape C presents the entire TOPT performance of a single examinee. Again, trainees are instructed to score each segment, assign a global rating, and then to check their work against the ratings and notes provided in the *Reference Guide*.

There is one final component to the kit. It is possible for trainees to receive a certificate of their achievement in learning how to score the TOPT. To do so, they write to CAL, which then sends them a calibration tape containing the performance of fifteen examinees responding to three items each. Instructions for scoring are the same as for the calibration test provided in the materials. Trainees have one week to rate each item and award a global rating to each examinee. They then return the tape and their ratings to CAL. CAL awards them a certificate of achievement if their performance in rating the examinees on the calibration tape is acceptable.

The manual suggests that as an alternative to working through these materials individually, individuals may want to listen to and discuss the ratings of the sample performances in a small group. In this way, the kit could serve as the basis for an in-service workshop for language teachers who want to learn how to score the TOPT or learn more about oral proficiency as represented by the ACTFL *Guidelines*.

The manual reminds trainees that learning to understand the proficiency scale and to rate speech samples is not a skill that can be quickly acquired. It suggests that it will take approximately ten hours to work through the kit materials.

Description of the Study

The current study uses data collected during the piloting of the rater self-training materials. Besides examining the efficacy of the materials, the pilot study also collected feedback from trainees to use in improving the materials. CAL staff carefully evaluated all of the trainees' comments and revised the materials on the basis of these comments. While the final revisions may improve the materials slightly, we believe that the data collected during the piloting can serve as a conservative indicator of the efficacy of these self-training materials.

Subjects

In an earlier project, CAL staff acquired names of individuals in the Washington, D.C. area who were interested in learning how to score the TOPT. CAL staff contacted these individuals to request their participation in the piloting of the self-training materials. CAL sent the kit to eleven persons who were work through the materials on their own. Nine of these returned formal written feedback. Eight participated in the calibration test and are the ones discussed in this article. Three could not participate in the calibration test due to scheduling conflicts. Seven of the eight who took the calibration test provided formal written feedback on the kit.

Materials

Data came from three sources. Qualitative data on the trainees' perceptions of the kit and its ability to train them to score the TOPT came from a five-page feedback form entitled the "TOPT Rater Training Kit Evaluation Form of the Preliminary Version." The trainees completed this form after working through all the materials. Qualitative data also came from personal interviews with the trainees following self-training.

Quantitative data was generated during the final calibration task. In this task, the trainees listened to the final

calibration tape, which contains the responses of fifteen examinees divided into five sets of three examinees each. Within each set, the examinees responded to the same three TOPT items. Items differed across the five sets. The only commonality between the sets was that two of the three items were at the Advanced and one at the Superior level. Trainees recorded their notes and their ratings to each taped segment on specially prepared notes-sheets; one notes-sheet was completed on each examinee. At the bottom of each sheet, the trainees recorded the global rating for the examinee.

Method

CAL sent the *TOPT Rater Training Kit* to each subject at home. They were allowed three weeks to work through the materials and complete the feedback form. At the end of three weeks, they made appointments to come to CAL where they took the calibration test and were debriefed on the kit. This took place in small groups ranging in size from one to three persons. At the meeting, the trainees returned all materials to CAL.

Subjects were volunteers and were not paid for their participation in the piloting. However, after revising the kit materials according to feedback obtained through the piloting, CAL sent each trainee a copy of the final version of the kit.

Results

Qualitative Analysis

All of the trainees were currently teaching Spanish. Seven of the eight were teaching at the college level, one as a teaching assistant. One trainee was teaching at the high school level. Seven of the eight were native or near-native English speakers; one was not a native speaker of English, but spoke Hungarian as a first language and Spanish as a second. This person did not return a completed feedback form.

There was a wide range of previous familiarity with the ACTFL Guidelines. At the upper end, one of the eight trainees indicated that he had been certified by ACTFL to administer the OPI (though his certification was no longer current). At the lower end, one indicated no familiarity with the Guidelines at all. Table 1 shows the level of familiarity with the ACTFL Guidelines for each of the eight trainees.

Table 1
Level of Familiarity with the ACTFL Guidelines

Trainee	Level of Familiarity
A-5	Previously ACTFL Certified
B-4	Attended 4-day training
C-4	Attended 4-day training
D-3	Received some training
E-2	Somewhat familiar
F-2	Somewhat familiar
G-2	Somewhat familiar
H-1	Not familiar

Table 1 shows that four of the eight had some training in the Guidelines, three had some familiarity but no training, and one was a complete novice. Note that the trainees have been coded according to their level of familiarity with the ACTFL scale, with 5 being most familiar and 1 being least familiar.

The following responses are based on the seven trainees who returned completed feedback forms. All trainees indicated that they had worked through all the materials, though some left out some portions that they felt were either already familiar to them or they felt were unimportant. The number of self-reported hours spent on the materials ranged from an estimate of 6-8 hours to 20 hours. The amount of time each trainee spent working on the materials is listed in Table 2.

Table 2
Number of Self-Reported Hours
Spent on the Self-Training Materials

Trainee	Time Spent
B-4	6-8 hours
H-1	8 hours
E-2	8-8.5 hours
C-4	10 hours
D-3	10-12 hours
A-5	13 hours
F-2	20 hours

Table 2 suggests that ten hours is a realistic estimate of the average time it takes to work through the materials. The trainees' comments suggest that some trainees worked through the materials in small spurts, while others worked in larger blocks of time. As can be seen in Table 2, there does not appear to be a strong relationship between the degree of familiarity with the Guidelines and the amount of time spent with the materials.

The feedback sheet asked the trainees to indicate on a 4-point scale (with 1 being low and 4 being high), their response

to several questions about the kit and its materials. Of interest here are responses to the four general questions.

The first question asked the trainees to indicate how well they felt the *TOPT Rater Training Kit* prepared them to score the *Test Kit Version* of the TOPT (i.e., how confident they would feel scoring examinees). Three of the trainees awarded this question a 4, and four gave it a 3. Only three trainees provided comments on this question. One said that he felt "fairly confident to score examinees." Another stated that she didn't do well on the distinction between Advanced-High and Superior. Finally, one trainee stated that she didn't feel confident yet as she was often one-level off in her ratings and still had a few questions about the scoring.

The second question asked the trainees to show how well they felt the *TOPT Rater Training Kit* acquainted them with the TOPT itself. All of the trainees gave this a 4. The single comment was "very thorough."

Another question asked the trainees how well they felt the *TOPT Rater Training Kit* helped them understand how the TOPT is scored. Four trainees gave it a 4 and three gave it a 3. The single comment was "good, but parts were a little unclear to me."

The final general question asked the trainees whether they would use the TOPT *Test Kit Version* with their own students, and to explain why or why not. Only the high school instructor felt he would not use the materials since language lab facilities at his school were inadequate. The others suggested specific uses, such as using it as part of a senior comprehensive exam, as part of a procedure for screening teacher certification candidates who received their language training at a different school, with student teachers, or as additional practice for students. One trainee added:

> I now think a tape mediated oral test is a real possibility--this gives me confidence to consider designing such a test for use in 1st, 2nd year classes. The rating of the test and training of raters is still a big issue, but I feel that giving students the same questions is an improvement in reliability over the more open interview format we've been using.

Another trainee, who had limited exposure to the ACTFL *Guidelines*, added:

> I now have a more definite idea of what I am looking for in their speaking skills and feel I can rate them more precisely now.

In general, the trainees were very enthusiastic about the

kit. In their written responses to other items, they gave us specific suggestions for improving the materials. They also elaborated on some of these suggestions during their debriefing sessions at CAL. Most of these comments served to make the materials more "user-friendly," to correct typos, or to ask for additional justification for a specific rating. On the whole, the trainees felt that they had gained ability in scoring the test.

Empirical Results

As described above, the trainees were administered a calibration tape consisting of fifteen examinees responding to three TOPT items each at CAL. As in the operational program, the trainees had their own tape and cassette player. They were allowed to listen to each response as many times as they wanted, and they had no time limit in which to complete the task. The fifteen examinees on the calibration tape did not appear in the training materials.

Each trainee awarded an ACTFL score to each of the fifteen examinees on each of the three items to which each examinee responded. Then the trainee assigned a global score based on the ratings of the three items. The standard for passing the calibration test follows that used in the TOPT operational program. To pass the calibration test, a trainee must have 20% or fewer discrepancies with the key. A discrepancy is defined as any disagreement that crosses the cut-score, which is the ACTFL Advanced level.[3]

The key to the calibration set was constructed by having two trainers for the TOPT operational program independently rate each segment and award global ratings. In addition, two CAL staff members who had undergone TOPT training also scored the calibration set. Finally, preliminary segment ratings (i.e., scores on each item for each examinee) that were awarded by TOPT trainers when they first reviewed the tapes were available. The key was based on the score awarded by the two trainers. In cases where they disagreed, the ratings of the CAL staff members and the preliminary segment ratings were consulted, and the key was

---

[3] There is another condition in the operational program that is considered a discrepancy. This is when the global score is two or more steps away from the key but did not cross the cut score. This only occurs when the trainee awards an Advanced and the key awards a Superior (or vice-versa) as a global rating. Since scores of Superior are converted to Advanced-High for score reporting purposes, in developing the TOPT *Rater Training Kit*, we did not allow a global score of Superior to be reported. Thus, this type of discrepancy could not appear in our calibration.

the modal rating of the five available ratings.

In the empirical analysis of this data, the ACTFL ratings were converted to integers as follows: Intermediate-Mid = 1; Intermediate-High = 2; Advanced = 3; Advanced-High = 4; and Superior = 5.

Table 3 shows the number of absolute agreements at the pass (Advanced or above)/fail (Intermediate High and below) level for each trainee. Less than 80% agreement would be considered failing the calibration set according to the rules followed in the operational program.

Table 3
Percent of Absolute Agreements on the Pass/Fail Level on the 15 Global Calibration Ratings

Trainee	Percent of Agreement
A-5	100%
B-4	93%
C-4	93%
D-3	93%
E-2	80%
F-2	100%
G-2	60%
H-1	73%

Table 3 shows that six of the eight trainees passed according to the criterion used in the operational scoring program. It may be noted that in the operational TOPT program, about 90% of the trainees pass the calibration set after ten to twelve hours of training. Given the small group in our study, these two figures may be comparable.

Table 3 suggests some relationship between the degree of familiarity with the Guidelines and passing the calibration tape. This is interesting in light of a comment made by a trainee, who had some training with the Guidelines. He commented that the materials should try to gage the trainee's familiarity with the Guidelines, adding more exposure and practice for those without much and allowing those with a high degree of familiarity to skip these sections.

Although not examined in the operational program, we examined the consistency of ratings awarded to the three individual items for each speaker. There were 45 such items on the calibration tape. Table 4 presents the percent of absolute agreements at the pass/fail level on these 45 segment ratings. According to the key, nine of the 45 segment ratings were below

pass, 36 were above pass.

Table 4
Percent of Absolute Agreements on the Pass/Fail Level on the 45 Segment Calibration Ratings

Trainee	Percent of Agreement
A-5	93%
B-4	90%
C-4	90%
D-3	90%
E-2	84%
F-2	100%
G-2	51%
H-1	76%

As expected, the trainees who passed on the global rating performed similarly well on the segment ratings. The same two trainees performed below the 80% agreement mark.

Another way of approaching the quality of these ratings is to examine the degree of agreement between the ACTFL level awarded by each trainee with that awarded by the key. One method to do this is by simple correlations between the trainee's rating and the key. A second way is to examine the number of agreements between the two. An absolute agreement indicates that both the trainee and the key awarded the same level on the ACTFL scale. A one step agreement indicates that adjacent levels on the scale were awarded (e.g., the key stated Advanced and the trainee gave Advanced-High). A two step agreement occurs when, for example, the key gives Advanced and the trainee gives Intermediate-Mid as a rating.

The first column in Table 5 presents the Pearson Product-Moment correlation between each trainee and the key. The next three columns show the percent of agreements: absolute agreement, one-step away, and two-steps away.

Table 5
Correlations and Percent of Absolute Agreements
on the ACTFL Level Assigned
on the 15 Global Calibration Ratings

Trainee	Correlation	Percent of Agreements		
		Absolute	1-Step	2-Steps
A-5	.84	69%	31%	0%
B-4	.92	87%	13%	0%
C-4	.81	47%	53%	0%
D-3	.81	47%	53%	0%
E-2	.50	47%	47%	6%
F-2	.83	60%	40%	0%
G-2	.67	33%	47%	20%
H-1	.78	33%	67%	0%

Table 5 shows that there was a wide range of correlations, ranging from .50 to .92. The percent of absolute agreements ranged from a low of 33% to a high of 87%. Only two raters had any ratings two steps away, though one of these had 20% such ratings.

Table 6 presents similar information on the 45 segment ratings. It is expected that these would be similar to those for the global ratings, though perhaps performance would be somewhat lower, since the unit of analysis is the individual performance rater than a global rating based on three performances.

Table 6
Correlations and Percent of Absolute Agreements
on the ACTFL Level Assigned
on the 45 Segment Calibration Ratings

Trainee	Correlation	Percent of Agreements		
		Absolute	1-Step	2-Steps
A-5	.70	52%	46%	2%
B-4	.81	67%	33%	0%
C-4	.78	36%	62%	2%
D-3	.70	51%	44%	5%
E-2	.61	53%	44%	2%
F-2	.74	44%	47%	9%
G-2	.65	29%	53%	18%
H-1	.72	40%	60%	0%

Table 6 shows a narrower range of correlations than in Table 5, ranging from .61 to .81. The percent of absolute agreements ranged from a low of 29% to a high of 67%. Only two raters had no ratings more than one step away from the key.

Tables 5 and 6 confirm Tables 3 and 4. The raters who

passed the calibration criterion had higher correlations and greater agreement with the key. However, there were a few surprises. In particular, Trainee E appears to have performed erratically when compared to the key. Although he "passed" by the standard criterion, and at the segment level on the pass/fail criterion, the correlation between his global ratings and that of the key was only .50. This trainee also had the lowest correlation with the 45 segment ratings (.61). However, on the segment ratings he had the second highest percentage of absolute agreements with the keys. In contrast, Trainee H did not pass by the criterion, but had correlations that were similar or better than some of those who did pass (.78 on the fifteen global ratings, .72 on the 45 segment ratings).

In working with ratings, it has been our experience that two main factors lead to discrepancies between trainees and the scoring key: the relative severity of the trainee in relation to the key, and a lack of internal consistency by the trainee. One or both factors may account for discrepancies. A trainee who is consistently much more severe than the key will tend to fail more candidates than the key. Similarly, a trainee who is more lenient than the key will tend to pass more candidates than the key. Both the trainee and the key may be rating in a similar fashion, but the trainee has centered himself or herself on the scale at a locus that is quite different from where the key is centered. Thus, in terms of an absolute criterion, the two appear in disagreement.

A lack of internal consistency (i.e., low intra-rater agreement) of a trainee can also lead to discrepancies when compared to the key. A large degree of internal inconsistency relative to the key and other raters would indicate that the trainee has not yet mastered or internalized the scoring scale.

We used a multi-faceted Rasch analysis (Linacre, 1989) to study these two factors. We performed the analyses using the computer software FACETS (Linacre & Wright, 1993). The object of interest in this analysis was only the relative severity and internal consistency of the trainees. Thus, only analyses of the trainees will be reported.

Table 7 shows the results of a FACETS analysis on the fifteen global ratings. For comparative purposes, the key is also included. Trainees are listed in order of severity in the first column, with the most severe at the top and the least severe at the bottom. The second column, labeled 'average', gives the average score on the fifteen global ratings on the converted (1-4) ACTFL scale. Under the column marked 'logit' are the logit calibrations for each trainee. This puts the trainees on a linear metric in terms of their severity. The key was anchored at zero for easy comparison. Under the column marked 'error' is the amount of measurement error (on the logit scale)

of each calibration. This gives an indication of how well the trainees are calibrated and can be used to determine how meaningful differences between calibrations are. The standardized infit and outfit statistics in the last two columns give an indication of the internal consistency of each trainee relative to the whole group. These statistics are usually only a concern if they are greater than +2 or less than -2. When both infit and outfit are negative for a trainee, this can be an indication of strong unanimity or overly-consistent behavior. When both are positive, this can be an indicator that the trainee over-used scale points (i.e., had little variability) or that the trainee was inconsistent with the key and the other trainees (cf. Linacre & Wright, 1993: 75).

Table 7
Facets Analysis on the 15 Global Ratings

Trainee	Average	Measure Logit	Measure Error	Standardized Infit	Standardized Outfit
G-2	2.2	3.82	0.53	-1	0
H-1	2.7	1.77	0.55	0	0
F-2	3.1	0.00	0.56	0	0
Key	3.1	0.00	0.56	-1	-1
B-4	3.2	-0.64	0.58	0	0
A-5	3.2	-0.95	0.64	0	2
E-2	3.4	-1.75	0.64	1	0
C-4	3.6	-3.16	0.75	0	0
D-3	3.6	-3.16	0.75	-1	0

Table 7 indicates that trainees awarded a wide range of average scores (2.2 to 3.6) as global ratings. This wide range is reflected in the logit calibrations of the trainees. However, none of the fit statistics are greater than 2 or less than -2. The fact that the key has two negative fit statistics confirms that it represents a central tendency of the trainees as a group. It is interesting that the two trainees at the extreme ends of the severity scale (trainees G and D) both have a negative infit value. This confirms that they were rating according to the same internalized scale. However, it is obvious that they had centered themselves very differently along that scale. Trainee A, the most experienced trainee, has a large outfit statistic. Trainee E, whose correlation with the key was so low, is the only other trainee with a non-zero fit statistic.

Table 8 shows information on the individual misfitting ratings for this analysis. The analysis requested all individual ratings with a standardized statistic of ±2 to be reported. The first column gives the trainee. The second column in Table 8 identifies the speaker by giving the calibration set and speaker number. The third column shows the rating awarded by the

trainee. The fourth column shows the rating that would be expected given the tendency toward leniency or severity of that particular trainee across all examinees and given the ratings awarded that particular examinee across all trainees. The next column gives the absolute residual (observed score minus expected score). The last column shows the residual standardized by its own standard error. This statistic is expected to approximate a unit normal distribution.

Table 8
Misfitting Ratings for the 15 Global Ratings

Trainee	Speaker	Score Awarded	Score Expected	Residual	Standardized Residual
A-5	Set5Spk1	3	4.0	-1.0	-7
E-3	Set2Spk1	3	3.9	-0.9	-2
E-3	Set4Spk1	3	1.6	1.4	2
E-3	Set5Spk2	2	3.2	-1.2	-2

Table 8 sheds light on the misfit statistics in Table 7. Only two of the eight trainees had any misfitting ratings. Trainee A awarded a lower score than expected to the first speaker in calibration set 5. Trainee E had three misfitting ratings, two in which he awarded higher ratings than expected, and one in which his rating was lower than expected. There is no apparent pattern in these misfitting ratings.

Tables 7 and 8 together indicate that Trainee E may have been more inconsistent than desired in his ratings. Although he passed on the criterion currently used, he may not have internalized the scale. Trainees G and H, however, who did not pass the calibration exercise, do seem to be internally consistent in their ratings. The fact that they were more severe than the key led them to fail more examinees than the key did.

To better understand the trainees, we can look at the same information for all 45 segment ratings. Table 9 presents the same information as Table 7.

Table 9
Facets Analysis on the 45 Segment Ratings

Trainee	Average	Measure Logit	Measure Error	Standardized Infit	Standardized Outfit
G-2	2.0	3.66	0.30	0	0
H-1	2.7	1.00	0.29	1	1
Key	2.9	0.12	0.28	-2	-2
E-2	3.2	-0.62	0.27	0	0
B-4	3.2	-0.69	0.27	-1	0
A-5	3.2	-0.80	0.28	0	0
F-2	3.3	-0.97	0.26	3	3
D-3	3.3	-1.18	0.26	0	0
C-4	3.6	-1.87	0.26	-2	-2

Again, we see the wide spread between the average ratings awarded by the trainees. Trainee G is particularly severe, yet the infit and outfit statistics show that he is internally consistent. Trainee C is lenient, yet the negative fit statistics indicate he is overly consistent with respect to the group as a whole. Again, the key has negative fit statistics, indicating that it embodies the consensus of the group as a whole. Trainee F is problematic. She has the only fit statistics above 2. Trainee H may also be problematic, though to a lesser extent. The fit statistics suggest that these trainees may be overusing extreme categories or generally be inconsistent with themselves and the group as a whole. A look at the misfitting ratings may be helpful.

Table 10 is similar to Table 8. However, the item is introduced as a third facet. The ratings in Table 10 have been flagged as misfitting on the basis of scores awarded by the individual trainee to all speakers on all items, scores awarded the individual speaker by all trainees on all items, and the scores awarded that specific item by all trainees for all speakers.

Table 10
Misfitting Ratings for the 45 Segment Ratings

Trainee	Speaker	Item	Awarded	Score Expected	Score Residual	Standardized Residual
A-5	Set1Spk3	P4	5	3.7	1.3	2
A-5	Set3Spk3	S2	2	2.9	-0.9	-2
A-5	Set5Spk1	S4	3	4.3	-1.3	-2
A-5	Set5Spk2	P4	4	3.0	1.0	2
C-4	Set1Spk1	P4	2	3.2	-1.2	-2
D-4	Set1Spk1	P4	2	3.1	-1.1	-2
D-4	Set2Spk3	S3	5	3.7	1.3	2
E-2	Set1Spk1	S1	4	3.0	1.0	2
E-2	Set4Spk1	P3	3	1.7	1.3	2
E-2	Set5Spk2	S4	2	3.4	-1.4	-2
F-2	Set1Spk1	T4	5	3.3	1.7	3
F-2	Set2Spk2	T2	1	2.2	-1.2	-2
F-2	Set2Spk3	S3	5	3.6	1.4	2
F-2	Set2Spk3	T2	5	3.3	1.7	3
F-2	Set2Spk3	T1	4	3.1	0.9	2
F-2	Set3Spk2	S2	1	2.3	-1.3	-2
F-2	Set3Spk3	P5	4	2.8	1.2	2
G-2	Set3Spk2	S2	2	1.1	0.9	3
G-2	Set5Spk1	S4	4	2.9	1.1	2
G-2	Set5Spk2	P4	3	1.6	1.4	2
H-1	Set2Spk1	T1	4	2.9	1.1	2
H-1	Set2Spk3	S3	2	3.1	-1.1	-2
H-1	Set3Spk1	P5	2	3.3	-1.3	-2
H-1	Set4Spk2	P3	4	3.1	0.9	2
H-1	Set4Spk3	T3	4	3.0	1.0	2
H-1	Set5Spk1	T2	4	2.8	1.2	2

Most of the trainees were involved in only a few misfitting ratings. Trainee B had none. Trainee F and Trainee H, who had the only two positive misfit statistics in Table 9, naturally are involved in largest number of unexpected ratings. There is an unmistakable pattern in Trainee F's data. In Set 2 for Speaker 3, she awarded unexpectedly high marks on all three segments. This speaker was a Hispanic from a border town in Texas, a native speaker of Spanish whose fluency was stronger than other characteristics of his speech. Although the key and the other trainees saw beneath his fluency, Trainee F apparently did not, and awarded him unexpectedly high marks. This speaker was involved in more individual misfitting ratings (five) than any

other speaker. Speaker 1 in Set 1 was in second place, being involved in four misfitting ratings. This may be due to a "warm-up" effect, since he was the first speaker heard in the task. There are no other clear patterns in the data for any other trainee, item or speaker that would suggest any other consistent problem in rating.

Summary

Our original question was to examine the efficacy of the rater self-training materials. In our investigation, we learned that each trainee is unique, bringing to the task of self-training his or her own unique background, personality, and personal goals. In order to illustrate this, we present a summary description of each trainee and his or her performance.

Trainee A

Trainee A appears able to score the TOPT reliably after working through the self-training materials. On the pass/fail criterion on the global ratings he scored 100%, and 93% on the segment ratings. The correlation on the global ratings was .84 and .70 on the segment ratings. Although somewhat more lenient than the key, he showed no pattern of inconsistency in his ratings. This trainee had already been certified by ACTFL to conduct the OPI, and had conducted many interviews. His previous experience led him to inform us that he skipped over some parts of the materials and felt that others were redundant to those with OPI training (though should be kept in for beginners). He estimated he spent some thirteen hours working on the materials. He felt the materials were excellent and adequately prepared him to score the TOPT. He plans to use the materials as part of his school's senior comprehensive examination. He appears to enjoy rating. About the practice calibration sets and the sample examinee tape included in the kit he writes "I loved this part. I was especially happy when I discovered that most--not all--of my ratings were on target with those of the TOPT developers."

Trainee B

Trainee B indicated that he had attended a four-day OPI training workshop, though he had not become certified by ACTFL to administer the OPI. He also appears able to score the TOPT accurately and reliably. He had 93% percent agreement on the pass/fail criterion on the global scores and 90% on the 45 individual segment ratings. His correlations with the key were the highest of all the trainees in both categories (.92 and .81 respectively). He also had the highest degree of absolute agreement with the key in both categories (87% and 67% respectively) and none of his ratings were ever more than one scale step away from the key. In the FACETS analysis, he had no misfitting ratings on either the 15 global ratings or the 45

segment ratings. His negative fit statistics reveal a tendency to be "unexpectedly consistent" with the group as a whole. He had a very positive attitude toward the kit materials and estimated spending the least amount of time with them (6-8 hours). He felt adequately prepared to use the materials, and plans to use them on teacher certification candidates at his university.

Trainee C

Trainee C indicated that she had received training in the ACTFL OPI, but was not ACTFL-certified. She also appears competent to rate the TOPT, scoring 93% on the pass/fail criterion for the global ratings, and 90% for the individual segment ratings. Her correlation with the key was moderate (.81 on the global ratings and .78 on the segment ratings, which was the second highest). However, she was most often one scale step away on her ratings (53% on the global ratings, 62% on the segment ratings). The FACETS analysis revealed that she was the most lenient of all trainees, though consistent in her ratings. On her segment ratings she had a negative fit statistic for both infit and outfit. This result indicates that she had a high degree of relative consistency with the key and the other trainees. She had only one misfitting rating from the 45 segment ratings. Her tendency toward leniency was not enough to restrict her from passing the criterion. She appears to have mastered the scoring procedures, though she has centered herself somewhat high on the scale. The kit materials helped her be aware of her tendency. She wrote on her feedback sheet: "I don't feel very confident yet--I was often one level off on my ratings."

Compared with the others, Trainee C appeared to be more metacognitavely aware of the rating process and more analytical of herself and her performance. She writes: "I wonder about the influence of one speaker's ability on the ranking [sic] of the next...Your rationale were clearly criterion based, but I caught myself doing cross-comparisons. Any way to avoid this?" She also wondered about her being inconsistent in the way she recorded her notes while listening to the examinees. She listed five specific ratings in the kit for which she wanted further explanations. She was also more desirous of additional support. She wrote that she liked the idea of working through the kit as a group, and suggested "assigning a mentor with whom the trainee could talk by phone."

Trainee D

Trainee D was the sole high school level Spanish teacher. He indicated that he had some training in the ACTFL *Guidelines*, but had not attended the 4-day OPI training workshop. He wrote that he considered himself "basically familiar" with the *Guidelines*. He appears competent to score the TOPT, passing the

global rating criterion with 93% and the segment criterion with 90%. His correlations with the global ratings were .81 and with the segment ratings .70. More than half of his global ratings, however, were off by more than one step on the scale. The results of the FACETS analysis for Trainee D parallel those for Trainee C. Like Trainee C, he was more lenient than the key (the second most lenient trainee after Trainee C). He scored the tapes with a high degree of internal consistency and was consistent with the scale. However, he centered himself fairly high on the scale. He had a positive attitude towards the kit materials, though he felt only somewhat confident with his ability to rate the TOPT. He would not be able to use the kit materials in his high school situation, however.

Trainee E

Trainee E was a graduate teaching assistant in Spanish at a local university. He indicated that he was only somewhat familiar with the *Guidelines* before beginning this project. He appears to be a borderline pass, scoring 80% on the global rating criterion, and 84% on the segment ratings. His correlation with the global ratings (.50) was surprisingly low, and two of the fifteen ratings were two-steps away on the scale. His correlation with the segment ratings (.61) was also the lowest of all the trainees, and 9% of these were two-steps away on the scale. The FACETS analysis shows that he was somewhat more lenient than the key. More importantly, it revealed that he had three misfitting global ratings. In two cases he awarded a score much higher than expected, and in one case much lower. On the segment rating level, however, he appeared more consistent and was only slightly more lenient than the key. Trainee E appears to be somewhat unstable in his global ratings, though adequate in the segment ratings. His comments on the kit were brief and positive, and he wrote that now he felt fairly confident to score examinees. He appeared to have little self-reflection on the rating task. His inconsistencies may be due to his lack of teaching experience.

Trainee F

Trainee F indicated that she had little exposure to the *Guidelines* before beginning this project, and that she had no prior background or experience on rating. However, she was very interested in testing oral proficiency. This was one of her responsibilities at her school and she felt somewhat inadequate on how to do it. She spent the most time of any trainee on the kit ("20 hours, maybe more"), but was perhaps the most enthusiastic about the experience. She indicated that the kit has given her an approach to rating the speaking skills of her students that she didn't possess before and that she was grateful to have had this experience. Her diligence with the materials appears to have been rewarded. On the criterion, she scored

100%. The experienced OPI tester was the only other one to score 100% on this criterion. On the segment ratings, she was the only one to have 100% agreement on the pass/fail basis. She had respectable correlations with both measures (.83 and .74, the third highest in each case). The FACETS analysis for her was revealing. On the global ratings, she was calibrated exactly at the same point as the key, though on the segment ratings she was somewhat more lenient than the key. On the segment ratings, however, she had a surprisingly high misfit statistic, which is often an indication of inconsistent rating behavior. However, when her single misfitting ratings were analyzed, it is clear that she had trouble with a single native Spanish-speaking speaker, whom she rated more highly than expected. Besides those ratings, only four others were misfitting, which is comparable to Trainee A. Her unexpectedly high rating of Speaker 3 in Set 2 did not affect her performance at the pass/fail level. Trainee F appears to be generally on target in her ratings and to have mastered the art of rating the TOPT (especially making pass/fail distinctions), even though she came into the experience as a novice.

Trainee G

We know little about Trainee G except that she indicated that she had only minimal experience with the Guidelines. She was not a native speaker of English and did not complete a feedback form. Although she attended a debriefing session, she did not say anything at the meeting, at which there were also two other trainees. She did not pass the criterion, scoring only 60% on the global ratings and 51% on the segment ratings. Her correlations were second to lowest. Most of her ratings were one step away, and 20% of her global ratings and 18% of her segment ratings were two-steps away on the scale. The FACETS analysis reveals that she was the most severe rater, but that she had few misfitting ratings. Thus, she was internally consistent as a rater and consistent with the others in the group. Her tendency toward severity is shown by the three misfitting ratings flagged from the analysis of the 45 segment ratings. In those three instances, she awarded scores that were higher than expected of her. She appears, then, to rate according to the scale, but to have centered herself much lower on the scale than the key. Because of the lack of feedback, it is hard to imagine why this was the case.

Trainee H

Trainee H was the only trainee to indicate that she had no prior knowledge of the ACTFL Guidelines. She spent the second shortest amount of time with the materials. Like Trainee G, she did not pass the criterion measure, though she had respectable correlations with the global ratings (.78) and the segment ratings (.72). Although most of her ratings were one step away

on the scale (67% of the global ratings, 60% of the segment ratings), she had no ratings that were two steps away. The FACETS analyses indicate that she was the second most severe rater, though not nearly as severe as Trainee G. She appeared consistent in her global ratings, but appeared towards awarding unexpected ratings in her segment ratings. Table 10 shows that she had the most number of individual misfitting ratings (discounting Trainee F's problems with one speaker). These ratings show no consistent tendency, indicating some general lack of internal consistency on her part. Her feedback form contained very few comments, though her ratings were generally positive. Her comments at the debriefing session were equally terse. The lack of thoughtful feedback and the short amount of time spent with the materials may indicate that she did not exert herself in the project. Her main comment, perhaps not surprisingly, was that there was "excessive detail" in the materials. In her debriefing session it became clear that she had little experience with the *Guidelines* or with holistic scoring in general. This was evidenced in her stressing the centrality of grammatical accuracy and her desire that more discrete elements be rated. She felt these elements should be clearly stated to examinees so that they know exactly what they will be rated on. To simplify scoring, these elements should also be made explicit to the raters.

Conclusions

Two main conclusions on the efficacy of rater self-training materials can be drawn from this study. First, it appears that it is possible for raters to train themselves in the holistic scoring of a performance assessment. On the criterion used in the operational TOPT program, six of the eight trainees in the study would have passed. The FACETS analysis reveals that all of the trainees were rating in a manner consistent with the ACTFL Guidelines and with the scoring key. Two of the raters, however, were quite severe in their rating. This led them to fail more examinees than the key, and thus to fail the required criterion.

The second conclusion is that the efficacy of rater self-training appears to depend to a large extent on background characteristics and motivation of the trainees. In this study, previous familiarity with the ACTFL *Guidelines* appears to have some effect on the success of the rater self-training. Those trainees who had more previous exposure to the Guidelines performed better on the final calibration task.

A second characteristic may be labeled interest or motivation. Trainee F in this study had very little previous exposure to the *Guidelines*. However, she had tremendous motivation to go through the materials. We sensed this motivation by the amount of time she spent with the materials,

her discussion of the problems she faced in her responsibility in assessing the speaking ability of her students, and in her sense of gratefulness for having participated in this project. This trainee performed extremely well. In contrast, the two trainees who performed worst showed the least interest. One did not bother to complete the feedback form. The other spent only eight hours on the task and wrote only cursory comments that indicated a lack of appreciation of her role in the piloting of the materials.

Language ability may have played a role with one of the two unsuccessful trainees. From the little face-to-face contact we had with Trainee G, we wondered about her level of proficiency in English. At least a college-level proficiency in English would be essential in successfully working through the materials. Also, we wonder if Trainee E's youth and lack of teaching experience played a role in his somewhat inconsistent ratings.

Of course, there may be other background characteristics that could be investigated for their influence on the efficacy of rater self-training. It would be interesting to compare the performance of these trainees with that of individuals who have not worked through the *TOPT Rater Training Kit*. The results would provide further evidence of the utility of the kit as a substitute trainer. Another study could investigate the influence of language proficiency by comparing the relationship between each trainee's proficiency in reading English and speaking Spanish and performance on the calibration test. If a formal measure of the rater's Spanish speaking ability were obtained, it might shed light on whether it would make sense to establish a minimum Spanish proficiency for raters before they begin working on the kit, and what that minimum should be.

In conclusion, we feel we have demonstrated that individuals can train themselves to score a performance assessment such as the TOPT reliably. Self-training should be considered a viable alternative to group training of raters. However, the results of our study remind us of the role of background variables on performance. We feel our experience with these trainees parallels the better-documented evidence on the role of background variables on examinee performance on language tests. Our study has shed some light on the influence of some of these characteristics in the efficacy of rater self-training. We look forward to further research that will help us better understand the role of background variables on success in rating.

References

American Council on the Teaching of Foreign Languages. 1986: *ACTFL proficiency guidelines*. Hastings-on-Hudson, NY: Author.

Baker, E.L. 1992: *The role of domain specifications in improving the technical quality of performance assessments. Project 2.2: alternative approaches to measuring liberal arts subjects: history, geography, and writing*. Final Report for the Office of Educational Research and Improvement. Los Angeles: Center for Research on Evaluation, Standards, and Student Testing. (ERIC Document Reproduction Service No. ED 346 133)

Braungart-Bloom, D.S. 1986, April: *Assessing holistic raters' perceptions of writing qualities: An examination of a hierarchical framework following pre-post training and live readings*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA. (ERIC Document Reproduction Service No. ED 270 477)

Educational Testing Service. 1982: *Speaking Proficiency English Assessment Kit*. Princeton, NJ: Author.

Haven, M., and Anderson, P. 1987: *"Write on, Illinois", a user's guide to scoring student essays*. Springfield, IL: Illinois State Board of Education. (ERIC Document Reproduction Service No. ED 291 103)

Linacre, J.M. 1989: *Many-facet Rasch measurement*. Chicago: MESA Press.

Linacre, J.M., and Wright, B.D. 1993: *A user's guide to facets: Rasch measurement computer program*. Chicago: MESA Press.

Mueller, L.Z., Dabney, V.M., Wilhide, J., and Mappus, L.L. 1984: *Teaching and testing our basic skills objectives (T & T). Writing: grades 4-12*. Columbia, SC: South Carolina State Dept. of Education, Columbia, Office of Research. (ERIC Document Reproduction Service No. ED 253 886)