DOCUMENT RESUME

ED 296 589 FL 017 471

AUTHOR Barnwell, David

TITLE Who Is To Judge How Well Others Speak? An Experiment with the ACTFL/ETS Oral Proficiency Scale.

PUB DATE 12 Oct 86

NOTE 11p.; Paper presented at the Eastern States Conference on Linguistics (Pittsburgh, PA, October 12, 1986).

PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS College Faculty; Higher Education; *Interrater Reliability; Language Teachers; *Language Tests; *Rating Scales; Second Language Instruction; Spanish; *Test Reliability

IDENTIFIERS *ACTFL ETS Language Proficiency Guidelines; *Oral Proficiency Testing; University of Pittsburgh PA

ABSTRACT

A study examined inter-rater reliability on the American Council on the Teaching of Foreign Languages/Educational Testing Service (ACTFL/ETS) oral language proficiency rating scale. Seven raters, all elementary or intermediate college Spanish teachers given only brief formal training in the use of the scale, evaluated recorded interviews with Spanish students at varying proficiency levels. The ratings were paired for comparison and the pairs were categorized as being in perfect agreement, acceptable disagreement (indicating disagreement by one subdivision of the rating scale), or total disagreement. Over 41 percent of the paired ratings were found to be in perfect agreement, almost 45 percent were in acceptable disagreement, and less than 14 percent were in total disagreement. The majority of disagreements were within a particular level rather than across levels. The results suggest a high degree of concordance between raters, with comparatively inexperienced raters reaching acceptable levels of agreement in most cases. The continuing need for native speaker input in the test construction and administration processes is emphasized. (MSE)

WHO IS TO JUDGE HOW WELL OTHERS SPEAK ?
AN EXPERIMENT WITH THE ACTFL/ETS ORAL PROFICIENCY SCALE

David Barnwell

Paper read at the Eastern States Conference on Linguistics, Pittsburgh, Pennsylvania, October 12, 1986.

WHO IS TO JUDGE HOW WELL OTHERS SPEAK ?
AN EXPERIMENT WITH THE ACTFL/ETS ORAL PROFICIENCY SCALE

David Barnwell
Columbia University, New York

1. The past few years have seen increasing interest in the use of proficiency ratings in the measurement of second language proficiency. A large number of workshops and training sessions have been held in a bid to bring the news of proficiency testing and teaching to a wider audience, and a considerable literature has grown up around the topic of proficiency. However, we still lack a convincing body of empirical research on the proficiency measurement now most widely used in the American academic setting, namely, the ACTFL/ETS oral interview.

The comparatively weak empirical base of the ACTFL/ETS scale stems from the very genesis of the oral interview. As is well known, the ACTFL/ETS procedure has its roots in the Foreign Service Institute oral interview. The ACTFL/ETS scale was developed in response to the need to adapt the FSI scale to use at college and high-school level; it is, in Liskin-Gasparro's (1984, 477) words, "an academic version of the government scale". As an offshoot of the FSI oral interview, the ACTFL/ETS interview has perhaps benefitted unduly from the FSI interview's long and successful history. It is striking, for example, that when arguing for the reliability of the oral interview, Liskin-Gasparro cites no study on the ACTFL/ETS scale. Rather does she refer to work on the FSI scale, without speculating on the degree to which FSI findings can be transferred to the ACTFL/ETS scale.

Given the close relationship between the ACTFL/ETS and FSI scales, it is worthwhile briefly to review some of the published research on inter-rater reliability in the FSI and similar oral interviews. In fact, inter-rater reliability has been high at FSI since the earliest days (Rice 1959). According to later studies, the two FSI judges agree to within a (+) of each other in 95% of cases (Clark 1978, 58-69), or at rates ranging from 87% to 92% (Adams 1978). Bachman and Palmer (1981) computed FSI inter-rater reliability at .88.

Of course the raters used at FSI are highly experienced and enjoy many opportunities to practice their trade. It is thus interesting to review the kinds of reliability figures reported for less experienced or non-specialist raters. A number of studies of oral rating have used raters who had not received FSI training, but who had learned to employ the FSI scale through informal training and exposure to FSI materials. Henning (1983) cites an inter-rater reliability figure of .93 on what he calls an "improvised FSI interview". Shohamy reports an inter-rater reliability value of .98 on FSI-type interviews in Hebrew (1983). Graham

(1978) found an inter-rater agreement rate of 93% in oral interview work at the Language Training Mission in Utah in the 1970s. Other studies, such as Clifford (1978) showed that oral interview reliability could be just as high as that found on the more 'objective' MLA Proficiency Tests. Schulz (1977) and Bartz (1979) working with non-interview oral communication tests, found inter-scorer agreement to be very high.

FSI itself has undertaken interesting research in this area. Frith (1979) describes a study in which a group of non-specialists in rating participated in a two-day FSI training course, and afterwards rated a number of sample interviews. They reached a level of agreement of 84% with trained FSI raters over the first eight interviews they rated, and this subsequently rose to 96%, with the benefit of consultation with FSI training personnel. Perhaps more impressive was the performance of a parallel group used in this study. These did not attend any training session at FSI, but merely studied an FSI training kit and listened to sample interviews. Rating independently, they too reached a concordance of 84% with the FSI raters over the first eight interviews, and this improved subsequently to 94% with the aid of consultation with the trainers.

Thus it can be seen that there is a lot of published evidence regarding the the ability of raters, be they formally or informally trained, to operate the FSI rating scales reliably. However, as indicated earlier in this paper, there is nothing like corresponding evidence on the ACTFL/ETS scale.

As is well known, the ACTFL/ETS variant introduces new subdivisions at the lower end of the scale. In this light, it should be remembered that any attempt to fine-tune a rating scale incurs the risk of diminishing reliability. Put simply, the more options that raters are given to choose among, the more chances they have to disagree with each other. Thus it seems clear that the degree to which the ACTFL/ETS scale can be used reliably should not be taken on trust, especially in view of the ambitious goals which have been put forward for proficiency testing (Buck and Hiple 1984). The extent to which raters can make reliable use of the ACTFL/ETS scale is a valid topic for research, and can no longer be taken for granted.

2. An experimental study on the use of the ACTFL/ETS scale.

In view of these considerations, it was decided to conduct an experiment which would yield information on the operation of the ACTFL/ETS scale. The primary data for this study were gathered from a series of oral interviews in Spanish, conducted at the University of Pittsburgh. The students interviewed--the candidates, as they will be called here--were drawn from a wide range of undergraduate classes at the university, ranging from Spanish 1 to Advanced Composition and Conversation classes. Participation was entirely voluntary. Interviews in Spanish with these students were recorded in the Language Laboratory of the University. These interviews lasted an average of about fourteen minutes. Care was taken not to allow the students be identified from the recordings.

The interviews were all carried out by the author of this paper, who closely followed the ACTFL/ETS guidelines for interviewers (Liskin-Gasparro 1982). The researcher had had extensive experience, both as a teacher and tester, in using oral interview techniques, and had participated in several workshops in which the ACTFL/ETS procedures were used and discussed. None of the candidates was a student of the researcher. Twenty-six interviews in all were carried out, with students drawn from a very wide range of classes.

The raters used in this investigation numbered seven, all teachers of elementary or intermediate Spanish classes. Three were native speakers of English, three others were native speakers of Spanish, while the seventh was a Spanish-Portuguese bilingual. None had received formal ACTFL/ETS training. Since the primary focus of this study was on rating behaviour, it was decided to separate the functions of interviewer and rater. Thus one variable could be eliminated from the research, since different interviewers, especially inexperienced ones, might reach different levels of competence. The focus was on how much rating agreement could be achieved, not on how well the interview itself could be conducted by informally trained personnel.

There are probably both advantages and disadvantages to the use of raters who evaluate from recordings, without actually taking part in an oral interview. On the positive side, such raters can devote all their attention to the rating task, without having to worry about how to conduct the interview. As the raters used in this study were inexperienced, it was seen as an advantage that they could concentrate on the rating.

Nevertheless, raters who are not also acting as interviewers do not have the opportunity to probe a candidate with questions and topics that the interviewer might have overlooked. It could be suspected that those who rate from recordings would rate more severely, since they have more time to notice errors, and have no affective interaction with the candidate. However, Lowe's study on the FSI scale showed, on balance, that there appeared to be no fundamental difference between the rating behaviour of these 'third raters' and that of those who actually participated in interviewing candidates (1978). Similarly, Ingram (1982), working on an Australian counterpart of the ACTFL/ETS scale, could find no easily generalizable differences between the rating behavior of those who participated in interviewing and those who did not.

Before undertaking the rating process, the raters used in this investigation were given a brief period of training. They were first issued with descriptions of the aims and procedures of the oral interview technique, using materials such as provided in the ETS Oral Proficiency Testing Manual (Liskin-Gasparro 1982). Having familiarized themselves with the interview format, as well as the detailed operational descriptions of each level, the raters met as a group for one training session of about two hours. At this session they listened to a selection of taped interviews supplied by ETS and FSI. Eight tapes were played, either in full or in part. Since official ETS and FSI ratings were available for each interview, these ratings were used as criterion references in the discussion.

Independent rating began about a week after the training class was held. Three group rating sessions were held, each lasting about two hours. For administrative reasons, seven recorded interviews were played at each session, thus producing a total of twenty-one interviews to be evaluated by the seven raters. The raters listened to each interview and then gave their rating in writing. The raters were not permitted to discuss their ratings with each other or to announce their ratings before they turned them in. The researcher made no ratings, and did not seek to influence the rating performance in any way.

3 Findings

Two types of correlations were established as a means of measuring the degree of agreement between raters. Firstly, raters were administratively divided into all possible pairs yielded by the group as a whole--forming twenty-one pairs. The concordance between the two raters in a pair was then assessed. In addition, the agreement between each individual rater and the group as a whole was measured.The seven raters used in this experiment can be seen, as has been said, as forming twenty-one pairs. Twenty-one interviews were rated, giving a total of 441 paired ratings of interviews. Following Adams' (1978) work with the FSI Scale, the concordance between the two raters in each pair was categorized as follows: Perfect Agreement, Acceptable Disagreement (defined as occurring when raters disagreed by one sub-division of the ACTFL/ETS scale), and Total Disagreement. In the present case, 41.5% of the paired ratings showed Perfect Agreement, 44.9% showed Acceptable Agreement, and 13.6% were in Total Disagreement. In other words, some 86% of paired ratings showed a basic agreement on the part of any two raters. This figure can be compared to those cited by Adams, who reported basic agreement rates of around 90% in her work with the FSI scale. Only very rarely in the present case were disagreements bigger than one subdivision of the scale produced. The majority of disagreements were within a particular level rather than across levels.

More statistically formal correlations between ratings were also computed (1). Of the twenty-one pairs of raters, correlations were in excess of .90 in two cases, between .80 and .90 in fourteen cases, and between .70 and .80 in four cases.The remaining pair inter-correlated at .58. When each individual rating was correlated to the group average, six of the seven yielded values of .90 or over. The seventh rater showed a correlation with the rest of the group of .81. Since the mean of this rater's ratings was somewhat lower than those of her colleagues, it appeared that she had been consistently more severe in her judgments than were the other raters.

Conclusions

The study reported here, informal and exploratory as it was, fills a gap in the literature on proficiency testing. It provides empirically-generated findings on the degree to which the ACTFL/ETS scale can be reliably used by informally trained personnel. The findings show a high degree of concordance among raters. Comparatively inexperienced raters reached acceptable levels of agreement in the great majority of cases. Given some minimal training in order to familiarize themselves with the scale to be used, it appears that people can judge how well others speak. This is not a revolutionary discovery. Indeed, Ingram, in his quite large-scale research in Australia, reported similar findings, leading him to express the belief that "the instrument may be used even by lay persons without special training". As was seen in our earlier review of findings on the FSI scale, there is a lot of evidence that raters do not need very protracted training periods before they can begin to rate reliably. The present study now confirms this observation in the case of the ACTFL/ETS scale. It seems that the greater sensitivity of this scale, as compared to that of the FSI scale, does not compromise raters' ability to agree on the level to which a candidate should be assigned.

Recent work on language proficiency has tended to stress the notion of proficiency *for what purpose*, the *Function* aspect of the ACTFL/ETS trisection. We now see references to the need to test the specific proficiency of groups such as bilingual teachers or medical or business personnel (Buck and Hiple 1984). It will also be remembered that the FSI test itself grew out of the need for a test of a candidate's ability to function in a certain (diplomatic) milieu. Were the interest in special types of proficiency to be maintained and developed, we would experience a need for more and more rating personnel who were drawn from the general population, from outside the confines of academic or measurement institutions. Almost by definition these people would have little background in language teaching or testing. Bodies which used these raters would consequently be faced with the choice of how much and what kind of training to give before accepting their judgments as reliable.

Even if specialized forms of proficiency testing never evolve, we are still faced with the need to incorporate more native speakers into the development of the generic test. There were several references to "the native speaker" in the 1982 ACTFL/ETS oral proficiency scale, perhaps less so in the 1986 update. it is far from clear how many non-academic native speakers were involved in drawing up the scales. This raises one of the central paradoxes of foreign language testing, one which has been in evidence since the days of Audiolingualism and even before. For the fact is that the more "sophisticated" or "technical" are our language testing procedures, the more they diverge from the kinds of judgments that native speakers really make. This is true of the ACTFL/ETS oral proficiency interview. The training and certification process for interviewers is at present quite costly in terms of time and money, and we really have no information as to how or why it was decided that such expense was necessary. How much training does the average native speaker get ? Thus, if the ACTFL interviewer decides one speaks Spanish badly but the barman in Madrid appears to think one speaks rather well, who is to be believed ? Who is the expert ? And what does it mean to be an expert in language testing ? One does not learn a language in order to talk to one's teacher, nor to one's ACTFL interviewer. The domain of proficiency is outside the classroom, indeed outside the testing milieu itself. We cannot continue to claim to gauge proficiency without seeking input from the "naive" native speakers with whom we hope our students will one day interact. The optimal quantity and quality of training needed by foreign language proficiency testers is a question that remains amenable to empirical investigation. For now, all we can do is reiterate that this study, like many others, has shown that long periods of training are not needed to teach people to judge how well other people speak. Other *people* should therefore be the judges of how well we speak, not other language teachers or testers.

FOOTNOTE

[1] In the present study, numerical eqivalences to the verbal ratings were assigned on the basis of allotting a value of zero to the lowest point on the scale, Novice Low, and proceeding through increments of one up to the highest point, Superior, which received a value of 8. Correlations were then calculated betweeen the ratings (Guilford 1954, 395-7). The whole question of how verbal ratings should be converted for statistical analysis has been passed over in the literature on proficiency testing. The statistical treatment of non-parametric scores has implications for the interpretation of reliability data. Indeed it is a criticism of FSI interview research that no one has pointed out that to some extent high reliability figures may be an artifact of researchers' failure to distinguish intra-boundary from inter-boundary disagreements. However, a justification of the procedure used in the present study may be found in Ingram (1982).

Table 1: Correlations between Raters

	F	JJ	AW	MS	GA	MC	PI	RG
F	1	.951	.967	.926	.929	.951	.900	.811
JJ	.951	1	.902	.867	.854	.873	.838	.815
AW	.967	.902	1	.889	.884	.957	.812	.775
MS	.926	.867	.889	1	.850	.854	.748	.847
GA	.929	.854	.884	.850	1	.826	.800	.777
MC	.951	.873	.957	.854	.826	1	.861	.723
PI	.900	.838	.812	.748	.800	.861	1	.584
RG	.811	.815	.775	.847	.777	.723	.584	1

Key: F = Overall averaged score for entire group

Other letters represent initials of raters.

Table 2. Inter-rater pair agreements for each interview
(21 possible pairs)

Interview	P.A.	A.D.	T.D.
A	4	11	6
B	7	10	4
C	11	10	0
D	7	12	2
E	5	10	6
F	6	12	3
G	21	0	0
H	7	12	2
I	10	10	1
J	10	6	5
K	4	11	6
L	7	10	4
M	7	8	6
N	7	12	2
O	7	10	4
P	7	12	2
Q	7	12	2
R	21	0	0
S	10	6	5
T	11	10	0
U	7	10	4
Total	183	194	64

Key: Column 1 gives interview identifying numbers, listed in order in which they were rated.
P.A. = Perfect Agreement
A.D. = Acceptable Disagreement
T.D. = Total Disagreement

REFERENCES

Adams, Marianne. 1978. Measuring foreign language proficiency: a study of agreement among raters. In Clark, Direct Testing, 129-49.

Bachman, Lyle, and Adrian Palmer. 1981. The construct validation of the FSI Oral Interview. Language Learning, 31, 1, 67-86.

Bartz, Walter. 1979. Testing oral communication in the foreign language classroom. Arlington, Va: Center for Applied Linguistics. ED 176590.

Buck, Kathryn, and David V Hiple. 1984. The rationale for defining and measuring foreign language proficiency programs for business. Foreign Language Annals, 17, 5, 525-28.

Clark, John L. 1978. Direct testing of speaking proficiency: theory and application. Princeton: ETS.

Clifford, Ray. 1978. Reliability and validity of language aspects contributing to oral proficiency of prospective teachers of German. In Clark, Direct Testing, 193-210.

Frith, James R. 1979. Testing the FSI Testing Kit. ADFL Bulletin, 11, 2, 12-14.

Graham, Stephen. 1978. Using the FSI Interview as a diagnostic evaluation instrument. In Clark, Direct Testing, 31-9.

Guilford, J.P. 1954. Psychometric methods. New York: McGraw-Hill.

Henning, Grant. 1983. Oral proficiency testing: comparative validities of interview, imitation and completion models. Language Learning, 33, 3, 315-32.

Ingram, D.E. 1982. Report on the formal trialling of the Australian Second Language Proficiency Ratings. Canberra: Australian Dept. of Immigration. Eric ED 230025.

Liskin-Gasparro, Judith E. 1982. Foreign language oral proficiency assessment manual. Princeton: ETS.

Liskin-Gasparro, Judith E. 1984. "The ACTFL Proficiency Guidelines: Gateway to Testing and Curriculum. Foreign Language Annals, 17, 5.

Lowe, Pardee. 1978. Third Rating of FSI Interviews. In Clark "Direct Testing" pp. 257-70.

Rice, F.A. FSI Tests.1959. Linguistic Reporter, 1: 4.

Schulz, Renate. 1977. Discrete-point versus simulated communication testing in foreign languages. Modern Language Journal, 32: 33-47.

Shohamy, Elana. 1983. Rater reliability of the Oral Interview speaking test. Foreign Language Annals, 16, 3, 219-22.