DOCUMENT RESUME

ED 388 039 FL 023 277

AUTHOR Thompson, Irene
TITLE A Study of Inter-rater Reliability of the ACTFL Oral Proficiency Interview in Five European Languages: Data from ESL, French, German, Russian, and Spanish.
PUB DATE 95
NOTE 26p.; Paper presented at the Annual Meeting of the American Association of Applied Linguistics (Long Beach, CA, March 25-28, 1995).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS English (Second Language); Evaluators; French; German; Interpersonal Communication; *Interrater Reliability; *Language Proficiency; *Language Tests; *Oral Language; Questioning Techniques; Russian; Spanish; Testing
IDENTIFIERS *ACTFL Oral Proficiency Interview

ABSTRACT

This report addresses the reliability of the American Council on the Teaching of Foreign Languages (ACTFL) Oral Proficiency Interview (OPI), not as a measure of speaking ability, but rather as practiced by testers trained by the ACTFL, such as by the Interagency Language Roundtable (ILR), in English as a Second Language (ESL), French, German, Russian, and Spanish. Inter-rater consistency was measured by Pearson product-moment correlation coefficients and by a modified Cohen's kappa. Pearson coefficients were highly significant and remarkably similar in all five languages; Cohen's kappa results were also significant. Study results also confirm that interaction with the interviewee presents a source of variance in the assessment of speaking ability and that some levels of speech performance are simply harder to rate than others. Findings suggest that similarities and differences existed in the five languages that were difficult to explain and that inter-rater disagreement was very frequent and dependent on the level. It is concluded that a large and heterogeneous group of ACTFL-trained oral proficiency interviewers can apply the OPI in the five languages tested with a fairly high degree of consistency. (Contains 20 references.) (NAV)

A Study of Inter-rater Reliability of the ACTFL Oral Proficiency Interview in Five European Languages: Data from ESL, French, German, Russian, and Spanish

Irene Thompson
The George Washington University

A Study of Inter-rater Reliability of the ACTFL Oral Proficiency Interview in Five European Languages: Data from ESL, French, German, Russian, and Spanish

Irene Thompson
The George Washington University

INTRODUCTION

The widespread use of the Oral Proficiency Interview (OPI) throughout the government, the academic community, and increasingly the business world, calls for an extensive program of research concerning theoretical and practical issues associated with the assessment of speaking proficiency in general, and the use of the OPI in particular. Its growing popularity notwithstanding, the OPI has yet to generate a solid body of empirical research regarding its validity and reliability (Bachman, 1988; Bachman and Clark, 1987; Clark and Lett, 1988; Clark and Clifford, 1988; Valdman, 1988). The purpose of this article is not to address the validity of the OPI as a measure of speaking ability, but to expand our knowledge about the reliability of the OPI as practiced by testers trained by the American Council on the Teaching of Foreign Languages (ACTFL).

REVIEW OF THE LITERATURE

Reliability of the OPI as practiced by the Interagency Language Roundtable (ILR)

In the ILR version of the OPI, two testers (the examiner who is in charge of the test, and a native speaking interviewer) work together to elicit a sample of examinee's speech for subsequent assessment. The examiner's role is to make sure that the interviewer elicits a ratable sample. After the interview, the two testers independently rate the examinee's performance. If their opinions differ by one step,[1] the lower of the two ratings is awarded. If their ratings differ by more than one step, they submit the tape and their

[1] One-step differences involve adjacent ratings, e.g., 1 and 1+. The difference between 1 and 2 is considered a two-step disagreement.

ratings for arbitration by a third rater. ILR testers regularly give interviews and have an opportunity to compare and discuss them with each other.

Adams (1978) studied the reliability of ILR interviews by having 4 German, 6 French and 11 Spanish testers from the Foreign Service Institute rate 50 prerecorded interviews in the three languages. She found an average correlation of 0.91 among examiners (presumably, the better trained and more experienced testers). She also found that French and Spanish examiners agreed among themselves better than French and Spanish interviewers but that the opposite was true in German. Disagreements in all languages mostly involved one-step differences. The percentage of disagreements that crossed boundaries between main levels was not reported.

In a study by Clark (1986), 115 examinees in French and German were interviewed by two-person teams composed of hand picked testers from the Defense Language Institute, the Central Intelligence Agency, and the Foreign Service Institute. Stansfield and Kenyon (1992) used Clark's raw scores to calculate test-retest reliabilities which had a range of 0.90-0.92 in French, and 0.84-0.87 in German. The variation in the assignment of proficiency ratings by interviewers from the three agencies suggested that tester groups may develop their own idiosyncratic testing and tester training procedures.

Reliability of the ACTFL oral proficiency interview

The ACTFL version of the OPI differs from the ILR version in three significant ways. In the first place, the ACTFL scale condenses ILR levels 3, 3+, 4, 4+, and 5 into a single category of "Superior," but makes more distinctions at the lower end of the scale (ILR level 0 is broken down into Novice Low and Novice Mid, ILR level 1 is divided into Intermediate Low and Intermediate Mid). In the second place, the test is administered by only one interviewer, who conducts and records the interview, and then rates it from tape immediately after the test or after some delay. Interviews conducted for "official" purposes are independently rated by a second, and in cases of significant disagreement, by a third rater. In the third place, ACTFL testers are scattered around the country, and

unlike ILR testers generally have few opportunities to maintain calibration by comparing and discuss their ratings with each other.

Magnan (1987) examined inter-rater reliability of an experienced ACTFL tester in French and 14 trainees who attended an intensive 4-day workshop, and then conducted 8 interviews on their own in Phase I, and 7 additional interviews in Phase II of their training. The interviews were recorded and submitted to the trainer who checked 6 of the 15 interviews. In both phases, Pearson's r between trainer and trainee ratings was 0.94, and Cohen's kappa was 0.53 for Phase I and 0.55 for Phase II. Magnan also found that the disagreements between trainer and trainee ratings were mostly confined to one step within the same main proficiency level.

In another study, Magnan (1986) interviewed 40 students of French. The taped interviews were then independently rated by two other ACTFL-certified testers from the Educational Testing Service (ETS). Cohen's kappa between the two raters was 0.72. All discrepancies in rating were one step apart within the same main level. The greater inter-rater reliability in this study compared to the 1987 study could have been due either to the greater experience of the ETS testers, and to the fact that both of them assigned ratings after listening to interviews that were conducted by someone else, whereas in the 1987 study one of the ratings came from the interviewer.

Based on 119 double-rated ACTFL interviews, Dandonoli and Henning (1990) reported alpha inter-rater reliabilities for mean of two raters that ranged from 0.85 to 0.98 in ESL, and between 0.89 and 0.97 in French. As in the Magnan 1987 study, one of the ratings was assigned by the interviewer. Dandonoli and Henning did not report what percentage of rating discrepancies was one step apart, two steps apart, etc.

Finally, Stansfield and Kenyon (1992) reported Pearson's inter-rater reliabilities of 0.81 in Chinese, 0.94 in Portuguese, 0.97 in Indonesian, and 0.97-0.99 in Hebrew for two

raters listening to prerecorded interviews. The number of interviews, or the nature of the rating discrepancies were not reported.

These small-scale studies of the ACTFL version of the OPI demonstrate high inter-rater reliabilities that are comparable to those of ILR testers. However, these studies involve very few testers and are based on a small number of interviews. With a growing number of ACTFL-certified testers in a number of different languages, we need to know whether the same high rate of inter-rater agreement holds for a larger and more representative sample of testers.

Moreover, the studies surveyed differ with respect to the way the two ratings were obtained. For instance, in Magnan (1986), and Stansfield and Kenyon (1992) both raters scored prerecorded interviews conducted by someone else. On the other hand, in Magnan (1987) and in Dandonoli and Henning (1990), one of the ratings was assigned by the interviewer. We need to investigate if the conditions under which ratings are obtained affect ratings. Clark and Lett (1988) suggested that audio-tape-based second ratings may be systematically biased by comparison with the original ratings in the direction of lower scores due to the fact that linguistic inaccuracies in the interviewees' speech may become more salient during an audio replay than during the real-time interview. We do not have any empirical data as to whether such a bias exists among ACTFL testers.

Hiple (personal communication) and Reed (personal communication) suggested that some levels are inherently more difficult to rate than others. Many testers will probably agree that one of the most difficult distinctions to make is that between High and the next higher level, e.g., Advanced High and Superior. However, there is no empirical evidence to support the notion that inter-rater agreement varies from level to level. Nor do we know whether testers in different languages have different patterns of agreement at different levels. For instance, do testers in a variety of languages find the Advanced High-Superior distinction the most troublesome, or does each language have its own problem level?

Magnan reported that disagreements between two experienced French raters scoring from a tape were always confined to the same proficiency level but we do not know whether this would also be true of a larger sample of raters and languages other than French

RESEARCH QUESTIONS

The present study was designed to consider the following questions:

1. What is the inter-rater reliability of ACTFL-certified testers in five European languages: Spanish, French, Russian, ESL, and German?
2. What is the relationship between interviewer-assigned ratings and second ratings based on audio replay of the interviews?
3. Does inter-rater reliability vary as a function of proficiency level?
4. Do different languages exhibit different patterns of inter-rater agreement across levels?
5. Are inter-rater disagreements confined mostly to the same main proficiency level?

SUBJECTS AND METHODOLOGY

This study is based on interviews in ESL, French, German, Russian, and Spanish made available by Language Testing International (LTI).[2] Out of a total of 822 interviews, 27 (3.33%) were unratable[3] and were excluded from the final analysis which is based on 795 ratable interviews. All interviews were conducted and rated by ACTFL-certified testers. The taped interviews were then independently second-rated by other ACTFL-certified testers. Table I gives the number of testers who conducted the interviews in each language. Testers came from many different institutions, varied in testing experience, and included both native and nonnative speakers.

[2] Unfortunately, Chinese and Japanese data had to be excluded from this study because of the small size of the samples.

[3] An interview was considered unratable if the second rater was unable to assign a rating.

TABLE I
Distribution of ACTFL-certified testers

Spanish	French	Russian	ESL	German	Total
80	58	11	14	11	174

Data for French, Spanish, English, and German are based on telephone interviews conducted by LTI. Two-thirds of the Russian sample are based on face-to-face interviews of Russian summer school students at Middlebury College and at the University of Iowa,[4] and one-third on telephone interviews conducted by LTI. The interviewees present a very broad spectrum of learners in terms of age, education, amount of exposure to the language, and type of language learning experience. Table II shows the distribution of ratings assigned by interviewers (first ratings).

TABLE II
Distribution of first ratings

	IM	IH	A	AH	S	Ratable	Unratable	Total
Spanish	29 6.58%	52 11.79%	99 22.45%	93 21.09%	168 38.10%	441	20 4.34%	461
French	15 9.09%	14 8.48%	40 24.24%	34 20.61%	62 37.58%	165	6 3.51%	171
Russian	19 23.46%	17 20.99%	11 13.58%	13 16.05%	21 25.93%	81	0 0.00%	81
ESL	8 13.11%	10 16.39%	23 37.70%	6 9.84%	14 22.95%	61	1 1.61%	62
German	13 27.66%	7 14.89%	10 21.28%	8 17.02%	9 19.15%	47	0 0.00%	47
Total	84 10.57%	100 12.58%	183 23.02%	154 19.37%	274 34.47%	795	27 3.33%	822 100.00%

[4] These data were collected for a validation study of the Russian Guidelines under a grant from the US Department of Education to the Educational Testing Service and ACTFL. Students in these summer programs came from programs all over the country.

87 interviews (10.94% of the total sample) were also rated by third raters. These were mostly cases when the ratings assigned by the interviewer and by the second rater were more than one step apart, or when the second rater found the interview particularly difficult to rate.

Before proceeding with a discussion of the results, several caveats are in order. In the first place, the size of the samples varied significantly from language to language, ranging from a high of 441 in Spanish to a low of 47 interviews in German. Secondly, the number of interviews at levels below the Intermediate Mid was too small in most of these languages to yield reliable statistics, therefore interviews at the Novice Low, Novice Mid, Novice High, and Intermediate Low levels had to be excluded from final analysis. As a result, only interviews at the Intermediate Mid, Intermediate High, Advanced, Advanced High, and Superior levels were considered. Thirdly, the number of interviews differed from level to level. In general, there were more interviews at higher than at lower proficiency levels because LTI clients are primarily interested in persons with 'usable' levels of language ability. Fourthly, this study is based in great part on interviews conducted on the telephone, and we simply do not know if ratings based on telephone interviews are different from those based on face-to-face tests.

RESULTS

In order to make the results comparable with other studies, inter-rater consistency was measured by two statistics. In the first place, Pearson product-moment correlation coefficients were computed between all pairs of raters[5]. These coefficients are given in column 1 of Table III below. They were highly significant and remarkably similar in all five languages. The estimated variance (square of correlation), given in column 2, accounted for by speaking ability, was between 0.87 and 0.70. These results are surprisingly robust even though this is not what North (1993:43) called a "lab" study of a

[5] The following numerical scores were assigned to the ACTFL levels: Novice Low=0.1, Novice Mid=0.3, Novice High=0.8, Intermediate Low=1.1, Intermediate Mid=1.3, Intermediate High=1.8, Advanced=2.3, Advanced High=2.8, Superior=3.3.

hand-picked group of experienced testers where inter-rater reliabilities are expected to be high.

TABLE III:
Product-moment correlation coefficients between first and second ratings

	r	R^2	df
Spanish	0.846*	0.781	439
French	0.873*	0.760	163
Russian	0.897*	0.870	79
ESL	0.839*	0.704	59
German	0.885**	0.783	45

* $p < 0.0001$

These reliability estimates are high because with only five nominal categories, the possibility of inter-rater agreement due to chance is not taken into account. Therefore, a modified Cohen's kappa (Fleiss, 1971) was also computed for each language to provide a more conservative measure of inter-rater consistency. Cohen's kappa is more appropriate for this type of data for the following reasons: (1) it is designed to measure the degree of agreement between two raters who independently rate a sample of subjects on a nominal scale; (2) it incorporates a correction for the extent of agreement expected by chance; (3) it measures agreement between a pair of raters where each subject is rated on a nominal scale, but where the raters rating one subject are not necessarily the same as those rating another one. Thus, a modified Cohen's kappa gives a more conservative estimate of inter-rater agreement than Pearson's r. Nevertheless, the results were also significant. The kappas in Table IV indicate that if a first rating is "x," the chances of a second rating being the same are over four in one in Spanish, Russian, and ESL, and over five in one in French and German.

TABLE IV
Inter-rater reliability as measured by modified Cohen's kappa

Spanish	French	Russian	English	German
0.474*	0.531*	0.443*	0.469*	0.516*

Table V shows the frequency of inter-rater agreement at different levels collapsed across languages in terms of both raw scores and percentages. Overall, inter-rater reliability was greatest at the Superior level, followed by Intermediate Mid, Advanced, Intermediate High, and Advanced High levels.

TABLE V
Relationship between first and second ratings collapsed across languages
(cells showing agreement between raters are highlighted)

Rater 1	Rater 2						
	Below Int Mid	**Int Mid**	**Int High**	**Advanced**	**Advanced High**	**Superior**	**Total**
Int Mid	12 14.29%	57 67.86%	13 15.48%	2 2.38%			84
Int High	2 2.00%	21 21.00%	56 56.00%	21 21.00%			100
Advanced		9 4.92%	41 22.40%	106 57.92%	24 13.11%	3 1.64%	183
Advanced High			9 5.84%	62 40.26%	60 38.96%	23 14.94%	154
Superior				22 8.03%	48 17.52%	204 74.45%	274

Table VI shows the frequency of agreement between raters for each language separately.

TABLE VI
Relationship between first and second ratings by language

Spanish

Rater 1	Rater 2						
	Below Int Mid	**Int Mid**	**Int High**	**Advanced**	**Advanced High**	**Superior**	**Total**
Int Mid	6 20.69%	20 68.97%	2 6.90%	1 3.45%			29
Int High	1 1.92%	10 19.23%	28 53.85%	13 25.00%			52
Advanced		6 6.06%	23 23.23%	57 57.58%	12 12.12%	1 1.01%	99
Advanced High			4 4.30%	37 39.78%	39 41.94%	13 13.98%	93
Superior				16 9.52%	31 18.45%	121 72.02%	168

French

Rater 1	Rater 2						
	Below Int Mid	Int Mid	Int High	Advanced	Advanced High	Superior	Total
Int Mid	2 13.33%	8 53.33%	4 26.67%	1 6.67%			15
Int High		2 14.29%	11 78.57%	1 7.14%			14
Advanced		1 2.50%	7 17.50%	26 65.00%	5 12.50%	1 2.50%	40
Advanced High			1 2.94%	16 47.06%	11 32.35%	6 17.65%	34
Superior				2 3.23%	9 14.52%	51 82.26%	165

Russian

Rater 1	Rater 2						
	Below Int Mid	Int Mid	Int High	Advanced	Advanced High	Superior	Total
Int Mid	4 21.05%	12 63.16%	3 15.79%				19
Int High	1 5.88%	6 35.29%	9 52.94%	1 5.88%			17
Advanced		1 9.09%	6 54.55%	4 36.36%			11
Advanced High			3 23.08%	5 38.46%	5 35.46%		13
Superior				2 9.52%	3 14.29%	16 76.19%	21

ESL

Rater 1	Rater 2					
	Int Mid	Int High	Advanced	Advanced High	Superior	Total
Int Mid	6 75.00%	2 25.00%				8
Int High	3 30.00%	5 50.00%	2 20.00%			10
Advanced	1 4.35%	4 17.39%	12 52.17%	6 26.09%		23
Advanced High			3 50.00%	3 50.00%	0 0.00%	6
Superior			2 14.29%	2 14.29%	10 71.43%	14

GERMAN

Rater 1	Rater 2					
	Int Mid	**Int High**	**Advanced**	**Advanced High**	**Superior**	**Total**
Int Mid	11 84.62%	2 15.38%				13
Int High		3 42.86%	4 57.14%			7
Advanced		1 10.00%	7 70.00%	1 10.00%	1 10.00%	10
Advanced High		1 12.50%	1 12.50%	2 25.00%	4 50.00%	8
Superior				3 33.33%	6 66.67%	9

In Spanish, French, and Russian, inter-rater concurrence peaked at the Superior, whereas in English and German, it was highest at the Intermediate Mid level. In Russian, inter-rater agreement was lower at the Advanced level than in any of the other languages, and in French, it was lower at the Intermediate Mid level than in the other four languages.

To examine the direction of the bias in second ratings, the number of second ratings that were lower and those that were higher than the interviewer-assigned ratings was computed for each language. The results are presented in Table VII. Spanish, French, Russian, and ESL second raters assigned ratings that were generally lower than those given by interviewers. Only in German was the opposite true. However, for the five languages combined, almost three times as many second ratings were lower than first ratings as those that were higher. The difference in the frequency of disagreements across languages was significant (chi-square 21.563, df 4, $p<0.0001$).

TABLE VII
Direction of disagreements between first and second raters

	Rater 2 lower than rater 1	Rater 2 higher than rater 1
Spanish	143 75.57%	43 24.43%
French	39 67.24%	19 32.76%
Russian	31 88.57%	4 11.43%
ESL	15 60.00%	10 40.00%
German	6 33.33%	12 66.67%
Total	224 71.79%	88 28.21%

Next, the distance between discrepant ratings was measured in terms of steps. Adjacent ratings are one step apart, e.g., Intermediate Mid and Intermediate High, or Intermediate High and Advanced, whereas Intermediate Mid and Advanced are two steps apart. Table VIII shows that an overwhelming majority of rating disagreements were one step apart. There was no difference in the proportion of one-step to two-step discrepancies due to language (chi-square 2.097, df 4, p=0.718). There were no three-step disagreements.

TABLE VIII
Rating disagreements in terms of steps

	One step disagreements	Two step disagreements
Spanish	141 83.43%	28 16.57%
French	50 89.29%	6 10.71%
Russian	24 80.00%	6 20.00%
ESL	22 88.00%	3 12.00%
German	16 88.89%	2 1.11%
Total	253 84.90%	45 15.10%

In order to take the analysis one step further, all pairs of discrepant ratings were broken down into those that stayed within the same main level (minor borders) and those that crossed borders between main levels (major borders). Table IX shows the breakdown of border crossings by language. The percentage of rating pairs that crossed major borders was quite similar at the Intermediate High, Advanced, Advanced High, and Superior levels in all five languages. All disagreements crossed a major border at the Superior level, while there were very few instances of rating disagreements that involved crossing of major borders at the Intermediate Mid level. Overall, more disagreements involved major border crossing. The difference in frequency of minor/major border crossing due to language approached significance (chi-square 9.242, df 4, p=0.055). Spanish and German rating pairs crossed major borders more frequently than French, Russian and ESL pairs.

TABLE IX
Frequency of minor and major border crossings

	Minor borders	Major borders
Spanish	68 38.64%	108 61.36%
French	29 50.00%	29 50.00%
Russian	19 54.29%	16 45.71%
ESL	14 56.00%	11 44.00%
German	4 22.22%	14 77.78%
Total	134 42.95%	178 57.05%

Table X shows the distribution of third ratings. The "Neither" column indicates the number of third ratings which were neither like the first, nor like the second rating. German topped the list in the percent of ratings that had to be submitted to arbitration by a third rater.

TABLE X
Distribution of third ratings

Language	Rater 3=Rater 1	Rater 3= Rater 2	Neither	Total 3rd ratings % of total sample
Spanish	11 28.20%	17 43.60%	11 28.20%	39 8.84%
French	9 52.94%	4 23.53%	4 23.53%	17 9.94%
Russian	1 9.09%	9 81.82%	1 9.09%	11 13.58%
ESL	4 57.14%	2 28.57%	1 14.29%	7 11.29%
German	6 46.15%	6 46.15%	1 7.70%	13 27.66%
Total	31 35.63%	38 43.68%	18 20.69%	87 100.00%

Overall, a higher percentage of third ratings agreed with second ratings than with first ratings, but this percentage varied from language to language. In Russian, an overwhelming majority of third ratings was identical to second ratings; in Spanish, third-raters were more likely to agree with second raters; in German, third raters were equally likely to agree with first and with second raters; and in ESL, third raters tended to side with first raters. Across languages, almost 21% of third ratings agreed neither with the first, nor with the second rating. The percentage of ratings in the "Neither" category ranged from a high of 23.5% in Spanish to a low of 8% in German.

DISCUSSION

This study provides some tentative answers to the research questions posed earlier.

1. What is the inter-rater reliability of ACTFL-certified testers in five European languages? Inter-rater reliability indices between first- and second ratings in Spanish, French, Russian, ESL, and German were significant both when Pearson's r and Cohen's kappa were used. Although Pearson's r was somewhat lower than reported by Adams (1978), it must be kept in mind that there are some important differences between these two studies.

In the first place, Adams obtained ratings from a relatively small group of hand-picked testers who work and test in close contact with each other. By comparison, this study involved a large group of testers who work in isolation. Unlike ILR testers all of whom test regularly, ACTFL testers vary in amount of testing experience. In addition, unlike ILR interviewers, who are native speakers of the language they test, ACTFL testers range in their speaking proficiency from native to baseline Superior—the minimum level required for certification. Finally, Adams based her study on ratings assigned by testers who merely listened to prerecorded interviews, whereas in this study, one of the ratings was assigned by the person who actually conducted the interview.

The inter-rater reliabilities in this study are also lower than those reported by Magnan (1986) for ACTFL interviews in French, and Dandonoli and Henning (1990) for ACTFL interviews in French and ESL. Magnan's study involved only two experienced ETS raters both of whom scored the interviews from listening to tapes. Dandonoli and Henning used the same methodology as this study, however, all their interviewers/raters were highly experienced, and their number was very small. On the other hand, Magnan's (1987) study of inter-rater agreement between trainees, who conducted the interviews, and trainer, who listened to these interviews on tape, obtained results that are almost identical with the French data in this study.[6]

2. What is the relationship between first and second ratings? The present study lends support to the hypothesis that interaction with the interviewee, whether face-to-face or by telephone, as opposed to listening to an audio replay of the interaction, presents a source of variance in the assessment of speaking ability. Thus, when investigating inter-rater reliability, we need to keep in mind the conditions under which the ratings were obtained. When second raters disagreed with interviewer-assigned ratings, they were three times as

[6] Magnan (1987) reported Cohen's kappa of 053-0.55; a modified Cohen's kappa for French in this study was 0531. This means that in both studies the chances of two raters assigning the same rating are over five to one.

likely to assign scores that were lower rather than higher. This finding is at variance with Lowe (1978) who reported that ratings based on audio replay of ILR interviews in Spanish, French, Russian, and German were significantly higher than the original scores. However, Lowe's study was based on third ratings of only those interviews, which resulted in test scores that were disputed by examinees presumably because they thought they deserved a higher, and not a lower score.

Theoretically, third ratings should be more like second ratings because both second and third raters score interviews from audio replay. Third ratings in this study were, indeed, more likely to agree with second ratings than with first ratings but the tendency varied from language to language. It should be remembered that third ratings are often called for when there is substantial disagreement between the first and second ratings. Such disagreements arise when there are problems with elicitation procedures, when examinees have an unusual profile, or when they fail to cooperate with the interviewer. On the whole, third-rated interviews are probably not representative of the sample as a whole.

How can we explain the fairly systematic difference between first and second ratings? Are certain aspects of speaking performance more salient during audio playback, while others are more prominent during interaction with the examinee? Unfortunately, the literature does not provide us with any clear answers. On the one hand, Halleck (1992) reported that ACTFL raters justified their ratings primarily in terms of functions and context. In his study, of the 180 reasons cited in support of ratings at the Intermediate and Advanced levels, 169 related to the speakers' communicative competency, and only 11 had to do with grammatical accuracy. On the other hand, Magnan (1988) found a linear relationship between grammatical accuracy and French proficiency scores assigned by two independent second raters for levels ranging from Intermediate Low to Advanced High. Raffaldini (1988) suggested that OPI ratings reflect primarily linguistic and discourse competence, thus, the bias towards assigning lower scores may be explained by the fact that second raters, removed from contact with examinees, focus their attention on grammatical and discourse aspects of the examinees' performance. This is exactly the

point made by Clark and Lett (1988) who suggested that interviewers may be swayed by functional and interpersonal aspects of the interviewees' performance, since they have less time to focus on the linguistic aspects of the candidates' speech. As a result of the difference between the two rating environments, judgments based on audiotape playback alone may tend to be lower than those assigned by interviewers.

3. Is inter-rater reliability a function of proficiency level? The results of this study provide support for the hypothesis that some levels of speech performance are simply harder to rate than others. Most testers will agree that the "High" levels are the most troublesome because speech performance at these levels is characterized by its "almost the next level" quality. Thus, Intermediate High is an inconsistent Advanced, and an Advanced High is an inconsistent Superior. The absence of quantifiable ways to estimate this "almostness" leaves plenty of room for raters to disagree, particularly in the case of imperfectly elicited samples. The present study, in fact, shows that Advanced High had by far the lowest interrater reliability.

The highest frequency of inter-rater agreement occurred at the Superior level. This may be explained by the fact that on the ACTFL scale this level encompasses a broad range of performances ranging from baseline Superior to native-like command of the language. In contrast to the ACTFL scale, the Interagency Language Roundtable (ILR) scale breaks up this range into five steps, namely 3, 3+, 4, 4+, and 5. It is possible that agreement would have been lower if the Superior interviews were rated on the ILR scale, following Clark (1988) who computed inter-rater reliability for scoring interviews in Chinese on a 13-point scale (which included levels 3, 3+, 4, 4+, and 5). Another possibility is that the samples included many near-native and native speakers who are easy to rate on the ACTFL scale.

4. Do different languages exhibit different patterns of inter-rater agreement across levels? The five languages exhibited both similarities and differences which are difficult to explain.

Intermediate Mid. The highest percentage of identical ratings occurred in German, and the lowest in French. In ESL and German all second ratings were higher than first ratings; in French there were twice as many higher second ratings than lower ones; in Spanish there was a tendency to rate lower; and in Russian, the number of lower and higher second ratings was comparable.

Intermediate High. Inter-rater agreement was highest in French and lowest in German. The pattern of second rater disagreements differed from language to language. Spanish and ESL raters assigned both lower and higher scores. Russian raters tended to rate lower; in German, all disagreements were biased in the direction of higher scores.

Advanced. Spanish and Russian raters generally rated lower, while French, ESL, and German raters assigned about an equal number of higher and lower ratings.

Advanced High. Spanish, French, Russian, and ESL second ratings were generally biased in the direction of lower scores. The opposite was true in German.

Superior. There were few rating disagreements at this level. In cases of disagreement, all second ratings were lower.

5. Are inter-rater disagreements mostly confined to the same main proficiency level? Magnan (1986, 1987) reported that cases of inter-rater disagreement were mostly confined within the same main proficiency level, however, the present data showed that crossing of major borders was not only very frequent, but also dependent on the level.

Intermediate Mid. Because of the placement of this level on the ACTFL scale, one-step disagreements in either direction kept them confined to the same main level.

Intermediate High. The picture varied from language to language. In Spanish and English, the ratio of major/minor border crossings was about the same. Because of tendency on the part of second raters to score lower, rating disagreements in French and Russian were confined to the Intermediate level. Since German second raters tended to assign higher scores, their disagreements tended to cross major boundaries.

Advanced.. Consistent with the tendency to rate lower after audio replay, rating disagreements at this level crossed major borders in all five languages at a ratio of approximately three to one. The trend was most pronounced in Russian, where all discrepancies crossed a major border.

Advanced High. Second raters had the option of assigning either lower or higher ratings. A lower rating keeps a disagreement confined to the Advanced level, while a higher rating crosses a major border. With the exception of German, second ratings tended to stay within the Advanced level.

Superior. One can disagree with a Superior rating only by assigning a lower score as there are no levels above the Superior on the ACTFL scale. Since the scale defines the Advanced High speaker as an inconsistent Superior, and since tester training emphasizes the need to rate lower in borderline cases, all rating disagreements at the Superior level crossed a major border.

Conclusion

This study has demonstrated that a large and heterogeneous group of ACTFL-trained oral proficiency interviewers can apply the ACTFL oral proficiency scale in Spanish, French, Russian, ESL, and German with a fairly high degree of consistency as measured by both a lenient and a conservative statistic. Future attempts at improving inter-rater reliability should consider the following:

- Develop a set of taped interviews at all levels in each language.
- Have all certified testers rate these interviews.
- Collect data on their performance.
- Identify levels which cause the greatest disagreement among raters.
- Establish reasonable norms for rating accuracy.
- Recalibrate raters who are too lenient or too strict.

To avoid the rather consistent bias in second ratings, several solutions aimed at improving inter-rater consistency should be considered.

- Interviewers not assign ratings immediately after conducting an interview, but do so only after listening to the tape. Although current training of testers emphasizes the need to do so, this requirement is difficult to enforce.
- All interviews be rated by two second raters—a solution that will control for the rating environment but will be more time-consuming and costly.
- Interviews be videotaped, instead of audiotaped, as suggested by Clark and Lett (1988). A videotape will provide second raters with more cues than an audio tape, and thus make the second-rating environment more similar to the live interview. This solution is likely to be both costly and intrusive. The benefits of video- over audiotaping will need to be studied. Needless to say, videotaping is not a practical solution in long-distance interviewing.

We also need to decide whose rating should be accepted in cases of disagreement. It can be argued that while audio replay may allow raters to devote more attention to various details of the interviewee's performance, this rating environment is less representative of real-life situations in which the interviewee's performance is likely to be judged. The alternatives are:

- Accept the interviewer's rating as more 'ecologically' valid, albeit less stringent.
- Report both ratings and the conditions under which they were obtained.

Finally, the current requirements for tester certification need to be reviewed. At present, they stipulate that there may not be any major border disagreements between the trainee's and the trainer's ratings. This requirement is unreasonable because there is no evidence that perfect agreement between raters is possible at any level, particularly in the case of "High" ratings and the ones of the next higher level. The alternatives are:

- Require that trainer's and trainee's ratings be no more than one step apart. i.e., contiguous on the scale.
- Require several interviews for each step on the ACTFL scale and establish a tolerance standard for disagreements. This solution will greatly increase the number of interviews that must be submitted for certification and recertification.

This study needs to be replicated with examinees below the Intermediate Mid level where most of academic testing takes place. Although ratings below this level are mostly measures of achievement and do not represent language usable in real life. nevertheless, the profession owes it to its students to collect data on the reliability of the instrument it uses to evaluate them.

Finally, it must be pointed out that inter-rater agreement does not necessarily mean greater accuracy in mapping speaking behaviors onto the ACTFL scale. It merely means that raters agree on the criteria they use to evaluate speech samples. If two raters rate an Advanced performance as Superior, it means that they share the same bias, and does not

mean that the performance is Superior. Thus, inter-rater agreement is not a goal unto itself. It is desirable only when raters can appropriately match behaviors with steps on the scale, and when the scale itself is an adequate representation of speaking behavior.

ACKNOWLEDGMENTS

The author wishes to thank the following individuals without whose help this research would not have been possible. Helen Hamlyn of LTI for making the data available. Inge Ceunen of LTI and Judy Morag of ETS for recording and compiling the ratings. Marriette Reed of ETS, Charles Stansfield of the Center for Applied Linguistics, and David Hiple of the University of Hawai'i for providing valuable comments about the manuscript and sharing their insights. All errors of interpretation are mine.

References

Adams, M.L. 1978. 'Measuring foreign language speaking proficiency: A study of agreement among raters' in John L.D. Clark (ed.): *Direct Testing of Speaking Proficiency: Theory and Application* (pp. 129-49). Princeton, NJ: Educational Testing Service.

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

Bachman, L.F. 1988. 'Problems in examining the validity of the ACTFL oral proficiency interview.' *Studies in Second Language Acquisition* 10/2: 149-64.

Bachman, L.F. and J.L.D. Clark. 1987. 'The measurement of foreign/second language proficiency.' *Annals of the American Academy of Political and Social Science* 490: 20-33.

Buck, Katherine (ed.). 1989. *The ACTFL Oral Proficiency Interview Tester Training Manual.* Yonkers, NY: American Council on the Teaching of Foreign Languages.

Clark, John L.D. 1986. *A Study of the Comparability of Speaking Proficiency Interview Ratings across Three Government Language Training Agencies.* Washington, DC: Center for Applied Linguistics.

Clark, John L.D. 1988. 'Validation of a tape-mediated ACTFL/ILR-scale based test of Chinese speaking proficiency.' *Language Testing* 5/2: 187-205.

Clark, John L.D., and Ray T. Clifford. 1988. 'The FSI/ILR/ACTFL proficiency scales and testing techniques.' *Studies in Second Language Acquisition* 10/2:129-47.

Clark, John L.D., and John Lett. 1988. 'A research agenda' in Pardee Lowe, Jr., and Charles W. Stansfield (eds.): *Second Language Proficiency Assessment: Current Issues* (pp. 53-82). Englewood Cliffs, NJ: Prentice Hall.

Dandonoli, Patricia and Grant Henning. 1990. 'An investigation of the construct validity of the ACTFL proficiency guidelines and oral interview procedure.' *Foreign Language Annals* 23/1: 11-22.

Fleiss, Joseph L. 1971. 'Measuring nominal scale agreement among many raters.' *Psychological Bulletin* 75/5: 378-87.

Halleck, Gene B. 1992. 'The oral proficiency interview: Discrete point test or a measure of communicative language ability?' *Foreign Language Annals* 25/3: 227-32.

Lowe, Pardee, Jr. 1978. 'Third rating of FSI interviews' in John L.D. Clark (ed.): *Direct Testing of Speaking Proficiency: Theory and Application* (pp. 161-69). Princeton, NJ: Educational Testing Service.

Magnan, Sally Sieloff. 1986. 'Assessing speaking proficiency in the undergraduate curriculum: Data from French.' *Foreign Language Annals* 19/5:429-438.

Magnan, Sally Sieloff. 1987. 'Rater reliability of the ACTFL oral proficiency interview.' *The Canadian Modern Language Review* 43/2: 267-79.

Magnan, Sally Sieloff. 1988. 'Grammar and the ACTFL oral proficiency interview: Discussion and data.' *The Modern Language Journal* 72/3: 266-76.

North, Brian. 1993. 'The development of descriptors on scales of language proficiency: Perspectives, problems, and a possible methodology based on a theory of measurement.' NFLC Occasional Papers. Washington, DC: National Foreign Language Center.

Raffaldini, Tina. 1988. 'The use of situation tests as measures of communicative ability.' *Studies in Second Language Acquisition* 10/2: 197-216.

Stansfield, Charles W. and Dorry Mann Kenyon. 1992. 'Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview.' *System* 20/3: 347-64.

Valdman, Albert. 1988. 'Introduction.' *Studies in Second Language Acquisition* 10/2: 121-28.