A Study of Interrater 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

ABSTRACT *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. The present study, based on 795 double-rated oral proficiency interviews, was designed to consider the following questions: (1) What is the interrater reliability of ACTFL-certified testers in five European languages: ESL, French, German, Russian, and Spanish? (2) What is the relationship between interviewer-assigned ratings and second ratings based on audio replay of the interviews? (3) Does interrater reliability vary as a function of proficiency level? (4) Do different languages exhibit different patterns of interrater agreement across levels? (5) Are interrater disagreements confined mostly to the same main proficiency level? With regard to the above questions, results show: (1) Interrater reliability for all languages in this study was significant both when Pearson's r and Cohen's modified kappa were used. (2) When second-raters disagreed with interviewer-assigned ratings, they were three times as likely to assign scores that were lower rather than higher. (3) Some levels of performance are harder to rate than others. (4) The five languages exhibited different patterns of interrater agreement across levels. (5) Crossing of major borders was very frequent, and was dependent on the proficiency level. As a result of these findings, several practical steps are suggested in order to improve interrater reliability.*

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 interrater reliability of the OPI as practiced by testers trained by the American Council on the Teaching of Foreign Languages (ACTFL).

Irene Thompson (Ph.D., The George Washington University) is Professor of Russian at The George Washington University, Washington, DC.

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 ratings for arbitration by a third rater. ILR testers interview regularly and have an opportunity to maintain their calibration by comparing and discussing them with each other.

Adams (1978) studied the reliability of ILR interviews by having four German, six 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, trained and more experienced testers). She also found that French and Spanish *examiners* agreed among themselves more 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 and found a range of 0.90-0.92 for French, and 0.84-0.87 for German. The variation in the assignment of proficiency ratings by interviewers from the three agencies suggested that tester groups tend to 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: (1) 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). (2) 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. (3) ACTFL testers are scattered around the country and, unlike ILR testers, generally have few opportunities to maintain calibration by comparing and discussing their ratings with each other.

Magnan (1987) examined interrater reliability of an experienced ACTFL tester in French and 14 trainees who attended an intensive four-day workshop, and then conducted eight interviews on their own in Phase I, and seven additional interviews in Phase II of their training. The interviews were recorded and submitted to the trainer, who checked six of the 15 interviews. In both phases, correlations between trainer and trainee ratings were significant (Pearson's r 0.94, Cohen's kappa 0.53 for Phase I and 0.55 for Phase II, all significant beyond the 0.05 level). 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 interrater reliability in this study compared with the 1987 study could have been due 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 was assigned by the interviewer.

Based on 119 double-rated ACTFL interviews, Dandonoli and Henning (1990) reported alpha interrater reliabilities for mean of two raters that ranged from 0.85 to 0.98 in ESL and between 0.89 to 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 interrater 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 was not reported.

These small-scale studies of the ACTFL version of the OPI demonstrate high interrater 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 interrater 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 audiotape-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 a "High" rating and the next higher level, e.g., Advanced High and Superior. However, there is no empirical evidence to support the notion that interrater 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 interrater 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 interrater reliability vary as a function of proficiency level?
4. Do different languages exhibit different patterns of interrater agreement across levels?
5. Are interrater 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 percent) 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 1 *(see page 417)* 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.

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 represent a very broad spectrum of learners in terms of age, education, amount of exposure to the language, and type of language learning experience. Table 2 *(see page 417)* shows the distribution of ratings assigned by interviewers (first ratings).

Eighty-seven interviews (10.94 percent 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, interrater 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 3 *(see page 418)*. They were highly significant and remarkably similar in all five languages. The estimated variance (square of correlation), given in column two accounted for by speaking ability, was between 0.70 and 0.87. These results are surprisingly robust even though this is not a "lab" study of a group of hand-picked experienced testers where, according to North (1993:43), interrater reliabilities are expected to be high.

These reliability estimates are high because with only nine nominal categories, the possibility of interrater 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 interrater 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 an ordinal 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 an ordinal 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 interrater agreement than Pearson's r. Nevertheless, the results were also significant. The kappas in Table 4 *(see page 418)* 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 5 *(see page 418)* shows the frequency of interrater agreement at different levels collapsed across languages in terms of both raw scores and percentages. Overall, interrater reliability was greatest at the Superior

level, followed by Intermediate Mid, Advanced, Intermediate High, and Advanced High levels.

Table 6 *(see pages 419-20)* shows the frequency of agreement between raters for each language separately.

In Spanish, French, and Russian, interrater concurrence peaked at the Superior level, whereas in English and German, it was highest at the Intermediate Mid level. In Russian, interrater 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 7 *(see page 420)*. 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, $p<0.0001$).

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 8 *(see page 421)* 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.

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 levels (minor borders) and those that crossed borders between main levels (major borders). Table 9 *(see page 421)* 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 at the Superior level crossed a major border. Because of its placement on the scale, very few instances of rating disagreements crossed 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 10 *(see page 422)* 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.

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 percent 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 28.2 percent in Spanish to a low of 7.7 percent in German.

Discussion

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

1. What is the interrater reliability of ACTFL-certified testers in five European languages? Interrater 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 worked and tested 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 interrater reliabilities in this study are also lower than those reported by Magnan (1986) for ACTFL interviews in French, and by 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 interrater 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.[7]

2. What is the relationship between first and second ratings? The present study lends support to the hypothesis that interacting 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 interrater 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 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 her 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 toward 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. Does interrater reliability vary as 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 interviews had by far the lowest interrater reliability.

The highest frequency of interrater 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 interrater 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, probably, easy to identify on the ACTFL scale because it lumps all high-level performances into one category of Superior.

4. Do different languages exhibit different patterns of interrater agreement across levels? The five languages exhibited both similarities and differences that 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. Interrater 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 interrater disagreements mostly con-

fined to the same main proficiency level? Magnan (1986, 1987) reported that cases of interrater 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 a tendency on the part of French and Russian 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. The findings of this study suggest several practical steps that could be taken to improve interrater reliability.

Tester Certification

The current requirements for tester certification need to be reviewed. At present, no major border disagreements between trainee and trainer ratings are permitted. This requirement is unreasonable because there is no evidence that perfect agreement between raters is possible at any level, particularly at the "High" levels. The alternatives to the present requirements are:

- Eliminate the major-minor border distinction and require instead that trainer's and trainee's ratings be no more than one step apart, i.e., accept only contiguous disagreements.
- Require that participants in ACTFL oral proficiency testing workshops listen to a set of standard prerecorded interviews and demonstrate ability to rate with reasonable accuracy before they proceed with conducting their own interviews.
- For Phase I and Phase II of certification, require several interviews for each step on the ACTFL scale and establish a tolerance standard for disagreements between trainee and trainer. This solution will greatly increase the number of interviews that must be submitted for certification, and, therefore, hardly seems practical.

Tester Recertification

In order to maintain rating accuracy and prevent drift, the following measures should be considered:

- Develop a set of standard prerecorded interviews at all levels in each language.

- Have all certified testers rate these interviews.
- Collect data on their performance.
- Identify levels that cause the greatest disagreement among raters.
- Establish reasonable norms for rating accuracy.
- Recalibrate raters who are too lenient or too strict.

Rating Procedures

To avoid the rather consistent bias in second ratings, several solutions aimed at improving interrater 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 audiotape 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 videotaping over audiotaping will need to be studied. Needless to say, videotaping is not a practical solution in long-distance interviewing.

Reporting of Ratings

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.

Research

This study needs to be replicated with examinees below the Intermediate Mid level where most 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.

A study of interrater reliability of other languages is also called for.

Finally, it must be pointed out that interrater 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, interrater 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: Ed Scebold of ACTFL and 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, David Hiple of the University of Hawai'i, and John Miles of ACTFL for providing valuable comments about the manuscript and sharing their insights; Carol Reisen for helping with the statistics and last, but not least, Richard Thompson for his patient reading of the manuscript. All errors of interpretation are mine.

NOTES

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

[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.

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

[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.

[6] Main levels are Novice, Intermediate, Advanced, and Superior. Adjacent ratings such as Intermediate Mid and Intermediate High are within the same main level and, therefore, cross a minor border. Contiguous ratings such as Intermediate High and Advanced belong to two different main level and, therefore, cross a major border.

[7] Magnan (1987) reported Cohen's kappa of 053-0.55; a modified Cohen's kappa for French in this study was 0.531. This means that in both studies the chances of two raters assigning the same rating in French are over five to one.

REFERENCES

ACTFL. 1986. *Proficiency Guidelines.* Hastings-on-Hudson, NY: ACTFL.

Adams, M.L. 1978. "Measuring Foreign Language Speaking Proficiency: A Study of Agreement Among Raters," 129-49, in John L.D. Clark, ed., *Direct Testing of Speaking Proficiency: Theory and Application,*. Princeton, NJ: Educational Testing Service.

Bachman, L.F. 1988. "Problems In Examining the Validity of the ACTFL Oral Proficiency Interview." *Studies in Second Language Acquisition* 10: 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: ACTFL.

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.

_____. 1988. "Validation of a Tape-mediated ACTFL/ILR-scale Based Test of Chinese Speaking Proficiency." *Language Testing* 5:187-205.

Clark, John L.D., and Ray T. Clifford. 1988. "The FSI/ILR/ACTFL Proficiency Scales and Testing Techniques." *Studies in Second Language Acquisition* 10:129-47.

Clark, John L.D., and John Lett. 1988. "A Research Agenda," 54-82, in Pardee Lowe, Jr., and Charles W. Stansfield, eds., *Second Language Proficiency Assessment: Current Issues.* 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: 11-22.

Fleiss, Joseph L. 1971. "Measuring Nominal Scale Agreement among Many Raters." *Psychological Bulletin* 75: 378-87.

Halleck, Gene B. 1992. "The Oral Proficiency Interview: Discrete Point Test or a Measure of Communicative Language Ability?" *Foreign Language Annals* 25: 227-32.

Hiple, David. 1994. Personal communication.

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

Magnan, Sally Sieloff. 1986. "Assessing Speaking Proficiency in the Undergraduate Curriculum: Data from French." *Foreign Language Annals* 19:429-38.

_____. 1987. "Rater Reliability of the ACTFL Oral Proficiency Interview." *The Canadian Modern Language Review* 43: 267-79.

_____. 1988. "Grammar and the ACTFL Oral Proficiency Interview: Discussion and Data." *The Modern Language Journal* 72: 266-76.

North, Brian. 1993. "The Development of Descrip-

tors 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: 197-216.

Reed, Marriette. 1994. Personal communication.

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: 347-64.

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

TABLE 1

Distribution of ACTFL-Certified Testers

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

TABLE 2

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%

IM=Intermediate Mid; IH=Intermediate High; A=Advanced; AH=Advanced High; S=Superior

TABLE 3

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

* $p < 0.0001$

TABLE 4

Interrater Reliability as Measured by Modified Cohen's Kappa

Spanish	French	Russian	English	German
0.474	0.531	0.443	0.469	0.516

TABLE 5

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 6

Relationship Between First and Second Ratings by Languages

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

TABLE 6 *(continued)*

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

TABLE 7

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%

TABLE 8

Rating Disagreements in Terms of Steps

	One Step Disagreements	Two Step Disagreements
Spanish	150 80.65%	36 19.35%
French	52 89.66%	6 10.34%
Russian	29 82.86%	6 17.14%
ESL	22 88.00%	3 12.00%
German	16 88.89%	2 1.11%
Total	269 83.54%	53 16.46%

TABLE 9

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 10

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%