Reliability and the ACTFL Oral Proficiency Interview: Reporting Indices of Interrater Consistency and Agreement for 19 Languages

Eric A. Surface
Surface, Ward & Associates

Erich C. Dierdorff
DePaul University

Abstract: *The reliability of the ACTFL Oral Proficiency Interview (OPI) has not been reported since ACTFL revised its speaking proficiency guidelines in 1999. Reliability data for assessments should be reported periodically to provide users with enough information to evaluate the psychometric characteristics of the assessment. This study provided the most comprehensive analysis of ACTFL OPI reliability to date, reporting interrater consistency and agreement data for 19 different languages. Overall, the interrater reliability of the ACTFL OPI was found to be very high. These results demonstrate the importance of using an OPI assessment program that has a well-designed interview process, a well-articulated set of criteria for proficiency determination, a solid rater training program, and an experienced cadre of testers. Based on the data reported, educators and employers who use the ACTFL OPI can expect reliable results and use the scores generated from the testing process with increased confidence. Recommendations for future research are discussed.*

Introduction

In 1999, the ACTFL revised the ACTFL Proficiency Guidelines—Speaking (Breiner-Sanders et al., 2000) "to make the document more accessible to those who have not received recent training in ACTFL oral proficiency testing, to clarify the issues that have divided testers and teachers, and to provide a corrective to what the committee perceived to have been possible misinterpretations of the descriptions provided in earlier versions of the Guidelines" (p. 14). One of the most significant changes to the 1986 Guidelines from a measurement perspective was "the division of the Advanced level into the High, Mid, and Low sublevels" (p. 14). Previously, there were two categories at the Advanced level (Advanced and Advanced-High). The recent 1999 Guidelines subsequently created three rating options to describe the Advanced level of proficiency—Advanced-High, Advanced-Mid,

Eric A. Surface (PhD, North Carolina State University) is a principal and researcher with Surface, Ward & Associates, an organizational consulting and research firm based in Raleigh, North Carolina, and serves as the Director of Training Research for the Special Operations Forces Language Office, Ft. Bragg, North Carolina, as part of his fellowship with the Army Research Institute's Consortium Research Fellows Program.

Erich C. Dierdorff (PhD, North Carolina State University) is a Visiting Professor in the College of Commerce at DePaul University and a Consortium Research Fellow with the Army Research Institute, Ft. Bragg, North Carolina.

and Advanced-Low—by dividing the previous Advanced proficiency category into two and retaining the Advanced-High level to form the current conceptualization of the Advanced major level.

Although adapting the scale did not change the Oral Proficiency Interview (OPI) process, modifying the rating scale used by ACTFL certified testers to describe speaking proficiency might have an impact on the measurement properties of the assessment. Therefore, it is important to determine whether the change in the Guidelines affected the psychometric characteristics (i.e., the validity and reliability) and quality of the ratings generated during the ACTFL OPI process.

The *Standards for Educational and Psychological Testing,* published by the American Educational Research Association (AERA) (1999), provide evaluative guidelines for the users, developers, and publishers of *tests*, referring to any "evaluative device or procedure in which a sample of an examinee's behavior in a specified domain [test content area] is obtained and subsequently evaluated and scored using a standardized process" (p. 3), not simply restricted to paper-and-pencil assessments. *Validity* refers to "the degree to which evidence and theory support the interpretations of the test scores entailed by proposed uses of tests" (p. 9), whereas *reliability* indicates the ability of the testing procedure to provide a consistent measure of the specified domain when repeated.

The validity and reliability of a testing procedure should be demonstrated periodically, especially if the procedure has been modified or the specified domain has been redefined in some meaningful way. Thus, our study assessed the reliability of the ACTFL OPI across and within 19 languages. As with previous reliability studies (e.g., Thompson, 1995), we did not address the validity of the ACTFL OPI as a measure of speaking proficiency. Establishing validity requires multiple studies that provide evidence supporting that a test or assessment effectively measures the construct it purports to measure and can be used for a specific purpose. Validity evidence can take many forms, depending on the use of the test and the purpose of the validation study (e.g., criterion-related validation if the assessment is used to predict future job performance). This type of research is beyond the scope of the current study. The information presented herein specifically addresses the requirements of the Standards related to presenting reliability data and will help users of the ACTFL OPI evaluate the ratings produced by the procedure.

Research Background and Literature Review

ACTFL Proficiency Guidelines—Speaking, Revised

The main impetus for reassessing the reliability of the ACTFL OPI comes from the revision of the proficiency guidelines (Breiner-Sanders et al., 2000). The Revised Guidelines made several modifications to the previous version of the ACTFL Proficiency Guidelines (ACTFL, 1986). Although the committee made several changes related to the presentation of the Guidelines, the primary modification was to divide what was previously defined as the Advanced level of proficiency into the Advanced-Mid and Advanced-Low sublevels and aggregate the two new categories, along with the existing Advanced-High level, into the current conceptualization of the Advanced major level. Refinement of the measurement scale in this way could substantially affect the psychometric properties of the ACTFL OPI, thus making it necessary to examine the reliability of ratings produced from the revised criteria. Of course, this reliability research could not be conducted until enough data were available for analysis.

The Guidelines provide an *a priori* set of criteria against which interviewers measure and evaluate an individual's functional competency in speaking a language, as demonstrated by the test taker's ability to accomplish linguistic tasks at the various proficiency levels. The ACTFL guidelines are based on the Interagency Language Roundtable (ILR) descriptions of language proficiency for use in governmental and military organizations and have been modified for use in academia and industry. The ACTFL rating scale describes four major levels of language proficiency—Superior, Advanced, Intermediate, and Novice—that are delineated according to a hierarchy of global tasks related to functional language ability (e.g., can narrate and describe in all major time frames).

Three of the major levels (Advanced, Intermediate, and Novice) are further divided into three sublevels—High, Mid, and Low. Superior is the only major category that is not divided into sublevels. Combining the ACTFL major levels and sublevels yields a total of 10 separate proficiency categories. A complete description of the proficiency categories is provided in the published Guidelines (Breiner-Sanders et al.,2000) and can also be obtained through the ACTFL Web site (www.actfl.org).

Reliability and Interrater Consistency

Consistency defined by the extent that separate measurements retain relative position is the essential notion of classical reliability (Anastasi, 1988; Cattell, 1988; Feldt & Brennan, 1989; Flanagan, 1951; Stanley, 1971; Thorndike, 1951). Simply put, reliability is the extent to which an item, scale, procedure, or instrument will yield the same value when administered across different times, locations, or populations. In the specific case of rating data, the focus of reliability estimation turns to the homogeneity of judgments given by the sample raters. One of the most commonly used forms of rater reliability estimation is interrater reliability, which portrays the overall level of consistency among the sample of raters involved in a particular judgment process. When interrater reliability estimates are high, the interpretation has a large degree of consistency across sample raters.

Another common approach to examining interrater consistency is to use measures of agreement. Whereas

interrater reliability estimates are parametric and correlational in nature, measures of agreement are nonparametric and assess the extent to which raters give concordant or discordant ratings to the same objects (e.g., interviewees). Technically speaking, measures of agreement are not indices of reliability per se, but are nevertheless quite useful in depicting levels of rater agreement and consistency of specific judgments, particularly when data can be considered ordinal or nominal.

Items, tests, raters, or procedures generating judgments must yield reliable measurements to be useful and have psychometric merit. Data that are unreliable are, by definition, unduly affected by error, and decisions based upon such data are likely to be quite tenuous at best and completely erroneous at worst. Although validity is considered the most important psychometric measurement property (AERA, 1999), the validity of an assessment is negated if the construct or content domain cannot be measured consistently. In this sense, reliability can be seen as creating a ceiling for validity.

The *Standards for Educational and Psychological Testing* (AERA, 1999) provide a number of guidelines designed to help test users evaluate the reliability data provided by test publishers. According to the Standards, a test developer or distributor has the primary responsibility for obtaining and disseminating information about an assessment procedure's reliability. However, under some circumstances, the user must accept responsibility for documenting the reliability and validity in its local population. The level of reliability evidence that is necessary to assess and to be reported depends on the purpose of the test or assessment procedure. For example, if the assessment is used to make decisions that are "not easily reversed" or "high stakes" (e.g., employee selection or professional school admission), then "the need for a high degree of precision [in the reliability data reported] is much greater" (p. 30).

Given the nature of the ACTFL OPI and our study, the following Standards (AERA, 1999) are particularly noteworthy: (1) reliability estimates should be reported for each test score, subscore, or combination of scores (Standard 2.1); (2) reliability coefficients from similar assessments (e.g., Defense Language Institute's [DLI] OPI) are not interchangeable unless their implicit definitions of measurement error are equivalent (Standard 2.5); (3) evidence of both interrater consistency and within examinee consistency over repeated measurements should be provided for assessments when subjective judgment enters into the scoring process (Standard 2.10); (4) test developers should document the process for the selection and training of raters as well as scorer reliability and drift over time (Standard 3.23); and (5) test developers and publishers are responsible for amending, revising, or withdrawing a test as new research data becomes available (Standard 3.25). Taken together, providers of OPIs or other test/assessment procedures have the responsibility to report and periodically update the reliability data for their procedures. Thus, the Standards provide a strong justification for the research in this study.

Previous OPI Reliability Research

Although several studies have investigated and reported reliability data for the ACTFL OPI (e.g., Magnan, 1987; Thompson, 1995), all available reliability evidence predates the Revised Guidelines and uses data collected under the 1986 criteria (ACTFL, 1986).

Magnan (1986) found that interrater agreement for a sample of 40 students of French rated by two ACTFL-certified testers from the Educational Testing Service was .72 (Cohen's kappa). All rater disagreements were one sublevel apart (e.g., Mid versus High) within the same major proficiency level (e.g., Intermediate). Magnan (1987) found that interrater reliability between trainer and trainee ratings of French proficiency in a two-phase study were significant (the coefficients were r = .94, .94; τ = .83, .86; K = .53, .55; and Γ = .94, .95 for both phases respectively), and rater disagreements were again within only one sublevel in the majority of the instances.

In a construct validity study, Dandonoli and Henning (1990) reported interrater reliability for ratings of speaking proficiency in English (r = .98) and in French (r = .97). Thompson (1995) presented the most comprehensive evaluation of interrater reliability for the ACTFL OPI under the 1986 Guidelines. Thompson (1995) provided coefficients (Pearson's correlations) for five languages: .87 for French, .85 for Spanish, .90 for Russian, .84 for English, and .86 for German. Modified Cohen's kappa coefficients and the percentages of absolute and partial agreement were reported as well. The current study builds upon and extends this body of research with the ACTFL OPI.

Although other organizations (e.g., DLI) provide assessments of speaking proficiency, no comprehensive reports of interrater reliability data were found in a review of the past decade's research. However, some studies (e.g., Jackson, 1999) have reported reliability data within the confines of their defined research scopes. Jackson (1999), who investigated the impact of test modality (e.g., telephone or face-to-face) on oral proficiency testing at DLI, reported Kendall's *tau-b* coefficients for Russian and Arabic OPI ratings across several modalities for a small sample of participants. The coefficients ranged from .90 to 1.00 for the original ratings within the same testing mode (see Jackson, 1999 for details).

Although previous research supports the reliability of ILR-based OPI ratings (e.g., Adams, 1978; Bachman & Palmer, 1981; Carroll, 1967; Clark, 1986), the lack of current and comprehensive reliability data from large samples makes comparisons inappropriate. Additionally, differences in the rating process between ACTFL and

other OPI processes (e.g., DLI) limit the comparability of the results as well. Therefore, our study will only discuss the current findings in the context of previous research specifically related to the ACTFL OPI.

Research Questions

With the previously discussed considerations in mind, the present study sought to investigate the levels of interrater consistency derived from experienced ACTFL-certified testers using the Revised Guidelines. The following six specific research questions were examined:

1. What is the overall interrater consistency and agreement for all languages tested with the same ACTFL Revised Guidelines and rating protocol? Overall interrater consistency and agreement refers to calculating the coefficients across all pairs of raters in all languages.
2. Do interrater consistency and rater agreement levels vary across languages that are more commonly tested compared to those that are less commonly tested?
3. Do interrater consistency and rater agreement levels vary according to language difficulty (i.e., the level of difficulty for learning a given language)?
4. Do ratings of particular languages show more consistency or greater agreement than others?
5. Does rater agreement vary across proficiency categories? If so, what is the nature of disagreement (i.e., within a major proficiency level or between two major proficiency levels)?
6. When the first and second raters disagree and a third rater must be utilized, is the third rater significantly more likely to resolve the disagreement in favor of one rater more often than the other (i.e., are interrater reliability and agreement higher between the first and third raters or the second and third raters)?

Methods

Participants and Rating Methodology

A total of 5881 interviews conducted and rated by experienced ACTFL-certified testers and using the ACTFL assessment procedure were included in this study. The ACTFL OPI assessment procedure, as described in the *ACTFL Oral Proficiency Interview Tester Training Manual* (Swender, 1999), consists of four phases (Warm Up, Level Checks, Probes, and Wind Down) that are designed to efficiently elicit a ratable sample.

This study used data from oral proficiency interviews in 19 different languages: English, Mandarin, French, German, Italian, Japanese, Russian, Spanish, Hebrew, Czech, Arabic, Vietnamese, Portuguese, Polish, Albanian, Hindi, Tagalog, Cantonese, and Korean. Table 1 provides the number of interviews included in the study by language. The data were made available by Language Testing International (LTI), the ACTFL testing affiliate.

Two characteristics of the tested languages were used to code each case (each interviewee's data represents a case) into groups used for subsequent analyses. The first language characteristic used in this study was *testing density*, which represented LTI's frequency of assessing a particular language. Cases in languages with high-testing volumes were coded as More Commonly Tested (MCT) languages, while languages with lower volumes were coded as Less Commonly Tested (LCT). LTI's own internal categorization was used to code testing density. All cases in English, Mandarin, French, German, Italian, Japanese, Russian, and Spanish were considered MCT languages. All cases in Hebrew, Czech, Arabic, Vietnamese, Portuguese, Polish, Albanian, Hindi, Tagalog, Cantonese, and Korean were considered LCT languages.

The second language characteristic was *language difficulty*, which was derived by applying the language difficulty categories used by the American Council on Education (ACE) in its recommendations for granting college credit for official ACTFL OPI ratings to each of the cases. These categories were labeled Category I through Category IV and represent the relative difficulty for learning the language from the perspective of a native English speaker. Higher categories represent more difficult languages to learn. For its purposes, ACE considers English a Category I language. However, the language category assignment should differ depending on the speaker's first language. Although coding English as a Category I language in our analyses could be potentially problematic, we chose to mirror the operationalization of the Categories in the ACE recommendations for college credit to maintain alignment with their use. The language difficulty categories are equivalent to the ones used by military and governmental organizations.

As stipulated by the standard procedure for all ACTFL OPI assessments, each case was rated by a pair of testers. Some cases required a third tester to serve as a "tie-breaker" in situations of discrepancy between the pair's proficiency ratings. In all cases, the first rater conducted and audiotaped the interviews. Subsequently, this rater judged the interviewee's speaking proficiency from the tape at some later time.

Next, the taped interviews were independently rated by a second rater. All raters used the ACTFL rating scale described in the ACTFL Proficiency Guidelines—Speaking, Revised (Breiner-Sanders et al., 2000) to describe the proficiency levels of the interviewees. If the independent ratings provided by the rating pair disagreed, a third rater was assigned as an arbitrator to rate the interview tape. This rater did not know the previously assigned scores, nor that he or she was the third rater. No fourth raters were needed to reach a final rating (i.e., the rating of the third rater always agreed with the rating of either the first or second rater).

Throughout this article, the "first" rater always corresponds to the tester who conducted the interview, whereas the "second" and "third" raters represent those who rated inter-

viewees from the audiotapes. All raters were ACTFL-certified, meaning that they had completed the ACTFL OPI tester certification process as described in the *ACTFL OPI Tester Certification Information Application Packet* (ACTFL, 2002). These testers are required to keep current through ongoing training, testing, and norming procedures. Testing experience varied across raters. Both native and nonnative speakers served as raters. The total number of certified testers also varied across languages.

Analytic Procedure

In order to more accurately assess the extent of interrater consistency, we used a multimethod approach. Interrater consistency can be conceptualized from several perspectives (e.g., interrater reliability, interrater agreement, and so forth) and, thus, a multimethod approach allows for a more complete picture of the level of rating consistency. We also sought to include similar statistics to those previously employed in prior research examining interrater consistency of the ACTFL OPI. The overall rationale was to expand the breadth of rater consistency assessment, as well as to yield estimates comparable to past assessments.

Pearson correlation. Sometimes called a product–moment correlation, Pearson correlation (r) is one the most widely used methods of assessing interrater reliability. This correlation assesses the degree to which ratings covary. In this sense, reliability can be depicted in the classical framework as the ratio of true score variance to total variance (i.e., variance in ratings attributable to true speaking proficiency divided by total variance of ratings).

Spearman rank–order correlation. This is another commonly used correlation for assessing interrater reliability, particularly in situations involving ordinal variables. Spearman rank–order correlation (R) has a interpretation similar to Pearson's r; the primary difference between the two correlations is computational, as R is calculated from ranks and r is based on interval data. This statistic is appropriate for the OPI data in that the proficiency categories are ordinal in nature.

Kendall's tau. Tau (τ) is equivalent to Spearman's R with regard to the underlying assumptions. However, *tau* and R carry different interpretations. R is a correlation and thus represents a proportion of variability accounted for, whereas tau is a measure of agreement and represents the difference between two probabilities. *Tau* is the difference between the probability that the cases are rated in the same order by the two raters and the probability that the cases are rated in different orders by the two raters.

Goodman and Kruskal's gamma. Similar to *tau, gamma* (Γ) is a probability-based measure of agreement. However, unlike *tau, gamma* does not penalize for ties in that they are computationally ignored. As it is desirable to have high interrater consistency (i.e., a large number of tied ratings), *gamma* can provide useful information beyond that given by tau in terms of interrater consistency. As tied ratings are computationally ignored, the result is that *gamma* is typically higher in magnitude than tau.

Cohen's kappa. Cohen's *kappa* (κ) is another commonly used measure of agreement, which compares the observed agreement to the agreement expected by chance. *Kappa* values range from 1.00, when agreement is perfect, to 0.00, when agreement is at the chance level. *Kappa* does not take into account the degree of disagreement between raters as all disagreements are considered to contribute equally to the total level of disagreement. Therefore, if rating categories are ordered, it is preferable to use a weighted version of *kappa*, which assigns different weights to ratees for whom the raters differ by i categories. Thus, different levels of disagreement can contribute proportionally to the overall value of *kappa*. Weighted *kappa* was used in this study.

Raw percentages of agreement. This agreement method assesses the extent to which raters display perfect agreement. It serves as an absolute agreement estimate of interrater consistency and is calculated as the number of identical ratings divided by the number of total rating opportunities. As some disagreements can be expected, it is important to assess percentages of partial agreement as well. Thus, we estimated three separate partial agreement percentages: (1) interrater agreement within plus or minus one proficiency category (e.g., Novice-Low versus Novice-Mid); (2) interrater agreement within plus or minus two proficiency categories (e.g., Intermediate-Low versus Intermediate-High); and, (3) interrater agreement within plus or minus three proficiency categories (e.g., Advanced-Low versus Superior). In addition, some disagreements can be viewed as more severe in terms of language proficiency determination. For example, a partial interrater agreement that spans a major proficiency category boundary (e.g., first rater judges an Intermediate-High, while second rater judges an Advanced-Low) could be a more problematic discrepancy than a partial agreement *within* a major proficiency category, such as one spanning a minor proficiency boundary (e.g., Intermediate-Low versus Intermediate-Mid). To account for the specific nature of rater disagreements, we calculated the overall frequencies of rater disagreements that spanned one or more major proficiency categories. Also, we examined the specific locations of these major boundary-crossing disagreements (e.g., disagreements crossing Intermediate and Advanced versus those crossing Advanced and Superior).

Results

Research question 1: What is the overall interrater consistency and agreement for all languages tested with the same ACTFL Revised Guidelines and rating protocol? As shown in Table 1, the overall interrater consistency across all rater pairs in all included languages was significant ($p < .05$) for each of the test statistics. As expected, *gamma* had the highest value and all consistency measures had values greater than .90.

Table 2 displays the raw agreement percentages for the

Table 1

INTERRATER CONSISTENCY ANALYSES

Data Type	*N*	*r*	*R*	τ	Γ	K_{wt}
Overall	5881	.978	.976	.940	.990	.920
Language Density						
MCT	5389	.978	.975	.940	.991	.918
LCT	492	.979	.978	.941	.981	.929
Language Difficulty						
Category I	4458	.975	.971	.934	.991	.912
Category II	216	.981	.976	.945	.994	.929
Category III	441	.985	.983	.954	.990	.941
Category IV	766	.978	.977	.938	.981	.920
Language						
English	725	.960	.957	.912	.984	.883
Mandarin	241	.989	.989	.966	.997	.951
French	626	.977	.979	.949	.949	.927
German	216	.981	.976	.945	.994	.929
Italian	219	.944	.938	.886	.978	.844
Japanese	307	.981	.971	.933	.984	.924
Russian	278	.974	.966	.922	.980	.902
Spanish	2777	.978	.970	.934	.991	.917
Hebrew	19	.996	.999	.993	1.00	.980
Czech	15	.999	.999	1.00	1.00	1.00
Arabic	140	.946	.943	.864	.940	.822
Vietnamese	42	.999	.999	1.00	1.00	1.00
Portuguese	111	.982	.976	.947	.994	.930
Polish	25	.999	.999	1.00	1.00	1.00
Albanian	8	.959	.992	.972	1.00	.889
Hindi	47	.999	.999	1.00	1.00	1.00
Tagalog	7	.978	.971	.946	1.00	.920
Cantonese	13	.981	.981	.953	1.00	.904
Korean	65	.999	.999	1.00	1.00	1.00

Note. *r* = Pearson correlation; *R* = Spearman rank-order correlation; τ = Kendall's *tau*; Γ = Goodman-Kruskal *gamma*; K_{wt} = Cohen weighted *kappa* coefficient; MCT = more commonly tested; LCT = less commonly tested; all statistics are significant ($p < .05$).

overall language data. Eighty percent of the ratings across all 19 languages showed perfect agreement, whereas about 18% of the ratings disagreed by one proficiency category. That is, four-fifths of all rater pairs gave identical proficiency ratings and nearly all rater pairs (99%) were within one proficiency category (e.g., Novice-Low versus Novice-Mid).

Research question 2: Do interrater consistency and rater agreement levels vary across languages that are more commonly tested compared to those that are less commonly tested? The results in Tables 1 and 2 show that there were very small differences in both rater consistency and rater agreements levels between languages that are more commonly tested and those that are less frequently tested.

Research question 3: Do interrater consistency and rater agreement levels vary according to language difficulty? The results of the consistency measures (Table 1) demonstrate that the language difficulty classifications had practically no moderating effects on the magnitude of rater consistency. Similarly, the raw agreement percentages (Table 2) did not show any substantial discrepancies across the four language difficulty groups. Category III languages produced slightly higher levels of agreement, but this difference was quite small relative to the other three categories.

Research question 4: Do ratings of particular languages

Table 2

PERCENTAGES OF INTERRATER AGREEMENT

Data Type	Agreement	Disagreement Distance		
	Absolute	1 Step	2 Steps	3 Steps
Overall	80.79	18.59	.58	.05
	(4751)	(1093)	(34)	(3)
Language Density				
MCT	80.63	19.00	.37	.
	(4345)	(1024)	(20)	.
LCT	82.52	14.02	2.85	.61
	(406)	(69)	(14)	(3)
Language Difficulty				
Category I	80.64	19.09	.27	.
	(3595)	(851)	(12)	.
Category II	82.87	16.67	.46	.
	(179)	(36)	(1)	.
Category III	83.90	14.97	1.13	.
	(370)	(66)	(5)	.
Category IV	79.24	18.28	2.09	.39
	(607)	(140)	(16)	(3)
Proficiency Category				
Novice				
Low	94.44	5.56	.	.
	(51)	(3)	.	.
Mid	76.40	21.35	2.25	.
	(68)	(19)	(2)	.
High	81.63	15.65	2.72	.
	(120)	(23)	(4)	.
Intermediate				
Low	72.76	26.25	1.00	.
	(219)	(79)	(3)	.
Mid	79.73	19.46	.82	.
	(586)	(143)	(6)	.
High	78.27	21.35	.38	.
	(616)	(168)	(3)	.
Advanced				
Low	76.49	22.83	.41	.27
	(563)	(168)	(3)	(2)
Mid	75.93	23.41	.55	.11
	(694)	(214)	(5)	(1)
High	75.47	23.73	.80	.
	(563)	(177)	(6)	.
Superior	92.71	7.15	.15	.
	(1271)	(98)	(2)	.

Table 2 (continued)

Data Type	Agreement	Disagreement Distance		
	Absolute	1 Step	2 Steps	3 Steps
Language				
English	77.93	21.52	.55	.
	(565)	(156)	(4)	.
Mandarin	86.31	13.69	.	.
	(208)	(33)	.	.
French	84.66	15.34	.	.
	(530)	(96)	.	.
German	82.87	16.67	.46	.
	(179)	(36)	(1)	.
Italian	73.97	25.57	.46	.
	(162)	(56)	(1)	.
Japanese	79.80	19.22	.98	.
	(245)	(59)	(3)	.
Russian	75.54	22.66	1.80	.
	(210)	(63)	(5)	.
Spanish	80.88	19.91	.22	.
	(2246)	(525)	(6)	.
Hebrew	94.74	5.26	.	.
	(18)	(1)	.	.
Czech	100.0	.00	.	.
	(15)	0	.	.
Arabic	56.43	32.14	9.29	2.14
	(79)	(45)	(13)	(3)
Vietnamese	100.0	.	.	.
	42	.	.	.
Portuguese	82.88	16.22	.90	.
	(92)	(18)	(1)	.
Polish	100.0	.	.	.
	(25)	.	.	.
Albanian	87.50	12.50	.	.
	(7)	(1)	.	.
Hindi	100.0	.	.	.
	(47)	.	.	.
Tagalog	85.71	14.29	.	.
	(6)	(1)	.	.
Cantonese	76.92	23.08	.	.
	(10)	(3)	.	.
Korean	100.0	.	.	.
	(65)	.	.	.

Note. Sample sizes shown in parentheses; Steps = 10 specific proficiency rating values; MCT = "more commonly tested"; LCT = "less commonly tested"; proficiency categories derived from *ACTFL Proficiency Guidelines–Speaking, Revised.*

show more consistency or greater agreement than others? When taken collectively, the results of the consistency analyses showed no substantially large differences across the 19 tested languages. For instance, values of *r* ranged from .94 to .99. The largest spread of any specific consistency statistic across languages was for the weighted kappa statistic (.82 to 1.0). Some small language effects were apparent for Italian and Arabic data, which both had slightly lower levels of rater consistency. Importantly, caution should be used when interpreting these results, in that several languages (e.g., Albanian) had very small sample sizes and were presented for the sake of illustration and completeness.

The bottom half of Table 2 provides the results of the raw agreement percentage analysis by language. The raw percentage for absolute agreement ranged from a low of 56% (Arabic) to a high of 100% (Czech, Vietnamese, Polish, Hindi, and Korean) across the 19 languages. The vast majority of languages had greater than 80% perfect rater agreement. For those languages with less than 100% rater agreement, the majority displayed differences of only one proficiency category between raters. Again, some of these results should be interpreted with caution due to small sample sizes.

Research question 5: Does rater agreement vary across proficiency categories? If so, what is the nature of disagreement? From the results shown in Table 2, ratings of Novice-Low proficiency tended to have the highest level of absolute agreement, followed by ratings of Superior proficiency (93%). Overall, the level of absolute agreement in the Novice proficiency level tended to be fairly high (94%, 76%, and 81% across the specific Novice sublevels). Intermediate and Advanced proficiency ratings showed very similar overall absolute and partial agreement percentages, as well as across their respective proficiency sublevels.

With agreement differences evident across the 10 proficiency categories, the nature of the rating disagreement became an important issue. Table 3 shows the results pertinent to this line of inquiry. After completing the initial analyses presented in Table 2, a more specific agreement percentage analysis was undertaken to examine the nature of disagreements between raters. Out of the 5881 total rater pairs, there were 1130 rater pairs that "disagreed." These pairs were further analyzed to capture the location along the ACTFL language proficiency categories where the disagreements were most prevalent (i.e., whether the disagreements were across major or minor boundaries). One focus of this analysis was on pairs of ratings where the disagreement was between ratings from different major categories or levels, that is, rater disagreements leading to incongruous categorical assignments of the interviewees (e.g., one rater giving a "Novice" and the other giving an "Intermediate" assignment). This is also referred to as "crossing a major boundary."

Table 3

PERCENTAGE OF DISAGREEMENT ACROSS MAJOR PROFICIENCY LEVELS

Data Type	%
Overall	
1 Boundary	41.50 (469)
2 Boundaries	0.00
Specific	
Novice–Intermediate	12.58 (59)
Intermediate–Advanced	39.45 (185)
Advanced–Superior	47.97 (225)

Note. Sample sizes in parentheses; overall disagreement cases = 1130 of 5881; total number of ratees in each proficiency category: novice (290), intermediate (1823), advanced (2396), superior (1371); proficiency categories derived from ACTFL Proficiency Guidelines–Speaking Revised 1999. The disagreements in this table only refer to crossing a major proficiency boundary.

As shown in Table 3, approximately 41% of disagreement cases crossed a single major proficiency level boundary. No disagreements were associated with crossing two major proficiency categories. Thus, close to three-fifths of all rater disagreements were within a given major proficiency category (i.e., the disagreements were within a single major level), which is also referred to as "crossing a minor boundary." Of the 41% disagreement cases that crossed a major level boundary, the majority (48%) were between the Advanced and Superior proficiency categories, followed by the Intermediate and Advanced (39%) and the Novice and Intermediate (13%) categories. These percentages matched the proportions of total test takers that fell within these categories. That is to say, a vast majority of interviewees were judged to be Advanced (n = 2396) or Superior (n = 1371), thus paralleling the larger percentage of disagreements spanning these two major proficiency categories.

Research question 6: When the first and second raters disagree and a third rater must be utilized, is the third rater significantly more likely to resolve the disagreement in favor of one rater more often than the other? In accordance with the ACTFL guidelines, disagreement between two raters necessitates a third rater to serve as a "tie-breaker." Table 4 shows the results of interrater reliability analysis, using Pearson's r, comparing consistency between any two original raters and the third arbitrating raters. The previous moderator variables (e.g., language) were also included in these reliability analyses. The results were quite consistent in showing that interrater reliability was clearly higher for second raters paired with third raters than it was for the first and third rater pairs. This relationship held across the testing density, language difficulty classifications, and specific languages as well.

To test whether or not the interrater reliability differences across rater combination were statistically significant, we used a modified t-test (Stieger, 1980). A modified t-test was chosen over a traditional t-test (i.e., Hotelling's "exact" t-test) because it is more appropriate for "correlated correlations" (Meng, Rosenthal, & Rubin, 1992), that is, correlations derived from samples that are not independent (from the same population). Across testing density and language difficulty, all interrater reliabilities were significantly different ($p < .05$, one-tailed). Within specific languages, only two interrater reliabilities were

Table 4

COMPARING INTERRATER RELIABILITY BY RATER COMBINATION

Data Type	*N*	*r*		% of Absolute Agreement	
		Raters 1 & 3	Raters 2 & 3	Raters 1 & 3	Raters 2 & 3
Overall	1130	.907	.960**	28.23	68.38
Language Density					
MCT	1044	.902	.961**	27.97	70.69
LCT	86	.935	.948*	31.40	53.49
Language Difficulty					
Category I	864	.884	.954**	28.59	70.95
Category II	36	.918	.966**	27.78	69.44
Category III	71	.872	.939**	30.99	60.56
Category IV	159	.926	.957**	25.16	65.78
Language					
English	160	.827	.921**	33.13	64.38
Mandarin	33	.900	.963**	18.18	78.79
French	96	.920	.925	46.88	53.13
German	36	.918	.966**	27.78	69.44
Italian	57	.832	.909**	35.09	64.91
Japanese	62	.913	.969**	24.19	72.58
Russian	68	.861	.935**	30.88	60.29
Spanish	532	.892	.968**	22.93	77.07
Arabic	61	.946	.948	31.15	47.54
Portuguese	19	.861	.932**	36.84	63.16
Cantonese	3	.999	.999**	0.000	100.0

Note. Raters 1 and 2 required, while rater 3 serves as a "tie-breaker" for disagreements; *r* = Pearson correlation; * denotes significant difference between correlations ($p < .05$), one-tailed; ** denotes significant difference ($p < .01$), one-tailed; MCT = more commonly tested; LCT = less commonly tested.

not significantly different (French and Arabic). Also shown in Table 4 are the raw percentages of absolute agreement for the two rater combinations. These agreement indices further emphasize the much larger levels of interrater consistency between the second and third rater combination as compared to the first and third rater combination.

Discussion

To place the consistency and agreement estimates found in the present study in perspective, two types of comparisons can be made: (1) general comparisons to "acceptable" levels of reliability derived from the educational testing and psychometric literature; and (2) specific comparisons to reliability levels found in previous research examining OPI raters. Regarding the first type of comparison, Nunnally (1978) suggested that an acceptable reliability for preliminary research is .70. Kaplan and Saccuzzo (1997) and Nunnally and Bernstein (1994) recommended a reliability benchmark of .80 for purposes of basic research and .90 to .95 for any applied research ventures. The estimates of interrater consistency found in this study were all above the recommendation for applied projects (.90). Moreover, none of the interrater reliabilities fell at or below the .70 level, which Murphy and Davidshofer (1994) call a "low level" of reliability.

As for the second type of comparison, the consistency estimates found within the present study were similar to, but generally higher than, estimates found in previous OPI research. For instance, using Pearson's *r*, Magnan (1987) found an interrater reliability of .94 for trainer-trainee ratings of French speaking proficiency. Dandonoli and Henning (1990) reported interrater reliabilities of .98 and .97 for English and French, respectively. Finally, in a study of proficiency ratings of English, French, German, and Spanish, Thompson (1995) found interrater reliabilities that ranged from .83 to .89. The reliability results presented herein for these languages ranged from .96 to .98. The improved reliability most likely results from rater training and higher levels of experience in the cadre of raters.

Combining the present study's results with similar findings from previous research provides evidence that bolsters

confidence in, and the generalizability of, the relatively high level of interrater reliability and consistency demonstrated by experienced ACTFL OPI interviewers. Moreover, the present study included more languages, a larger sample of raters, and a more comprehensive approach than previous studies.

Important as well is that this study directly addresses the *Standards for Educational and Psychological Testing* (AERA, 1999) related to reporting reliability evidence to users of an assessment. Overall, the results provide good news for those who use the ACTFL OPI to make decisions about speaking proficiency in the 19 languages examined in our study. We recommend that ACTFL conduct and publish the results of interrater consistency and agreement analyses every three to five years to continue to meet the guidelines established by the Standards. This becomes particularly salient for those languages, such as Tagalog and Albanian, that contained small samples sizes. Furthermore, we openly encourage other providers of OPI assessment to follow this suggestion as well.

Research questions 2 through 5 were included to provide additional information related to the functioning of the ACTFL interview protocol under the Revised Guidelines for speaking proficiency. If interrater reliability and agreement are not affected by these other characteristics (e.g., language difficulty), then these findings provide additional evidence that the rating scale and protocols are functioning as intended and with reasonable precision. Question 2 addresses whether interrater consistency remains similar when testing frequency in a language is considered. Since the results indicate that the levels of interrater consistency by testing density are virtually identical across the categories (MCT and LCT), we can conclude that the protocol is not unduly affected by the density of testing. If it had been, the reasons for this difference would be a point for future investigation.

Research question 3 addresses the issue of whether the difficulty level of the language tested (in terms of language learning) has an impact on the reliability of the OPI ratings. Again, the results indicate that language difficulty has no significant impact on interrater consistency and agreement, suggesting that this is not an issue for further investigation.

Research question 4 investigates interrater reliability within each language. For the 19 languages in this study, the interrater consistency and agreement results were above the acceptable range and very consistent across the different indices. The results for two languages, Italian and Arabic, were slightly lower than for the majority of the languages. With the data available to us, we were unable to empirically determine why this might be the case. A number of different factors could be affecting the consistency and agreement of the ratings, including characteristics of the raters, characteristics of the ratees, characteristics of the language or dialects, or the interaction of these factors. Of course, it could be a case of simply having aberrant raters who are in need of more training. We recommend that ACTFL investigate this issue by examining the most likely factors. Investigating the functioning of individual raters and pairs of raters would a good place to start.

Question 5 addresses whether or not the interrater agreement results vary across the major proficiency levels (Novice, Intermediate, Advanced, and Superior) and the nature of the disagreement (i.e., within or between major proficiency levels). Unlike our expectations for questions 2 through 4, we expected question 5 to demonstrate a difference between proficiency categories. Whenever subjective ratings are being made across a continuum and rater agreement is calculated, the highest agreement between raters should be expected for the extreme scale points or values, as the extremities of performance are generally the easiest to detect and consequently rate. Our results demonstrate this pattern because Novice-Low and Superior have the highest percentage of absolute agreement. In terms of the nature of the disagreements, virtually all of the disagreements were within one scale point or step (e.g., Novice-Low versus Novice-Mid or Novice-High versus Intermediate-Low). The majority of the disagreements (58.5%) were within the same major level, and the disagreements that crossed a major level boundary were spread across the three boundaries. Additionally, no disagreements spanned two major proficiency category boundaries. Overall, the results of research questions 2 through 5 provide additional evidence that the ACTFL rating scale and protocols are functioning as intended. This further bolsters our confidence in the reliability of the ACTFL OPI procedure.

The results for question 6 demonstrate that when the first and second raters disagree, the third rater has a tendency to "break the tie" more often in the favor of the second rater. This finding held across languages and characteristics in this study. The findings are quite robust on this point, as is apparent in the absolute agreement indices found in Table 4. As noted earlier, second and third raters always rate from the audiotape without having telephonic contact with the ratee, whereas the first rater conducts the interview and then rates from the audiotape at a later time. This could explain the results for question 6. Several factors could be driving this effect. The important question is whether conducting the interview as well as rating it affects the psychometric characteristics of the assessment, especially the validity of the assessment. However, given the high initial agreement between the first and second raters, there may be no impact of the differential roles of testers (i.e., interviewing and rating as opposed to rating only) on the overall validity at all. The findings from question 6 could be a function of the specific raters involved in the disagreements, not a function of the role differences. Therefore, research should be conducted to determine if the validity is affected and why.

In general, the ratings generated by any of the OPI interview procedures should validly and reliably measure the construct of interest (language speaking proficiency) and describe proficiency regardless of the testing mode (e.g., in-person, telephonic, or video conferencing) and whether or not the rater conducted the interview. In other words, mode and rater role (i.e., rater only versus interviewer and rater) should not affect the ratings assigned to the interviewee's proficiency by the raters. When assessment procedures depend on human

judgments, every effort should be made to maintain rater independence and reduce the interaction of rater–ratee characteristics and of rater–rater characteristics that might bias or contaminate the ratings. "Criterion contamination" refers to the condition when an assessment produces scores or ratings that measure other constructs or factors beyond the one of interest and constitutes a major threat to validity. Additionally, criterion contamination does not necessarily have an impact on reliability—in other words, a process that produces a biased or contaminated score can be reliable. Although not an issue for the ACTFL OPI, the interaction of rater–rater characteristics can be pertinent for procedures whereby both raters are simultaneously present and participate in the interview together. This could potentially undermine the independence of the ratings even when the protocol makes an effort to have the raters separately judge the proficiency prior to discussing the interview. In light of this information, research into this issue is justified and needed.

The results of question 6 suggest a potential OPI process change to be studied. We recommend that providers of OPI assessments, regardless of testing modality, research and evaluate moving to a protocol with differentiated roles in which there are *interview specialists* and *rating specialists*. Differentiation of the interview and rating roles would definitely eliminate any bias that may be introduced into the measurement system by the first tester conducting and rating the interview. The interviewer would elicit the best sample of speaking performance, and two independent raters would rate the audio or video record of the interview. To ensure a ratable sample, the training, evaluation, and compensation of the interview specialists would need to be aligned with the new process. However, the downside to this suggestion is that separating interviewer and rater roles will likely increase the cost of the OPI. Therefore, research should be conducted to determine and evaluate the impact of the modification. We suspect that validity and reliability would be improved because the process of conducting the interview likely interacts with certain rater characteristics to influence (or bias) the ratings. However, before the process change is executed, research should determine if construct-related validity and reliability are significantly improved through role differentiation. If they are not significantly improved, then the increased cost is clearly not justified. Additionally, if significant improvements are found from role differentiation and the underlying mechanisms affecting the ratings are discovered, the same improvements might be achieved by modifying rater training without the need for role differentiation. After data are available, ACTFL should be able to weigh the cost effectiveness of the options. Finally, given the high interrater reliability and agreement between the first and second raters, this suggestion should be viewed as an interesting research question and a potential improvement, not as a necessity.

Before making our recommendations for future research, we should acknowledge some limitations of this study. First, some languages in our study had a small number of cases, and results for these languages should be viewed with caution. Second, the data did not include characteristics of raters and ratees; therefore, we could not test for the influences of individual differences characteristics like race, gender, age, education, and length and breadth of testing experience. Third, the data did not include the testing context (i.e., employment, academia, etc.), and this would have allowed us to assess consistency and agreement of specific testing contexts. Finally, we did not assess the functioning of individual raters and rating combinations (i.e., pairs of raters) because it was beyond the scope of this article.

In addition to the recommendations made throughout this section, we recommend that ACTFL and language researchers consider the following future research studies:

1. A meta-analysis of reliability and validity across all types speaking proficiency assessments (e.g., ACTFL OPI, DLI OPI, and so forth).
2. A reliability assessment of individual raters (including intrarater reliability) and rating combinations across all languages tested with the ACTFL OPI.
3. A study where relevant ratee and rater characteristics are collected to determine if their interaction affects rating validity and reliability.
4. A study of the consistency of OPI ratings over repeated measures (with the same ratees) within a time frame in which learning might not be expected to be a factor.
5. A study to determine whether testing mode and rater role affect the ratings (see above).
6. A validity study capturing the policy and mental models of raters.
7. A series of construct-related validity studies with the Revised Guidelines for different test uses and testing contexts.
8. A series of criterion-related validity studies (both predictive and concurrent) to determine the validity of the ACTFL in relation to relevant criteria such as job performance.
9. A longitudinal study of language proficiency in academic and work contexts using methods such as latent growth modeling.

These are only a few research recommendations. In general, we suggest that ACTFL and other language researchers collaborate to address issues related to proficiency measurement in language learning and job-related language performance. It is important to note that claims about any assessment cannot be substantiated without a robust body of empirical evidence. This evidence can only come from well-designed research.

To conclude, we strongly reiterate that the results of our study are very positive for users of the ACTFL OPI and support its reliability as an assessment of speaking proficiency. Our study provides the most comprehensive investigation to date of interrater consistency and agreement for the ACTFL OPI; therefore, extending the reliability evidence available to test users and researchers. These results demonstrate the

importance of having an OPI assessment program that has a well-designed interview process, well-articulated criteria for rating, a solid rater training program, and an experienced cadre of testers.

Based on the data reported, educators and employers who use the ACTFL OPI can expect reliable results and use the scores generated from the process with increased confidence. In terms of future research, we encourage ACTFL and language researchers to conduct validity studies to ensure that revision of the Guidelines did not adversely affect assessment validity and to satisfy the recommendations of the *Standards for Educational and Psychological Testing* (AERA, 1999). This is one of the most pressing research needs in measuring language speaking proficiency.

Acknowledgments

The authors thank Ray Clifford, Helen Hamlyn, Ward Keesling, and Elvira Swender for their assistance with and comments related to this article.

References

Adams, M. (1978). Measuring foreign language speaking proficiency: A study of agreement among raters. In J. L. D. Clark (ed.), *Direst testing of speaking proficiency: Theory and application* (pp. 131–49). Princeton, NJ: Educational Testing Service.

American Council on the Teaching of Foreign Language (1986). *ACTFL proficiency guidelines*. Yonkers, NY: Author.

American Council on the Teaching of Foreign Language (2002). *ACTFL oral proficiency interview tester certification information application packet*. Yonkers, NY: Author.

American Educational Research Association (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

Bachman, L. F., & Palmer, A. S. (1981). The construct validity of the FSI oral interview. *Language Learning*, *31*, 67–86.

Breiner-Sanders, K. E., Lowe, P., Miles, J., & Swender, E. (2000). ACTFL proficiency guidelines—Speaking, revised 1999. *Foreign Language Annals*, *33*, 13–18.

Carroll, J. B. (1967). Foreign language proficiency levels attained by language majors near graduation from college. *Foreign Language Annals*, *1*, 131–51.

Cattell, R. B. (1988). The meaning and strategic use of factor analysis. In R. B. Cattell & J. R. Nesselroade (eds.), *Handbook of multivariate experimental psychology: Perspectives on individual differences*, 2nd ed. (pp. 131–203). New York: Plenum Press.

Clark, J. D. L. (1986). *A study of the comparability of speaking proficiency interview ratings across three government language training agencies*. Washington, DC: Center for Applied Linguistics.

Dandonoli, P., & Henning, G. (1990). An investigation of the construct validity of the ACTFL Oral Proficiency Guidelines and Oral Interview Procedure. *Foreign Language Annals*, *23*, 11–22.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (ed.), *Educational measurement*, 3rd ed. (pp. 105–46). Washington, DC: American Council on Education.

Flanagan, J. C. (1951). Units, scores, and norms. In E. F. Lindquist (ed.), *Educational measurement* (pp. 695–763). Washington, DC: American Council on Education.

Jackson, G. L. (1999). *Oral proficiency testing modality study* (DLIFLC Research Report No. 99-01). Monterey, CA: Defense Language Institute Foreign Language Center.

Kaplan, R. W., & Saccuzzo, D. P. (1997). *Psychological testing: Principles, applications, and issues*. 4th ed. Belmont, CA: Brooks and Cole.

Magnan, S. S. (1986). Assessing speaking proficiency in the undergraduate curriculum: Data from French. *Foreign Language Annals*, *19*, 429–38.

Magnan, S. S. (1987). Rater reliability of the ACTFL Oral Proficiency Interview. *The Canadian Modern Language Review, 43*, 267–76.

Meng, X. L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin*, *111*, 172–75.

Murphy, K. R., & Davidshofer, C. O. (1994). *Psychological testing: Principles and applications*. 3rd ed. Englewood Cliffs, NJ: Prentice-Hall.

Nunnally, J. C. (1978). *Psychometric theory,* 2nd ed. New York, NY: McGraw Hill Book Company.

Nunnally, J. C., Bernstein, I. H. (1994). *Psychometric theory*, 3rd ed. New York, NY: McGraw Hill Book Company.

Stanley, J. C. (1971). Reliability. In R. L. Thorndike (ed.), *Educational measurement*, 2nd ed. (pp 356–442). Washington, DC: American Council on Education.

Stansfield, C. W., & Kenyon, D. M. (1992). Research on the comparability of the Oral Proficiency Interview and the Simulated Oral Proficiency Interview. *System*, *20*, 347–64.

Steiger, J. H. (1980). Test for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*, 245–51.

Swender, E. (ed.) (1999). *ACTFL oral proficiency interview tester training manual*. Yonkers, NY: ACTFL.

Thompson, I. (1995). A study of interrater reliability of the ACTFL Oral Proficiency Interview in five European languages: Data from English, French, German, Russian, and Spanish. *Foreign Language Annals*, *28*, 407–22.

Thorndike, R. L. (1951). Reliability. In E. F. Lindquist (ed.), *Educational measurement* (pp. 560–620). Washington, DC: American Council on Education.