Reliability Study of the ACTFL OPI® in Chinese, Portuguese, Russian, Spanish, German, and English for the ACE Review

Prepared for:

American Council on the Teaching of Foreign Languages (ACTFL)
White Plains, NY

Prepared by
SWA Consulting Inc.

801 Jones Franklin Road
Suite 270
Raleigh, NC 27606

919.835.1562

http://www.swa-consulting.com

EXECUTIVE SUMMARY

This report documents the inter-rater reliability and agreement of the American Council on the Teaching of Foreign Languages (ACTFL) Oral Proficiency Interview (OPI®) from January 2009 to December 2011 to satisfy a review requirement of the American Council on Education College Credit Recommendation Service (CREDIT) program. The ACTFL OPI® is an assessment of functional speaking proficiency in a foreign language, which is delivered by trained and certified raters in an interview format across numerous languages. Comparisons of ACTFL OPI® inter-rater reliability and agreement were made across six languages: Chinese, Portuguese, Russian, Spanish, German, and English. Comparisons were also made across language categories (i.e., language difficulty) and interview years (i.e., 2009, 2010, and 2011 in this sample). For inter-rater agreement, concordance was further investigated by major proficiency level and sub-level.

METHOD

Given the ordinal nature of the ACTFL proficiency scale and ACTFL OPI® scores, inter-rater reliability was measured by the Spearman's *R* correlation, which is a coefficient of reliability appropriate for ordinal data. Inter-rater agreement was measured by the extent to which ratings exhibited absolute (i.e., exact) and/or adjacent (i.e., +/- one level) agreement. The combination of Spearman's *R* and absolute/adjacent agreement results provides sufficient information about reliability.

FINDINGS

Overall, the ACTFL OPI® exceeded the minimum inter-rater reliability and agreement standards. Further, the findings are consistent with results from Surface and Dierdorff (2003), indicating the ACTFL OPI® process yields relatively stable reliability results over time.

- Inter-rater Reliability
  - Spearman *R*s exceeded the standard for use, ranging from .95 to .98 across languages.
  - Inter-rater reliability was similar across language category and interview year.

- Inter-rater Agreement
  - Overall absolute agreement was higher than 70% for all languages.
  - Absolute agreement was similar across language categories and years.
  - Absolute agreement was greatest for Superior and Novice Low.

Overall, the findings support the reliability of the ACTFL OPI® as an assessment of speaking proficiency. Areas for continued improvement include increasing rater agreement within the Advanced level and the Novice High-Intermediate Low border. Findings are presented in more detail in the report.

TABLE OF CONTENTS

Reliability Study of the ACTFL OPI® in Chinese, Portuguese, Russian, Spanish, German, and English for the ACE Review

SECTION 1: PURPOSE

Test developers have a responsibility to demonstrate the effectiveness of their assessments by investigating and documenting their measurement properties (AERA, APA, & NCME, 1999). Among the fundamental measurement properties that should be documented is reliability, which refers to the consistency of test scores. 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 (AERA, APA, & NCME, 1999). Various methods are used to calculate and estimate reliability depending on the test type and purpose. This report documents the inter-rater reliability and agreement of the ACTFL Oral Proficiency Interview (OPI®) assessment, which is an assessment of functional speaking proficiency using an interview format rated by trained and certified experts. This report satisfies a review requirement of the American Council on Education CREDIT program. Inter-rater reliability and agreement were calculated across six interview languages—Chinese (Mandarin), Portuguese, Russian, Spanish, German, and English—and across three years—2009 through 2011. For inter-rater agreement, concordance was further investigated by major proficiency level and sub-level.

This report is divided into five total sections. Section 2 provides background on the ACTFL OPI®, a review of the American Council on Education (ACE) process, previous inter-rater reliability and agreement research on the ACTFL OPI®, and the primary research questions addressed in this report. Section 3 describes the methods, and Section 4 summarizes the results of the current study. Finally, Section 5 presents interpretations and conclusions based on these results. References are provided at the end of the report. Any questions about this report and study should be directed to Dr. Eric Surface (esurface@swa-consulting.com).

SECTION 2: BACKGROUND

THE ACTFL OPI®

The ACTFL OPI® is a live interview conducted telephonically between an ACTFL Certified OPI® Tester and the individual whose language proficiency is being assessed. The interview lasts between 20 and 30 minutes depending on the proficiency level of the test taker. A ratable sample is elicited through a series of personalized questions that adhere strictly to a standardized elicitation protocol designed to establish the speaker's highest level of sustained ability as well as the level at which the speaker is no longer able to sustain all the assessment criteria for the level. The elicited speech sample is digitally recorded and rated by the tester according to a standardized ACTFL OPI® rating protocol. The elicited speech sample is then compared to the descriptors contained in the ACTFL Proficiency Guidelines – Speaking and a rating is assigned. Each sample is independently rated by a minimum of two ACTFL Certified Testers. The two ratings must agree exactly. Any rating discrepancy is arbitrated by a third Certified Tester and an Official ACTFL OPI® rating is assigned when two ratings agree exactly.

ACE PROCESS

The American Council on Education (ACE) aims to foster greater collaboration and new partnerships within and outside the higher education community to help colleges and universities anticipate and address the challenges of the 21st century and contribute to a stronger nation and a better world. ACE is the major coordinating body for all the nation's higher education institutions. Among the missions of ACE is the commitment to support the advancement of adult learners through the Center for Lifelong Learning. One way in which the Center addresses this objective is through the CREDIT, a quality evaluation that translates professional workplace learning into college credit recommendations.

For over 30 years, ACE CREDIT has successfully worked with thousands of corporate learning programs offered by businesses and industry, labor unions, associations, government agencies and military services. The CREDIT recommendations are designed to provide adult learners with the opportunity to receive academic credit for courses completed outside the traditional university classroom. The ACE CREDIT recommendation carries benefits for each of the program's three participants: the Organization, the Adult Learner, and the Postsecondary Institution.

This report was commissioned to satisfy ACE CREDIT review requirements.

PREVIOUS ACTFL OPI® RESEARCH ON INTER-RATER RELIABILITY AND INTER-RATER AGREEMENT

The reliability of the ACTFL OPI® has generally been well supported. In the first published study, Magnan (1986) reported a Cohen's Kappa of .72 for a sample of 40 French students rated by two ACTFL-certified testers from the Educational Testing Service. All ratings had either absolute (i.e., were exactly the same) or adjacent (i.e., were off by only one sublevel) agreement. Subsequent studies provided further evidence for the reliability of the ACTFL OPI® (e.g., Dandonoli & Henning, 1990; Thompson, 1995; Thompson, 1996). In a comprehensive study assessing ACTFL OPI® reliability across 19 languages, Surface and Dierdorff (2003) found the inter-rater reliability to be very high. Spearman *R*s for the 19 languages ranged from .938 to .999. Absolute agreement exceeded the minimum standard for operational use (i.e., 70%) for all languages, with the exception of Arabic (which met the adjacent agreement standard). All languages met or exceed the adjacent agreement standard. In response to the Arabic findings, ACTFL retrained the certified ACTFL OPI® testers in Arabic.

RESEARCH QUESTIONS

This report addresses research questions related to the inter-rater reliability and inter-rater agreement of the ACTFL OPI®. These research questions are:

1. What is the inter-rater reliability of the ACTFL OPI® in Chinese, Portuguese, Russian, Spanish, German, and English?

2. Are there any differences in overall ACTFL OPI® inter-rater reliability levels by language category[1] and assessment year (2009-2011)?

3. What is the inter-rater agreement of the ACTFL OPI® in Chinese, Portuguese, Russian, Spanish, German, and English?

4. Are there any differences in overall ACTFL OPI® inter-rater agreement levels by language category, assessment year (2009-2011), and proficiency level?

[1] Language category is a proxy for language difficulty (Surface & Dierdorff, 2003). Given the languages in the study—only one language per categories II, III and IV—we decided to aggregate and analyze as Categories I/II and III/IV.

SECTION 3: METHOD

Reliability is an important psychometric property that all assessments should demonstrate (Flanagan, 1951; Thorndike, 1951; Stanley, 1971; Anastasi, 1988; Cattell, 1988). 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. The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) provides a number of guidelines designed to help test administrators evaluate the reliability data provided by test publishers. The level of reliability evidence that is necessary to assess and to be reported depends on the purpose of the test or assessment procedure. Reliability is particularly important because it can limit the validity of an assessment.

For assessments like the ACTFL OPI®, which uses raters, one of the most commonly used forms of reliability estimation is inter-rater reliability, which reflects the overall level of consistency among the raters. When inter-rater reliability estimates are high, it suggests a large degree of consistency across the raters. Raters must yield reliable measurements in order for the data to be useful. Data that are unreliable contain error, and decisions based on such data should be made with caution.

There are two types of inter-rater reliability evidence for rater-based assessments—inter-rater reliability coefficients and inter-rater agreement (concordance of ratings). Although there are many types of reliability analyses, the choice of specific technique should be governed by the nature and purpose of the assessment and its data. Also, simplicity is desired in communicating results to laypeople.

Inter-rater Reliability: Spearman's Rank Order Correlation

Spearman's rank-order correlation (R) is a commonly used correlation for assessing inter-rater reliabilities, and correlations should be at or above .70 to be considered sufficient for test development and .80 for operational use (e.g., LeBreton et al., 2003). Spearman's R is the most appropriate statistic for evaluation of the ACTFL OPI® data because the proficiency categories used for OPI® ratings are ordinal in nature.

Spearman's rank-order correlation is another commonly used correlation for assessing inter-rater reliability, particularly in situations involving ordinal variables. Spearman rank-order correlation (R) has an 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.

Inter-rater Agreement: Absolute and Adjacent Agreement

Another common approach to examining reliability is to use measures of inter-rater agreement. Whereas inter-rater reliability assesses how consistently the raters rank-order test-takers, inter-rater agreement assesses the extent to which raters give the same score for a particular test-taker. Since rating protocol assigns final test scores based on agreement (concordance) between raters rather than rank-order consistency, it is important to assess the degree of interchangeability in ratings for the same test taker. Inter-rater reliability can be high when inter-rater agreement is low, so it is important to take both into account when assessing a test.

Inter-rater agreement can be assessed by computing absolute agreement between rater pairs (i.e., whether both raters provide exactly the same rating). Standards for absolute agreement vary depending on the number of raters involved in the rating process. When two raters are utilized, there should be absolute agreement between raters more than 80% of the time, with a minimum of 70% for operational use (Feldt & Brennan, 1989). Absolute agreement closer to 100% is desired, but difficult to attain. Each additional rater employed in the process decreases the minimum acceptable agreement percentage. This accounts for the fact that agreement between more than two raters is increasingly difficult. Adjacent agreement is also assessed in this reliability study. Adjacent agreement occurs when raters are within one rating level in terms of their agreement (e.g., rater 1 gives a test taker a rating of Intermediate Mid and rater two gives a rating of Intermediate Low). In the ACTFL process, when there is not absolute agreement, an arbitrating third rater will provide a rating that resolves the discrepancy. Some foreign language proficiency interviews use an adjacent agreement standard and award the lower of the two adjacent ratings, which is different and not as rigorous as the ACTFL process.

Language Categories

ACTFL OPI® inter-rater reliability and agreement results are also reported across language difficulty levels. According to a categorization used by the US Government, a language is assigned to a category based on how difficult it is for a native English speaker to learn that language. Categories are distinguished by numerals, which range from I to IV. More difficult languages are assigned to categories with higher numerals (Category IV being the most difficult). Spanish and Portuguese are assigned to Category I; German is in Category II; Russian is in Category III; and Chinese (Mandarin) is in Category IV. For simplicity in reporting, English was included in Category I. For the purposes of this report, Categories I and II were collapsed into a single category: Category I/II (Spanish, Portuguese, English, and German). Categories III and IV were also collapsed into a single category: Category III/IV (Russian and Chinese).

SECTION 4: RESULTS

Research Question 1 - *What is the inter-rater reliability of the ACTFL OPI® in Chinese, Portuguese, Russian, Spanish, German, and English?*

Inter-rater reliability was calculated per language using Spearman's *R*. The correlation coefficient indicates the level of consistency between raters and should be at or above .70 to be considered sufficient for test development and above .80 for operational use (LeBreton et al., 2003). Coefficients closer to 1.00 are preferred. As shown in Table 1, all Spearman's *R* coefficients were statistically significant and exceeded the .80 standard, demonstrating high inter-rater reliability. Inter-rater reliability differed little across languages, as indicated by the small range of the correlations (0.953 to 0.983). The current results are highly consistent with Surface & Dierdorff (2003).

Table 1
Spearman's Correlations by Language

	Current Study		**Surface & Dierdorff (2003)**	
	N	***R***	***N***	***R***
Chinese	2907	.983	241	.989
Portuguese	476	.965	111	.976
Russian	1115	.976	278	.966
Spanish	8691	.964	2777	.970
German	647	.967	216	.976
English	3986	.953	725	.957
Overall	17823	.971	5881	.976

Note: All correlations in the current study were significant at the *p* > .000 level. The current study differed from the Surface and Dierdorff's study (2003) in the total number of languages. Therefore, the reader should use caution when comparing the overall inter-rater reliability between the studies.

Research Question 2 - *Are there any differences in overall ACTFL OPI® inter-rater reliability levels by language category and assessment year (2009-2011)?*

As shown in Table 2, the results by language category were all above the .80 standard, demonstrating high inter-rater reliability. Again, because Categories II, III & IV had only one language each, we chose to aggregate Category I & II languages and Category III & IV languages for analysis. Although Surface & Dierdorff (2003) reported results for each individual language category, the current results are very consistent with the previous findings.

Spearman's *R* coefficients for interview years 2009, 2010, and 2011 were calculated on the sample aggregated across all languages to determine if year had an overall impact. As shown in

Table 3, all correlations exceeded the 0.80 standard, demonstrating high inter-rater reliability across all three years in the current study. Inter-rater reliability was nearly identical across years.

Table 2
Spearman's Correlations by Language Category

	N	**Spearman's *R***	
		R	***p***
Category I/II	13801	.968	.000
Category III/IV	4022	.982	.000

Table 3
Spearman's Correlations by Year

	N	**Spearman's *R***	
		R	***p***
2009	5864	.966	.000
2010	6676	.969	.000
2011	5283	.979	.000

<u>Research Question 3</u> - *What is the inter-rater agreement of the ACTFL OPI® in Chinese, Portuguese, Russian, Spanish, German, and English?*

Both absolute and adjacent agreements were calculated for each language. As shown in Table 4, absolute agreement for all languages exceeded 70% (78% to 83%), indicating a fairly high level of concordance between raters (minimum standard for use is 70%). Further, absolute and adjacent agreements were similar across languages. When absolute and adjacent percentages are added, 98% to 100% of cases (depending on language) fall within the union of the sets.

Table 4
Absolute/Adjacent Agreement by Language

	N	**Absolute Agreement (exact)**	**Adjacent Agreement (+/- 1)**	**None (+/- 2)**
Chinese	2907	81%	19%	1%
Portuguese	476	83%	17%	1%
Russian	1115	79%	19%	2%
Spanish	8691	78%	22%	0%
German	647	81%	19%	1%
English	3986	81%	19%	1%
Overall	17823	79%	20%	1%

Note. Percentages are rounded to the nearest whole number, and thus may not always add up to 100%.

<u>Research Question 4</u> - *Are there any differences in overall ACTFL OPI® inter-rater agreement levels by language category, assessment year (2009-2011), and proficiency level?*

Both absolute and adjacent agreements were calculated for Category I/II and Category III/IV languages. As shown in Table 5, absolute agreement was satisfactory (and nearly identical) for Category I/II and Category III/IV languages. This is consistent with previous research.

Table 5
Absolute/Adjacent Agreement by Language Category

	N	**Absolute Agreement (exact)**	**Adjacent Agreement (+/- 1)**	**None (+/- 2)**
Category I/II	13801	79%	21%	0%
Category III/IV	4022	80%	19%	1%

Note. Percentages are rounded to the nearest whole number, and thus may not always add up to 100%.

Both absolute and adjacent agreements were calculated for each interview year (i.e., 2009, 2010, and 2011) for sample aggregated across all languages to determine if year impacted agreement. As shown in Table 6, absolute agreement was above the minimum threshold for operational use (i.e., 70%) for all years and was greatest in 2011. The trend shows improvement across the years.

Table 6
Absolute/Adjacent Agreement by Year

	N	**Absolute Agreement (exact)**	**Adjacent Agreement (+/- 1)**	**None (+/- 2)**
2009	5864	76%	23%	1%
2010	6676	78%	22%	1%
2011	5284	84%	16%	0%

Note. Percentages are rounded to the nearest whole number, and thus may not always add up to 100%.

Both absolute and adjacent agreements were calculated for each major proficiency level. As shown in Table 7, absolute agreement was above 70% for all major proficiency levels. Consistent with findings from Surface and Dierdorff (2003), absolute agreement was notably higher for the Superior proficiency level.

Table 7
Absolute/Adjacent Agreement by Major Proficiency Level

	N	Absolute Agreement (exact)	Adjacent Agreement (+/- 1)	None (+/- 2)
Novice	570	78%	22%	1%
Intermediate	5712	78%	22%	1%
Advanced	7465	74%	25%	1%
Superior	4076	90%	9%	0%

Note. Percentages are rounded to the nearest whole number, and thus may not always add up to 100%.

Both absolute and adjacent agreements were calculated for each proficiency sublevel. As shown in Table 8, absolute agreement was above the threshold of 70% for all proficiency levels (73% to 90%). Consistent with Surface and Dierdorff (2003), the highest agreement occurred at the extreme ends of the proficiency scale. That is, agreement was highest for the Superior level proficiency (90%) and the Novice Low sublevel proficiency (86%).

Table 8
Absolute/Adjacent Agreement by Sublevel Proficiency

	N	Absolute Agreement (exact)	Adjacent Agreement (+/- 1)	None (+/- 2)
Novice Low	29	86%	14%	0%
Novice Mid	219	82%	18%	0%
Novice High	322	74%	25%	1%
Intermediate Low	663	73%	26%	1%
Intermediate Mid	2075	78%	21%	0%
Intermediate High	2974	79%	21%	1%
Advanced Low	2409	73%	27%	1%
Advanced Mid	2879	75%	24%	1%
Advanced High	2177	75%	25%	1%
Superior	4076	90%	9%	0%

Note. Percentages are rounded to the nearest whole number, and thus may not always add up to 100%.

SECTION 5: INTERPRETATIONS AND CONCLUSIONS

Overall, the ACTFL OPI® exceeded inter-rater reliability and inter-rater agreement minimum standards. Overall, the inter-rater reliability was quite high (*R*=.97). The Spearman's *R* correlations ranged from .95 to .98 across all the six languages. Inter-rater reliability was similar across language categories and interview year. There was evidence of acceptable inter-rater agreement for operational use. Absolute agreement was higher than 70% for all comparisons and the overall agreement level was just below 80%. Absolute agreement was similar across interview language, language category and interview year. The highest agreement occurred at the extreme ends of the proficiency scale. That is, agreement was highest for the Superior proficiency level (90%) and the Novice Low proficiency sublevel (86%). Overall, the reliability evidence in the current study is consistent with previous findings (e.g., Surface & Dierdorff, 2003) and supports the operational use of the ACTFL OPI® in Chinese (Mandarin), Portuguese, Russian, Spanish, German, and English. Areas for continued improvement include increasing rater agreement within the Advanced level and the Novice High-Intermediate Low border.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, DC: Author.

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

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.

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.

LeBreton, J. M., Burgess, J. R. D., Kaiser, R. B., Atchley, E. K., & James, L. R. (2003). The restriction of variance hypothesis and inter-rater reliability and agreement: Are ratings from multiple sources really dissimilar? *Organizational Research Methods, 6*(1), 80-128.

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

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

Surface, E. A., & Dierdorff, E. C. (2003). Reliability and the ACTFL Oral Proficiency Interview: Reporting indices of interrater consistency and agreement for 19 languages. *Foreign Language Annals, 36*, 507-519.

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.

Thompson, I. (1996). Assessing foreign language skills: Data from Russian. *Modern Language Journal, 80,* 47-65.

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

ABOUT SWA CONSULTING INC.

SWA Consulting Inc. (formerly Surface, Ward, and Associates) provides analytics and evidence-based solutions for clients using the principles and methods of industrial/organizational (I/O) psychology. Since 1997, SWA has advised and assisted corporate, non-profit and governmental clients on:

- Training and development
- Performance measurement and management
- Organizational effectiveness
- Test development and validation research
- Program/training evaluation
- Work/job analysis
- Needs assessment
- Selection system design
- Study and analysis related to human capital issues
- Metric development and data collection
- Advanced data analysis

One specific practice area is analytics, research, and consulting on foreign language and culture in work contexts. In this area, SWA has conducted numerous projects, including language assessment validation and psychometric research; evaluations of language training, training tools, and job aids; language and culture focused needs assessments and job analysis; and advanced analysis of language research data.

Based in Raleigh, NC, and led by Drs. Eric A. Surface and Stephen J. Ward, SWA now employs close to twenty I/O professionals at the Masters and PhD levels. SWA professionals are committed to providing clients the best data and analysis upon which to make evidence-based decisions. Taking a scientist-practitioner perspective, SWA professionals conduct model-based, evidence-driven research and consulting to provide the best answers and solutions to enhance our clients' mission and business objectives.

For more information about SWA, our projects, and our capabilities, please visit our website (www.swa-consulting.com) or contact Dr. Eric A. Surface (esurface@swa-consulting.com) or Dr. Stephen J. Ward (sward@swa-consulting.com).

The following SWA Consulting Inc. team members contributed to this report (listed in alphabetical order):

Mr. Hyder Abadin
Mr. David Fried
Ms. Gwendolyn Good
Dr. Eric Surface