Reliability Study of ACTFL OPIc® in Spanish, English, and Arabic 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) computerized version of the Oral Proficiency Interview (OPIc®) 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 OPIc® is an assessment of functional speaking proficiency in a foreign language, which is delivered in a computer-based format. Comparisons of ACTFL OPIc® inter-rater reliability and agreement were made across three languages: Spanish, English and Arabic. 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, rater concordance was further investigated by major proficiency level and sub-level.

METHOD

Given the ordinal nature of the ACTFL proficiency scale and ACTFL OPIc® 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 OPIc® exceeded the minimum inter-rater reliability and agreement standards.

- Inter-rater Reliability
  - Spearman *R*s exceeded the standard for use, ranging from .95 to .97 across languages.
  - Inter-rater reliability was similar across language category and interview year.

- Inter-rater Agreement
  - Absolute agreement was higher for Spanish and English (80% for both) than it was for Arabic (71%).
  - The Arabic data showed an improving trend in agreement from 2009 (28%; *N*=7) to 2011 (73%; *N*=152).

Overall, the findings support the reliability of the ACTFL OPIc® as an assessment of speaking proficiency. Areas for continued improvement include increasing rater agreement within the Advanced proficiency levels and increasing rater agreement for the Arabic ACTFL OPIc®. Findings are presented in more detail in the report.

TABLE OF CONTENTS

Reliability Study of ACTFL OPIc® in Spanish, English, and Arabic 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 American Council on the Teaching of Foreign Languages (ACTFL) Oral Proficiency Interview - computer (OPIc®) assessment, which is an assessment of functional speaking proficiency in an internet-delivered interview format (embodied agent is the interviewer) and subsequently 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 three languages—Spanish, English and Arabic—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 OPIc®, a review of the American Council on Education (ACE) process, previous inter-rater reliability and agreement research on the ACTFL OPIc®, 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 OPIc®

The ACTFL OPIc® is a semi-direct test of functional spoken language proficiency, delivered via the Internet, and designed to elicit a 20 to 40 minute sample of ratable speech. Each test is unique and individualized through the selection of tasks within topic areas tailored to each test taker's linguistic ability, work experience, academic background, and interests. The test taker interacts with an embodied agent on the computer screen by responding to prompts in the target language. The elicited speech sample is digitally recorded and rated by a minimum of two ACTFL Certified OPIc® Raters. Certified ACTFL OPIc® raters compare the sample to the descriptions contained in the ACTFL Proficiency Guidelines – Speaking and assign one of ten possible ratings. The two ratings must agree exactly. Any rating discrepancy is arbitrated by a third tester and an Official ACTFL OPIc® 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 College Credit Recommendation Service (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 OPIc® RESEARCH

In 2005, the American Council on the Teaching of Foreign Languages (ACTFL) developed the ACTFL Oral Proficiency Interview – computerized (OPIc®), a computer-delivered assessment of speaking proficiency. The ACTFL OPIc® is rated according to the *ACTFL Proficiency Guidelines – Speaking (Revised 1999)*. The English ACTFL OPIc® demonstrated acceptable levels of reliability and validity through inter-rater reliability, test-retest reliability, and construct validity evidence (Surface, Poncheri, & Bhavsar, 2008). The success of this initial ACTFL OPIc® led to development in additional languages, such as Spanish. More specifically, previous studies have indicated that analyzing and documenting the inter-rater reliability of test-taker language proficiency for high-stakes language proficiency testing like the ACTFL OPIc® is imperative.

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 OPIc® in Spanish, English and Arabic?

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

3. What is the inter-rater agreement of the ACTFL OPIc® in Spanish, English and Arabic?

4. Are there any differences in overall ACTFL OPIc® 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 were only in Categories I 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 OPIc®, 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 OPIc® data because the proficiency categories used for ACTFL OPIc® 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 OPIc 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 OPIc® 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, since there was not any category II and III languages included, we reported agreement for categories I and IV only.

SECTION 4: RESULTS

Research Question 1 - *What is the inter-rater reliability of the ACTFL OPIc® in Spanish, English and Arabic?*

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 (0.947 to 0.970). English results are consistent with previous research.

Table 1
Spearman's Correlations by Language

	N	**Spearman's *R***	
		R	***p***
Arabic	173	.947	.000
Spanish	3470	.957	.000
English	1245	.970	.000

Note: Spearman's *R* for the aggregated languages was .963.

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

As shown in Table 2, all correlations were above .80; inter-rater reliability was nearly identical between Category I and Category IV.

Table 2
Spearman's Correlations by Language Category

	N	**Spearman's *R***	
		R	***p***
Category I	4715	.961	.000
Category IV	173	.947	.000

Note: There are two Category I languages (English and Spanish), and Arabic is Category IV. There are no Category II and III languages in this study.

Inter-rater reliability correlations were calculated using Spearman's *R* for interview years 2009, 2010, and 2011. As shown in Table 3, all correlations were above .80; inter-rater reliability was nearly identical across years.

Table 3

Spearman's Correlations by Year

	N	Spearman's *R*	
		R	***p***
2009	551	.974	.000
2010	1526	.962	.000
2011	2811	.959	.000

Research Question 3 - *What is the inter-rater agreement of the ACTFL OPIc® in Spanish, English and Arabic?*

Both absolute and adjacent agreements between the two raters were calculated for each language. As shown in Table 4, absolute agreement statistics were above the 70% threshold. Arabic was acceptable but lower than desired. It should be noted that Arabic is a relatively new assessment, and the inter-rater agreement has increased from 28% in 2009 (*N*=7) to 73% in 2011 (*N*=152).

Table 4

Absolute/Adjacent Agreement by Language

	N	**Absolute Agreement (exact)**	**Adjacent Agreement (+/- 1)**	**None (+/- 2)**
Arabic	173	71%	25%	5%
Spanish	3470	80%	17%	3%
English	1245	80%	16%	4%

Note. Percentages are rounded to the nearest whole number, and thus may not always add up to 100%. The absolute agreement across all languages is 79.7%.

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

Both absolute and adjacent agreements between the two raters were calculated for Category I/II and Category III/IV languages. As shown in Table 5, absolute agreement was above 70% for Category I and Category IV.

Table 5

Absolute/Adjacent Agreement by Language Category

	N	**Absolute Agreement (exact)**	**Adjacent Agreement (+/- 1)**	**None (+/- 2)**
Category I	4715	80%	17%	3%
Category IV	173	71%	25%	5%

Note. Percentages are rounded to the nearest whole number, and thus may not always add up to 100%. There are two Category I languages (English and Spanish), and Arabic is Category IV. There are no Category II and III languages in this study.

Both absolute and adjacent agreements between the two raters were calculated for each interview year. As shown in Table 6, absolute agreement was above the minimum threshold for operational use (i.e., 70%) for all years.

Table 6

Absolute/Adjacent Agreement by Year

	N	**Absolute Agreement (exact)**	**Adjacent Agreement (+/- 1)**	**None (+/- 2)**
2009	551	86%	11%	3%
2010	1526	79%	18%	4%
2011	2811	79%	18%	3%

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

Both absolute and adjacent agreements between the two raters were calculated for each major proficiency level. As shown in Table 7, absolute agreement was above the minimum of 70% for all major proficiency levels.

Table 7

Absolute/Adjacent Agreement by Major Proficiency Level

	N	**Absolute Agreement (exact)**	**Adjacent Agreement (+/- 1)**	**None (+/- 2)**
Novice	587	86%	13%	1%
Intermediate	2592	80%	17%	3%
Advanced	1452	75%	20%	5%
Superior	257	86%	13%	1%

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

Both absolute and adjacent agreements between the two raters were calculated for each sublevel proficiency level. As shown in Table 8, absolute agreement by test-taker proficiency level ranged from 60% to 94%. The Advanced Low and Advanced High categories were 66% and 60%, respectively. Absolute agreement for all other proficiency ratings was above 70%.

Table 8

Absolute/Adjacent Agreement by Sublevel Proficiency

	N	**Absolute Agreement (exact)**	**Adjacent Agreement (+/- 1)**	**None (+/- 2)**
Novice Low	194	94%	6%	0%
Novice Mid	198	88%	11%	1%
Novice High	195	76%	22%	2%
Intermediate Low	449	74%	23%	3%
Intermediate Mid	1024	84%	16%	1%
Intermediate High	1119	79%	15%	6%
Advanced Low	302	66%	31%	3%
Advanced Mid	878	83%	11%	6%
Advanced High	272	60%	36%	4%
Superior	257	86%	13%	1%

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 OPIc® exceeded inter-rater reliability and inter-rater agreement minimum standards. Overall, the inter-rater reliability was quite high (*R*=.963). The Spearman's *R* correlations ranged from .947 to .970 across the three 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. Inter-rater agreement for Arabic was acceptable but lower than desired. However, it is the newest language in the study and showed an improving trend. Absolute agreement was similar across interview language and language category. The highest agreement occurred at the extreme ends of the proficiency scale. That is, agreement was highest for the Superior proficiency level (86%) and the Novice Low (94%) and Novice Mid proficiency levels (88%). Overall, the reliability evidence in the current study supports the operational use of the OPIc®. Areas for continued improvement include increasing rater agreement within the Advanced proficiency levels and increasing rater agreement for the Arabic ACTFL OPIc®.

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.

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.

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., Poncheri, R. M., & Bhavsar, K. S. (2008, March). *Two studies investigating the reliability and validity of the English ACTFL OPIc® with Korean test takers: The ACTFL OPIc® validation project technical report*. Raleigh, NC: SWA Consulting Inc.

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