



---

## **Reliability Study of the ACTFL WPT<sup>®</sup> in Arabic, Russian, and Spanish 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) Writing Proficiency Test (WPT<sup>®</sup>) from 2008 to 2011 to satisfy a review requirement of the American Council on Education College Credit Recommendation Service (CREDIT) program. The ACTFL WPT<sup>®</sup> is an assessment of functional writing proficiency in a foreign language which is delivered by trained and certified experts in a writing format across numerous languages. Comparisons of ACTFL WPT<sup>®</sup> inter-rater reliability and agreement were made across three languages: Arabic, Russian, and Spanish. Comparisons were also made across language categories (i.e., language difficulty) and interview years (i.e., 2008 to 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 WPT<sup>®</sup> 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

The ACTFL WPT<sup>®</sup> exceeded the minimum inter-rater reliability and agreement standards. Further, the findings are fairly consistent with results from Surface, Dierdorff, and Poncheri (2006), indicating the ACTFL WPT<sup>®</sup> process yields relatively stable reliability results over time.

- Inter-rater Reliability
  - Spearman *R*s exceeded the standard for use, ranging from .92 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 and lowest for Novice High.
  - Absolute agreement was similar across language categories.
  - Absolute agreement was highest for Novice Mid and Superior.

Overall, the findings support the reliability of the ACTFL WPT<sup>®</sup> as an assessment of writing 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

<b>EXECUTIVE SUMMARY .....</b>	<b>2</b>
<b>SECTION 1: PURPOSE .....</b>	<b>4</b>
<b>SECTION 2: BACKGROUND.....</b>	<b>4</b>
<b>SECTION 3: METHOD.....</b>	<b>7</b>
<b>SECTION 4: RESULTS .....</b>	<b>9</b>
<i>Research Question 1 - What is the inter-rater reliability of the ACTFL WPT® in Arabic, Russian, and Spanish?.....</i>	<i>9</i>
<i>Research Question 2 - Are there any differences in overall ACTFL WPT® inter-rater reliability levels by language category and assessment year (2008-2011)?.....</i>	<i>9</i>
<i>Research Question 3 - What is the inter-rater agreement of the ACTFL WPT® in Arabic, Russian, and Spanish?.....</i>	<i>10</i>
<i>Research Question 4 - Are there any differences in overall ACTFL WPT® inter-rater agreement levels by language category, assessment year (2008-2011), and proficiency level?.....</i>	<i>10</i>
<b>SECTION 5: INTERPRETATIONS AND CONCLUSIONS .....</b>	<b>13</b>
<b>REFERENCES.....</b>	<b>14</b>
<b>ABOUT SWA CONSULTING INC. ....</b>	<b>15</b>

## **Reliability Study of the ACTFL WPT<sup>®</sup> in Arabic, Russian, and Spanish 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 Writing Proficiency Test (WPT<sup>®</sup>) assessment, which is an assessment of writing proficiency using a test 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 three interview languages— Arabic, Russian, and Spanish —and across four years—2008 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 WPT<sup>®</sup>, a review of the American Council on Education (ACE) process, previous inter-rater reliability and agreement research on the ACTFL WPT<sup>®</sup>, 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](mailto:esurface@swa-consulting.com)).

### **SECTION 2: BACKGROUND**

#### **THE ACTFL WPT<sup>®</sup>**

The ACTFL WPT<sup>®</sup> is an assessment of functional writing ability that measures how well a person spontaneously writes in the target language in response to four carefully constructed prompts dealing with practical, social, and professional topics that are encountered in true-to-life informal and formal contexts. The individual whose writing proficiency is being evaluated is presented with tasks and contexts that represent the range of proficiency levels from Novice to

Superior. All instructions and prompts are written in English; responses are written in the target language. The ACTFL WPT<sup>®</sup> can be administered in booklet form or via the Internet. The ACTFL WPT<sup>®</sup> lasts for 80 minutes (with an additional 10 minutes allotted for instructions). The elicited writing sample is compared to the descriptions contained in the ACTFL Proficiency Guidelines - Writing, and a rating is assigned. Each test is rated by two ACTFL Certified WPT<sup>®</sup> Raters. The two ratings must agree exactly. Any rating discrepancy is arbitrated by a third rater and an official ACTFL WPT<sup>®</sup> 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 WPT<sup>®</sup> RESEARCH ON INTER-RATER RELIABILITY AND INTER-RATER AGREEMENT**

Previous studies provided psychometric support for the use of writing proficiency measures developed according to the ACTFL proficiency guidelines. In 1990, Dandonoli and Henning presented the results of a multitrait-multimethod validation study, which included tests of speaking, writing, listening and reading in French and English as a Second Language (ESL). The inter-rater reliabilities for the writing test for the English and French samples were strong (reported Pearson *rs* of .87 and .89, respectively).

Thompson (1996) presented results from Russian speaking, reading, listening and writing proficiency assessments. The study used two samples of students: one from the University of Iowa and one from the Middlebury Russian Summer program. The inter-rater reliabilities for both the Iowa and the Middlebury samples were statistically significant, Pearson's  $r = .91$  and  $.72$ , respectively. In a recent conference report, Surface, Dierdorff, and Poncheri (2006) found strong support for favorable inter-rater reliability for the WPT, especially for the Spanish version. Further, the majority of rater pairs were making identical proficiency level judgments when scoring the WPT®.

## RESEARCH QUESTIONS

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

1. What is the inter-rater reliability of the ACTFL WPT® in Arabic, Russian, and Spanish?
2. Are there any differences in overall ACTFL WPT® inter-rater reliability levels by language category<sup>1</sup> and assessment year (2008-2011)?
3. What is the inter-rater agreement of the ACTFL WPT® in Arabic, Russian, and Spanish?
4. Are there any differences in overall ACTFL WPT® inter-rater agreement levels by language category, assessment year (2008-2011), and proficiency level?

---

<sup>1</sup>Language category is a proxy for language difficulty (Surface & Dierdorff, 2003). Given the languages in the study—only one language per Categories I, III and IV (none for Category II)—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 WPT<sup>®</sup>, 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 WPT<sup>®</sup> data because the proficiency categories used for WPT<sup>®</sup> 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 WPT 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 WPT<sup>®</sup> 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 is assigned to Category I; Russian is in Category III; and Arabic is in Category IV. There was no language in Category II for this study. For the purposes of this report, Categories I and II were collapsed into a single category: Category I/II (Spanish). Categories III and IV were also collapsed into a single category: Category III/IV (Russian and Arabic).



## SECTION 4: RESULTS

**Research Question 1** - *What is the inter-rater reliability of the ACTFL WPT<sup>®</sup> in Arabic, Russian, and Spanish?*

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 (0.923 to 0.982). The current results are highly consistent with the conference report by Surface, Dierdorff, and Poncheri (2006). The Spanish correlation was .921 ( $N=395$ ).

Table 1  
*Spearman's Correlations by Language*

	Current Study	
	<i>N</i>	<i>R</i>
<b>Arabic</b>	230	.982
<b>Russian</b>	180	.967
<b>Spanish</b>	3233	.923
<b>Overall</b>	3643	.935

*Note:* All correlations in the current study were significant at the  $p < .000$  level.

**Research Question 2** - *Are there any differences in overall ACTFL WPT<sup>®</sup> inter-rater reliability levels by language category and assessment year (2008-2011)?*

As shown in Table 2, the results by language category were all above the .80 standard, demonstrating high inter-rater reliability. Again, because Categories I, III & IV had only one language each, we choose to aggregate Category I & II languages and Category III & IV languages for analysis.

Spearman's *R* coefficients for interview years 2008-2011 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 four years in the current study. Inter-rater reliability was nearly identical across years.

Table 2  
*Spearman's Correlations by Language Category*

	<i>N</i>	<b>Spearman's <i>R</i></b>	
		<i>R</i>	<i>p</i>
<b>Category I/II</b>	3233	.923	.000
<b>Category III/IV</b>	410	.981	.000

*Note:* Category I/II is the same as Spanish only, Category III is Russian and Category IV is Arabic.

Table 3  
*Spearman's Correlations by Year*

	<i>N</i>	<b>Spearman's <i>R</i></b>	
		<i>R</i>	<i>p</i>
<b>2008</b>	35	.945	.000
<b>2009</b>	996	.926	.000
<b>2010</b>	1233	.940	.000
<b>2011</b>	1378	.934	.000

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

Both absolute and adjacent agreements were calculated for each language. As shown in Table 4, absolute agreement for all languages exceeded the minimum standard for use of 70%, indicating a relatively high level of concordance between raters. Furthermore, absolute and adjacent agreements were similar across languages. These findings are consistent with Surface, Dierdorff, and Poncheri (2006). When absolute and adjacent percentages are added, 98% to 99% of cases (depending on language) fall within the union of the sets.

Table 4  
*Absolute/Adjacent Agreement by Language*

	<i>N</i>	<b>Absolute Agreement (exact)</b>	<b>Adjacent Agreement (+/- 1)</b>	<b>None (+/- 2)</b>
<b>Arabic</b>	230	82%	17%	1%
<b>Russian</b>	180	81%	17%	2%
<b>Spanish</b>	3233	79%	20%	1%
<b>Overall</b>	3643	79%	20%	1%

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

**Research Question 4** - *Are there any differences in overall ACTFL WPT® inter-rater agreement levels by language category, assessment year (2008-2011), 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*

	<i>N</i>	<b>Absolute Agreement (exact)</b>	<b>Adjacent Agreement (+/- 1)</b>	<b>None (+/- 2)</b>
<b>Category I/II</b>	3233	79%	20%	1%
<b>Category III/IV</b>	410	81%	17%	2%

*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., 2008, 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%). The trend shows a slight improvement across the years.

Table 6  
*Absolute/Adjacent Agreement by Year*

	<i>N</i>	<b>Absolute Agreement (exact)</b>	<b>Adjacent Agreement (+/- 1)</b>	<b>None (+/- 2)</b>
<b>2008</b>	35	77%	20%	3%
<b>2009</b>	996	75%	24%	1%
<b>2010</b>	1233	82%	18%	1%
<b>2011</b>	1378	80%	20%	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 major proficiency level. As shown in Table 7, absolute agreement was above 70% for all major proficiency levels, with the exception of Novice (69%). These findings are fairly consistent with Surface, Dierdorff, and Poncheri (2006).

Table 7  
*Absolute/Adjacent Agreement by Major Proficiency Level*

	<i>N</i>	<b>Absolute Agreement (exact)</b>	<b>Adjacent Agreement (+/- 1)</b>	<b>None (+/- 2)</b>
<b>Novice</b>	26	69%	23%	8%
<b>Intermediate</b>	1375	83%	16%	1%
<b>Advanced</b>	1995	75%	24%	1%
<b>Superior</b>	247	88%	11%	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, with the exception of Novice High (64% to 100%). That is, agreement was highest for the Novice Mid sublevel proficiency (100%) and the Superior sublevel proficiency (88%).

Table 8  
*Absolute/Adjacent Agreement by Sublevel Proficiency*

	<i>N</i>	<b>Absolute Agreement (exact)</b>	<b>Adjacent Agreement (+/- 1)</b>	<b>None (+/- 2)</b>
<b>Novice Low</b>	---	---	---	---
<b>Novice Mid</b>	4	100%	0%	0%
<b>Novice High</b>	22	64%	27%	9%
<b>Intermediate Low</b>	56	73%	25%	2%
<b>Intermediate Mid</b>	343	84%	15%	1%
<b>Intermediate High</b>	976	83%	15%	1%
<b>Advanced Low</b>	861	70%	29%	1%
<b>Advanced Mid</b>	817	79%	20%	1%
<b>Advanced High</b>	317	78%	21%	1%
<b>Superior</b>	247	88%	11%	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 WPT<sup>®</sup> exceeded inter-rater reliability and inter-rater agreement minimum standards. Overall, the inter-rater reliability was quite high ( $R=.935$ ). The Spearman's  $R$  correlation was .982 for Arabic, .967 for Russian, and .923 for Spanish. Inter-rater reliability was high 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, with the exception of the Novice major level (69%) and Novice High sublevel proficiency (64%). Absolute agreement was similar across interview language and language category. There was a slight improvement in inter-rater agreement from 2008 to 2011. The highest agreement occurred at the extreme ends of the proficiency scale. That is, agreement was highest for the Novice Mid sublevel proficiency (100%) and Superior proficiency level (88%). Overall, the findings support the reliability of the ACTFL WPT<sup>®</sup> as an assessment of writing proficiency. Areas for continuous 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.
- 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., & Poncheri, R.M. (2006, May). *The “write” stuff: A preliminary assessment of a writing proficiency test*. Paper presented at the 21<sup>st</sup> annual conference of the Society for Industrial and Organizational Psychology, Dallas, TX.
- 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](http://www.swa-consulting.com)) or contact Dr. Eric A. Surface ([esurface@swa-consulting.com](mailto:esurface@swa-consulting.com)) or Dr. Stephen J. Ward ([sward@swa-consulting.com](mailto: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**