

Surface, E. A., & Poncheri, R. M. (2006, May). *Developing and validating a web-based assessment of speaking proficiency*. Symposium presented at the Computer Assisted Language Instruction Consortium Conference, Honolulu, HI.

# **Developing and Validating a Web-based Assessment of Speaking Proficiency**

Eric A. Surface  
SWA Consulting Inc.

Reanna M. Poncheri  
SWA Consulting Inc.



**MAY 2006**

APPROVED FOR PUBLIC RELEASE;  
DISTRIBUTION UNLIMITED

# Copyright Notice

This document and its content is copyright ©1997-2010 of SWA Consulting Inc. All rights reserved.

Any redistribution or reproduction of part, or the entire document in any form is prohibited except for: (1) you may print or download to a local hard disk extracts for your personal and non-commercial use only, and (2) you may copy the content to individual third parties for their personal use, but only if you acknowledge the website and author(s) as the source of the material. You may not, except with our express written permission, distribute or commercially exploit the content, nor may you transmit it or store it on any other website or other form of electronic retrieval system.



ACTFL

AMERICAN COUNCIL ON THE  
TEACHING OF FOREIGN LANGUAGES

# Developing and Validating a Web-based Assessment of Speaking Proficiency

CALICO – May 18, 2006

Honolulu, HI

SWA

Consulting

**Eric A. Surface, Ph.D.**

**Reanna M. Poncheri, M.S.**

---

# Overview

---

- Background on ACTFL Proficiency Testing
- ACTFL OPIc
- ACTFL OPIc Reliability and Validity Study
  - Background
  - Methodology
  - Findings by Research Question
- Future Research & Discussion



ACTFL

AMERICAN COUNCIL ON THE  
TEACHING OF FOREIGN LANGUAGES

# Background ACTFL Proficiency Testing

SWA

Consulting

---

# History of ACTFL Proficiency Tests

1986	ACTFL Oral Proficiency Interview (OPI)
1993	Business Writing Test
1994	Texas Oral Language Proficiency Test
1996	SST Standard Speaking Test of English
2001	Writing Proficiency Test (WPT)
2002	ACTFL Advanced Level Check
2005	Reading and Listening Screening Tests
2006	ACTFL OPIc

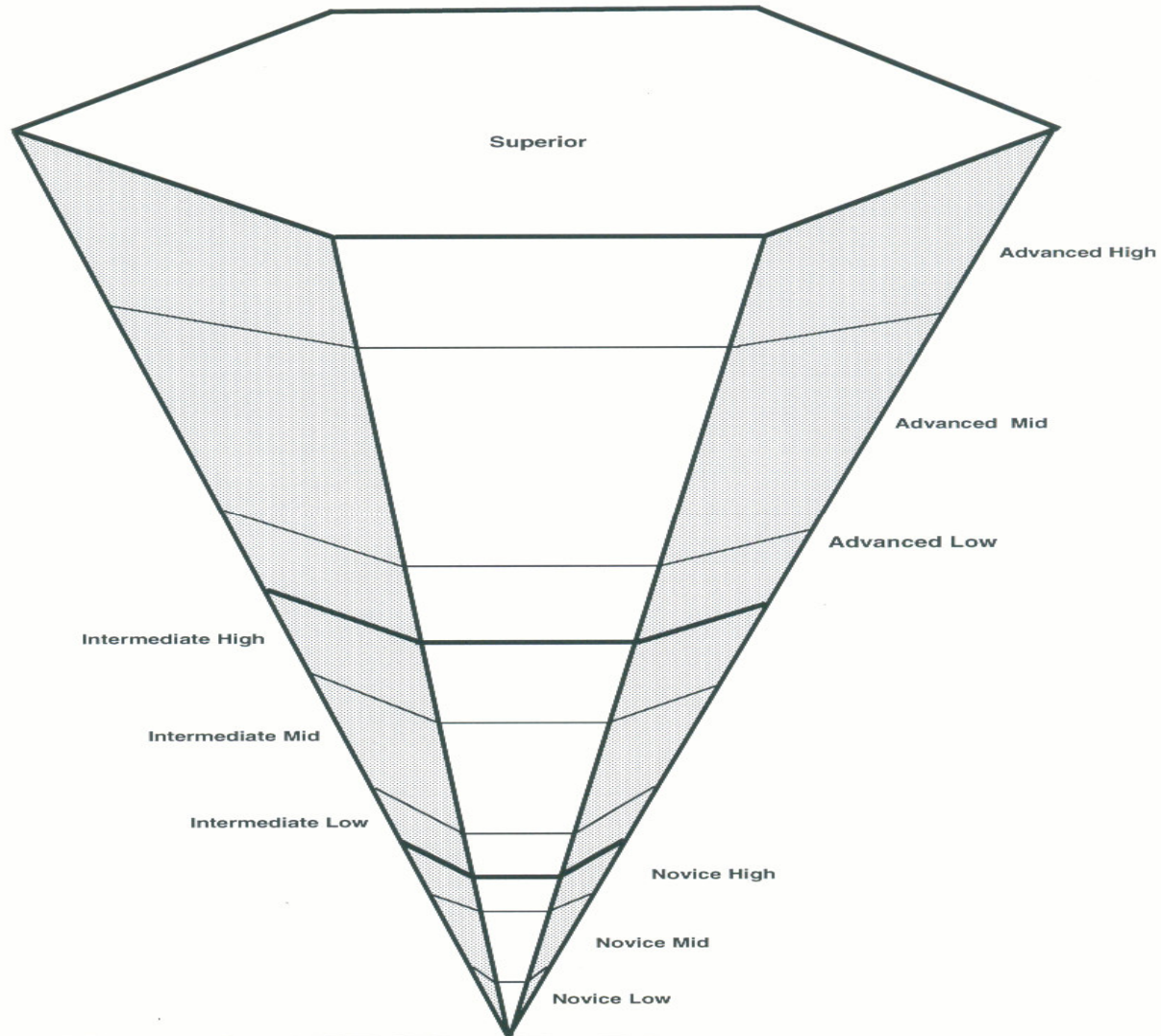
# ACTFL Proficiency Tests

- Valid and Reliable
- Assess Functional Language Ability
  - What a person “can do” with language in real life
- Not Curriculum or Program Dependent
- Recognize Competence Gained in Non-Traditional Settings
- Fair and Non-threatening
  - Adaptive and Interactive
  - Open ended - Free response
- ACTFL Proficiency Guidelines provide basis for rating

# ACTFL Proficiency Guidelines

- Based on the ILR (Interagency Language Roundtable) language descriptors
- Published in 1986
  - Speaking, Writing, Reading, Listening
  - Speaking Revised in 1999; Writing Revised in 2001
- 10-point scale over four major levels
  - Novice, Intermediate, Advanced, Superior
- Global characterizations of integrated performance
- Arranged in hierarchical order
- Identify stages of proficiency

# ACTFL Rating Scale



# ACTFL Ratings

- Certified Testers and Raters
  - Language educators from all levels of instruction
  - Superior level and native speakers/writers
  - Rigorous training and certification process
  - Ongoing tester quality assurance
  - Annual refresher training
  - Must re-certify every four years
- Rating the Language Sample
  - Look for the best match between language sample and level descriptions
  - Determine major level first, then the sublevel
  - Assign one of ten possible ratings

# ACTFL Oral Proficiency Interview (OPI)

- Face-to-face or telephonic assessment
- Interactive and adaptive
- Measure of consistent functional ability
- Standardized structure and protocol
  - Topics from the interests & experiences of interviewee
- Criterion-referenced evaluation
  - Function, Content/context, Text type, Accuracy
- Rating based on the *ACTFL Proficiency Guidelines – Speaking (Revised 1999)*
  - 10-point scale from Novice Low through Superior
  - Scored by certified ACTFL Testers and Raters



ACTFL

AMERICAN COUNCIL ON THE  
TEACHING OF FOREIGN LANGUAGES

# ACTFL OPIc

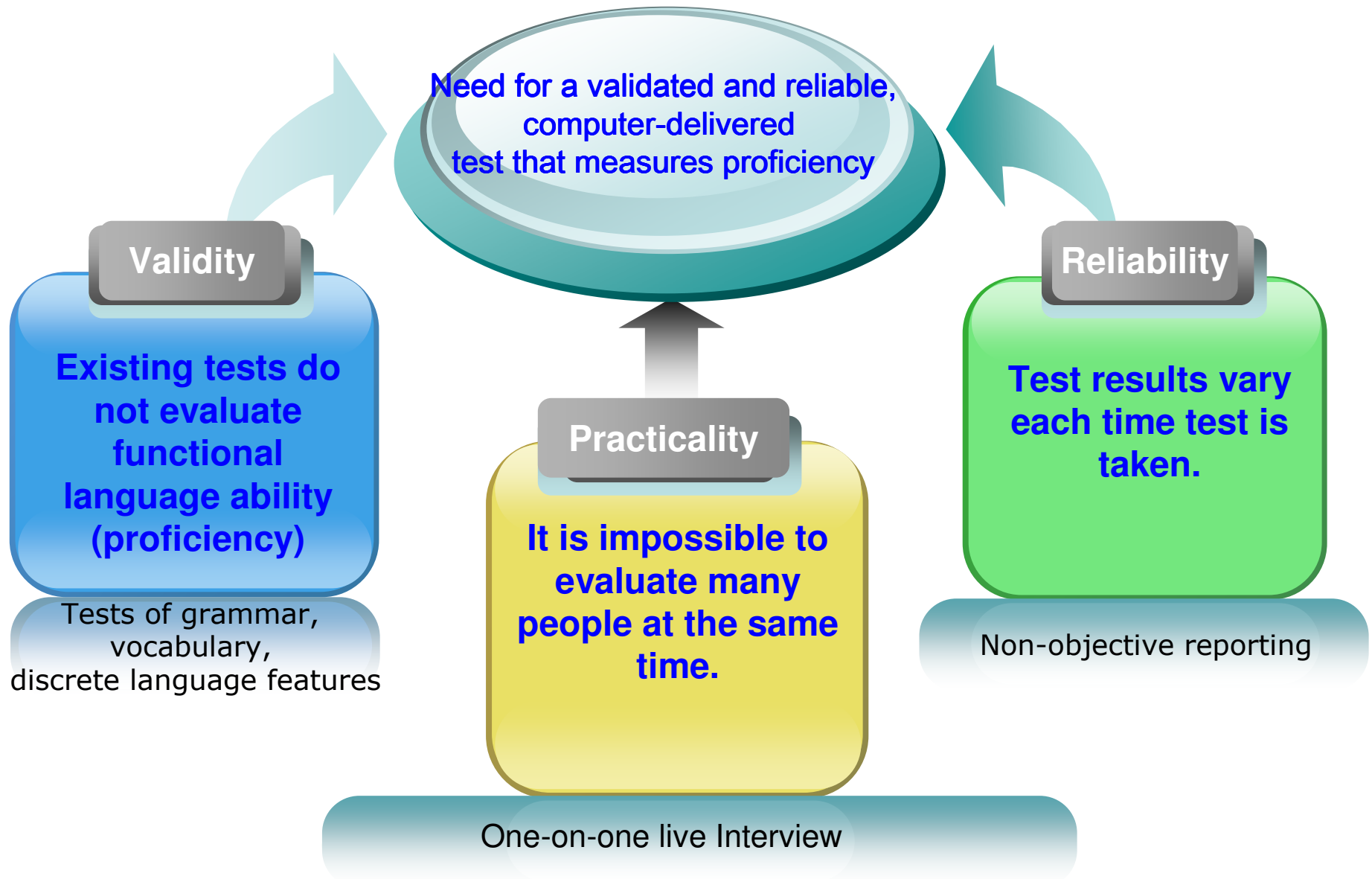
## Adapting the ACTFL OPI to a Computer-delivered Format

SWA

Consulting

---

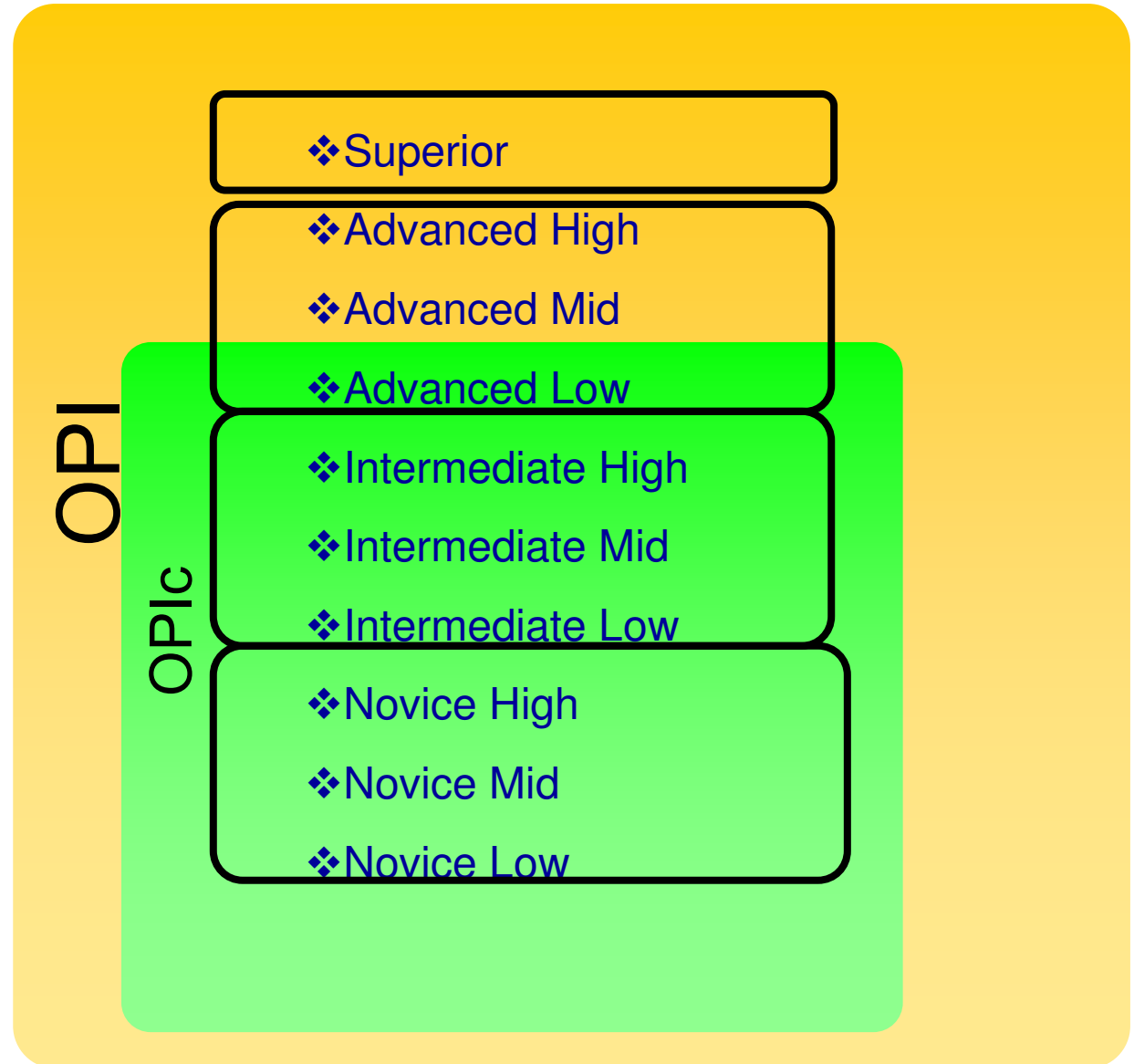
# Why Develop a New Test?



# What is the ACTFL OPIc?

- A semi-direct English oral proficiency test administered by computer
  - Accessed via the Internet
- Based on the structure of the ACTFL OPI
  - Adaptive and interactive
  - Test takers complete background survey and self assessment
- Valid and reliable according to industry standards
- Rated against same criteria as the ACTFL OPI
  - Scores correspond to OPI scores
- Rated by certified ACTFL raters

# OPIc and The ACTFL® Proficiency Rating Scale



# How is OPIc different from existing speaking tests?

1

## Interview Format

- ❖ Is like a conversation with a live tester
- ❖ Uses the interview format of the OPI
- ❖ Produces a sample of speech that can be rated

Conversation  
format

2

## Adaptive Features

- ❖ Each test is unique
- ❖ Background survey answers determine topic choices
- ❖ Self-assessment determines initial level

Each test is  
individualized

3

## ACTFL Certified Rating System

- ❖ ACTFL trained and certified raters
- ❖ Ongoing Quality Assurance

Secure  
Valid  
Reliable

# ACTFL OPic Language Selection

OPic

• Language Selection

In what language do you want to read test instructions?

시험 응시방법과 설문조사에 사용될 언어선택

ENGLISH       한국어 (Korean)

The language you select for reading instructions will not affect your English speaking test results.

시험 응시 방법과 설문등을 한국어로 선택하셔도 시험결과에는 어떤 영향도 주지 않습니다.

Next ▶

# ACTFL OPic Volume Setting

The screenshot displays the ACTFL OPic interface for setting volumes. At the top center is the OPic logo. Below it, the text "Setting Volumes" is displayed. A central white box contains a numbered list of instructions, with the text "Volume Setting Instruction" overlaid in red. To the left of the instruction box is a "PLAY" button with a play icon and a volume slider below it. To the right is a "RECORD" button with a record icon and a "REPLAY" button below it, also with a volume slider. At the bottom are "Previous" and "Next" navigation buttons.

OPic

• Setting Volumes

1. [blurred text]  
2. [blurred text]  
3. [blurred text]  
4. [blurred text]  
5. [blurred text]  
6. [blurred text]

**Volume Setting Instruction**

PLAY

RECORD

REPLAY

Previous

Next

# ACTFL OPIC Background Survey

## · Background Survey

To get started, please answer some survey questions about your work, studies, home and leisure time activities.

**1. Do you work? Check one**

- Full-time
- Part-time
- Do not work (if you checked "Do not work", go to question 5)

**2. Where do you work? Check one**

- Outside the home
- At home
- Do not work

**3. How many different companies have you worked for in your career? Check one**

- 1
- 2-3
- 4+
- Do not work

**4. Are you a manager who supervises other people? Check one**

- Yes

# ACTFL OPIc Self-Assessment

## • Self Assessment

Please read all 6 statements below.

Then choose the one that best describes **what you can do in spoken English**.

- I can say only a few words and memorized phrases in English.
- I can name basic objects, colors, days of the week, foods, clothing items, numbers, etc. I cannot always make a complete sentence or ask simple questions.
- I can give some basic information about myself, work, familiar people and places, and daily routines speaking in simple sentences. I can ask some simple questions.
- I can participate in simple conversations about myself, daily routines, work/school and hobbies. I can easily produce a series of simple sentences on these familiar topics and routines. I can also ask questions to get what I need.
- I can participate in conversations about familiar topics and activities related to home, work/school, personal and community interests. I can speak in connected discourse about things that have happened, are happening and will happen. I can explain when asked. I can handle routine situations, even when there may be an unexpected complication.

# ACTFL OPIc Test Description

## • Test Description

The OPIc is a computer administered test of spoken English.

The test will take between 10 and 20 minutes.

The OPIc is scored according to the *ACTFL® Proficiency Guidelines - Speaking (Revised 1999)*.

Next ►

# ACTFL OPIc General Instructions

## • Overview of OPIc

- 1) You will hear a series of recorded questions.
- 2) There is a **[Repeat]** button to replay the question once.
- 3) When you are ready to record your answer, press the **[Start Recording]** button.
- 4) Once you press the **[Start Recording]** button, you will not be able to reply or re-record.
- 5) When your answer is complete, press the **[Finish Recording]** button.

Next ▶

# ACTFL OPIc Sample Question



# ACTFL OPIc Test Start Page

## • Start Test

You are about to begin the OPIc.

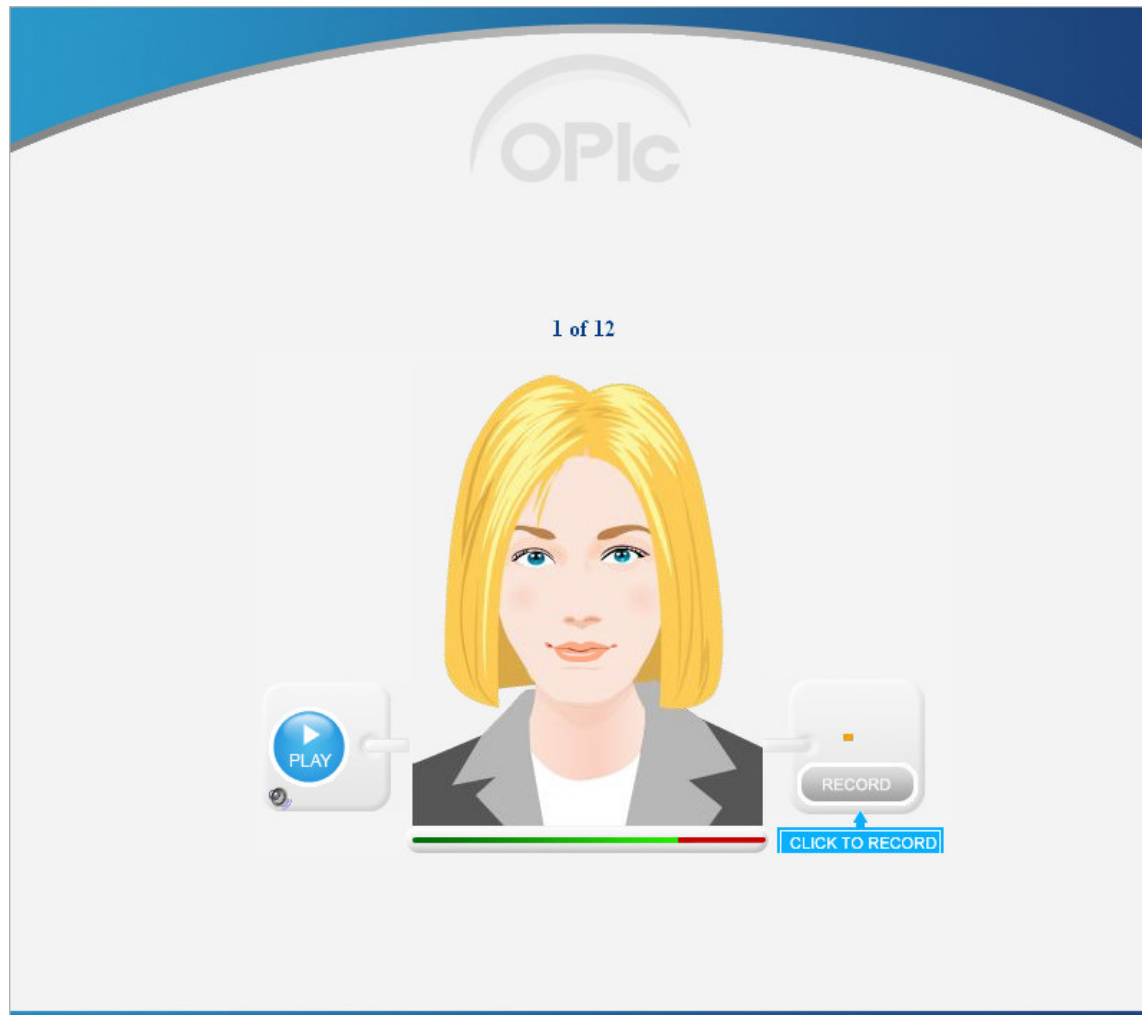
Remember to give as complete and detailed an answer as you can, to show your English ability at its best.

[Revisit OPIc Instruction](#)

[Revisit Sample Question](#)

Next ▶

# ACTFL OPIc First Question



# ACTFL OPIC Final Page

• End Test

You have now completed the ACTFL OPIC.

Any further instructions will be provided by your proctor.

Thank you and goodbye.

Retry

**Please wait until all your replies are transmitted to ACTFL.**

Log In



ACTFL

AMERICAN COUNCIL ON THE  
TEACHING OF FOREIGN LANGUAGES

# ACTFL OPIc Reliability and Validity Study

SWA

Consulting

---

# Background

- Test publishers should document the psychometric properties (e.g., reliability and validity) of their instruments according to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999)
- The *Standards* provide guidelines for presenting reliability and validity information about tests
- A test is any “evaluative device or procedure in which a sample of an examinee’s behavior...is obtained and subsequently evaluated...” (AERA et al., 1999)
  - Not restricted to paper-and-pencil assessments

# Background

- Although the *Standards* document provides the primary source of guidance for evaluating the acceptability of testing and assessments, **it is not a prescriptive document**
- The *Standards* acknowledges and advocates the use of and need for professional judgment
- The goal of our study was to assess the initial reliability and validity of the ACTFL OPIc in accordance with the guidance provided by the *Standards* and our professional judgment

# Reliability Evidence

- Reliability
  - Extent to which an item, scale, or procedure will yield the same value when administered across different times, locations, or populations
  - For ratings data, reliability refers to the homogeneity of judgments given by the sample of raters
  - Interrater reliability – overall level of consistency among a sample of raters
  - Test-retest reliability – repeatability of ratings across multiple administrations
  - Purpose and consequences determine acceptable level.

# Validity Evidence

- Validity
  - “the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed purpose” (AERA, APA, NCME, 1999, p. 11)
  
- Validity Evidence
  - Test content (content validity)
    - Addressed during development and initial testing
  - Internal structure (construct validity)
    - An evaluation of the degree to which ratings are related and are representative of the construct in question
  - Relations to other variables (convergent validity)
    - Relationship between scores on a test and scores on measures that assess the same or similar constructs

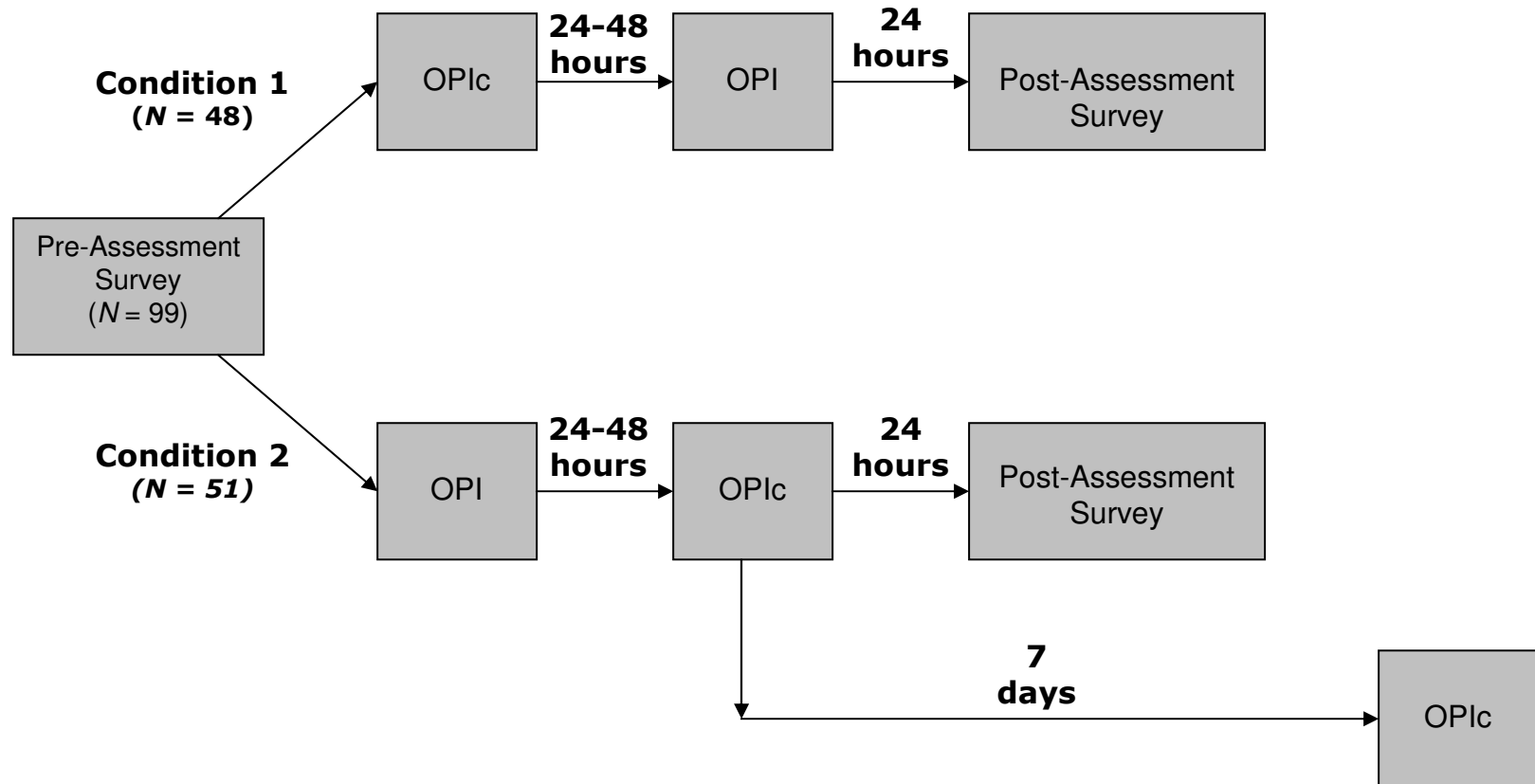
## Method: Participants

- Overall Sample ( $N = 151$ )
  - Gender: 67 males (44.4%), 84 females (55.6%).
  - Age: 21-42 (Mean age = 29)
  - Education: 83.4% university graduates
- Pilot Study 1 ( $N = 20$ )
  - Pilot test the ACTFL OPIc
  - Capture samples of speech for rater training
- Pilot Study 2 ( $N = 100$ )
- Holdout Sample ( $N = 31$ )

# Methods: Measures

- OPI
  - Validity evidence (e.g., Dandonoli & Henning, 1990)
  - Reliability evidence (e.g., Surface & Dierdorff, 2003)
  - Modified Protocol - three raters instead of standard two
- OPIc
  - Used three raters instead of standard two
  - OPIc raters are first trained as OPI raters
- Web-based pre-assessment survey
  - Attitudes toward computerized/telephonic tests, test-taking self-efficacy, visual and auditory learning styles, goal orientation
- Web-based post-assessment survey
  - Reactions to ACTFL OPIc, reactions to the ACTFL OPI, comparison of ACTFL OPI and ACTFL OPIc

# Method: Study Design



## What is the overall interrater reliability and consistency of the ACTFL OPIc?

- Analysis
  - Intraclass class correlations (ICCs; Shrout & Fleiss, 1979)
  - Estimate of maximal reliability ( $R_{Max}$ ; Drewes, 2000)
- Findings
  - First administration
    - ICC = .94 ( $F = 46.63, p = .00, n = 96$ )
    - $R_{Max} = .98$
  - Second administration
    - ICC = .79 ( $F = 12.40, p = .00, n = 42$ )
    - $R_{Max} = .98$

## How does the interrater reliability of the ACTFL OPIc compare to the ACTFL OPI?

- Analysis
  - Comparison of  $R_{\text{Max}}$  and ICCs for ACTFL OPI and ACTFL OPIc
- Findings
  - ACTFL OPI
    - ICC = .93 ( $F = 40.35, p = .00, n = 96$ )
    - $R_{\text{Max}} = .98$
  - ACTFL OPI and ACTFL OPIc have similar reliability results

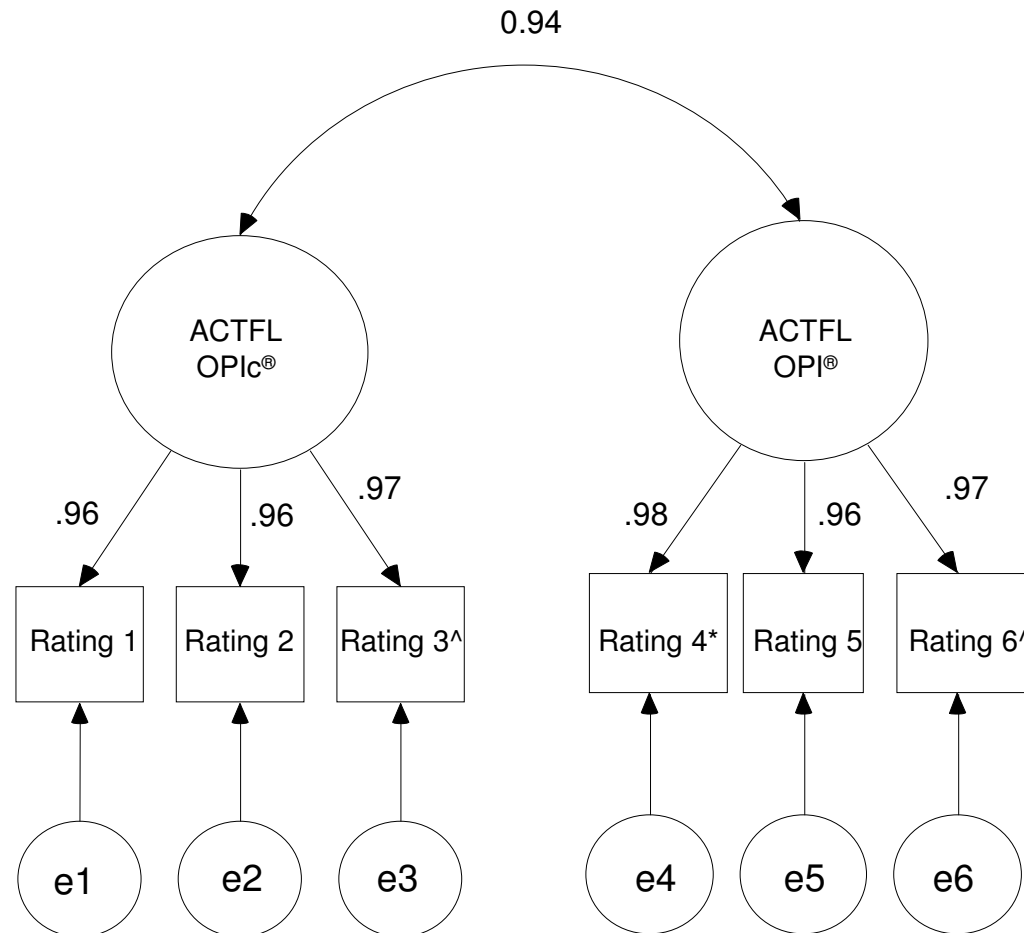
## What is the relationship between the ACTFL OPIc and the ACTFL OPI final ratings?

- Analysis
  - Correlations between final ratings of the ACTFL OPI and the ACTFL OPIc (1<sup>st</sup> and 2<sup>nd</sup> administrations)
- Findings
  - ACTFL OPI and ACTFL OPIc (1<sup>st</sup> administration)
    - $r = .92, p = .00; R = .91, p = .00$
  - ACTFL OPI and ACTFL OPIc (2<sup>nd</sup> administration)
    - $r = .94, p = .00; R = .94, p = .00$

## How do the underlying constructs of the ACTFL OPI and ACTFL OPIc compare?

- Analysis
  - Confirmatory factor analysis (CFA)
  - Correlated two-factor model to assess the relationship between the ACTFL OPI and the ACTFL OPIc
- Model Fit
  - $\chi^2 = 9.36 (8, 95)$
  - CFI = 1.00, TLI = 1.00
  - RMSEA = .04 [.00 - .13], SRMR = .01

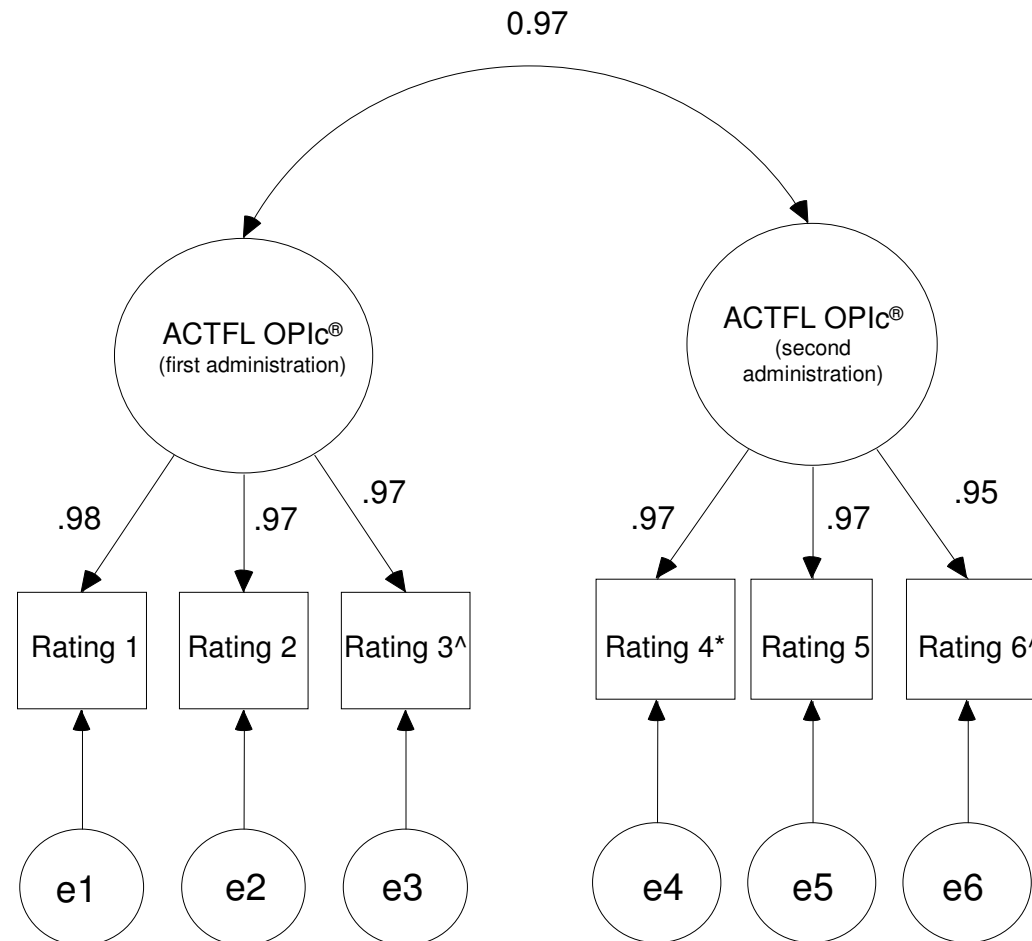
# Correlated Two-Factor Model



## How do the underlying constructs of the two administrations of the ACTFL OPIc compare?

- Analysis
  - Confirmatory factor analysis (CFA)
  - Correlated two-factor model to assess the relationship between the first and second administration of the ACTFL OPIc
- Model Fit
  - $\chi^2 = 4.14 (7, 95)$
  - CFI = 1.00, TLI = 1.00
  - RMSEA = .04 [.00 - .13], SRMR = .01

# Correlated Two-Factor Model



## What is the test-retest reliability of the ACTFL OPIc across two administrations?

- Analysis
  - Correlation between the final ratings obtained from the first and second administration of the OPIc
  - Correlation between the common rater for the first and second administration of the OPIc
- Findings
  - Correlation between 1<sup>st</sup> and 2<sup>nd</sup> administrations
    - $r = .94, p = .00; R = .91, p = .00$
  - Correlation between common raters for 1<sup>st</sup> and 2<sup>nd</sup> administration
    - $r = .89, p = .00; R = .89, p = .00$

## Additional Findings: Order of Administration

- Had no impact on final ratings
  - ACTFL OPI ( $F = .22, p = .64$ )
  - ACTFL OPIc (1<sup>st</sup> administration;  $F = .02, p = .89$ )
- Had no impact on agreement between the ACTFL OPI and ACTFL OPIc (1<sup>st</sup> administration)
  - $\chi^2 = 2.23, df = 3, p = .53$

# Additional Findings: Self-Assessment

- OPIc prompts are generated based on participants' self-assessment of proficiency
- Absolute agreement between the OPIc and the OPI was not as high as expected
  - Most disagreements were result of OPIc underestimated OPI rating
  - Most disagreements occurred when participants self-assessed at the lowest level of proficiency
  - Our analysis demonstrated that the under assessment of proficiency led to decreased concordance between OPIc and OPI ratings
  - **However, decreased absolute agreement did not impact reliability or validity coefficients**
- Recommendations were made to prevent under or over self-assessments of proficiency

# Summary of Findings

- ACTFL OPIc has acceptable levels of validity and reliability across all analyses. Majority of validity and reliability coefficients were above .90
- Reliability
  - Interrater reliability and maximum reliability coefficients for the OPIc indicated sufficient agreement
  - Test-retest reliability was demonstrated across two administrations of the OPIc
- Validity
  - OPIc was strongly related to the OPI
  - Construct validity coefficients were very high for the OPIc (.99)
- Absolute agreement was slightly lower than desired because of the self-assessment issue

## Study Recommendations for OPIc

- Adjust the self-assessment process to include more proficiency categories for participants and sound file examples for each proficiency category
- Ensure that the core ACTFL OPIc assessment provides sufficient opportunities or prompts for test takers to demonstrate speaking proficiency regardless of self-assessment
- Additional training for all ACTFL OPIc raters
- ACTFL and LTI should engage in an iterative development process using empirical data to improve the ACTFL OPIc
- Improve the user interface

# Limitations

- Technical issue with post-assessment data collection and a number of open-ended responses were lost
- Sample size
- Proficiency levels in the sample were not equal across scale
- “Artificial” nature of study may have impacted responses on pre- and post-assessments
- Speech samples were rated by three raters instead of the two raters – difficult to obtain absolute agreement
- Findings may be idiosyncratic to this sample

# ACTFL OPIc Follow-up Study

- Embraced the iterative improvement process
- Followed recommendations to improve the OPIc
- Study Goals:
  - Sample covering full range of proficiency
  - Approximate a real testing situation
- Study Design:
  - Sample of approximately 30 Korean professionals in a large organization
  - Administration order: Pre-Assessment survey, OPIc, and OPI
  - Used the standard two raters
- Preliminary results look favorable

## Future Research

- Assess the impact of any modifications made to the ACTFL OPIc as a result of this study (*in process*)
- Extend the validity and reliability evidence for the ACTFL OPIc across multiple contexts (*future as testing program expands*)
- Assess the ACTFL OPIc protocol in additional languages and cultures (*future as testing program expands*)
- Investigate the individual differences that might potentially influence an individual's ACTFL OPIc rating (e.g., computer test-taking anxiety). Can use current and future data



ACTFL

AMERICAN COUNCIL ON THE  
TEACHING OF FOREIGN LANGUAGES

# Discussion & Questions

Thank you.

SWA

Consulting

---



ACTFL

AMERICAN COUNCIL ON THE  
TEACHING OF FOREIGN LANGUAGES

## Contact Information

### *About the study:*

Dr. Eric A. Surface

[esurface@swa-consulting.com](mailto:esurface@swa-consulting.com)

### *About ACTFL or the ACTFL OPI or OPIc:*

Dr. Elvira Swender

[eswender@actfl.org](mailto:eswender@actfl.org)

### *About Scheduling an ACTFL Test:*

Ms. Helen Hamlyn

[hamlyn@languagetesting.com](mailto:hamlyn@languagetesting.com)

SWA

Consulting

---

## **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
- 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 with which to make solid data-driven 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. SWA has competencies in measurement, data collection, analytics, data modeling, systematic reviews, validation, and evaluation.

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)).