

Swender, E., Hamlyn, H., Surface, E. A., & Clifford, R. T. (2009, March). *Title VI grants: Laying the groundwork for improved language skills assessment*. Session presented at the Title VI 50th Conference, Washington, D.C.

Title VI Grants: Laying the Groundwork for Improved Language Skills Assessment

Elvira Swender
ACTFL

Helen Hamlyn
ACTFL

Eric A. Surface
SWA Consulting Inc.

Ray T. Clifford
Brigham Young University



MARCH 2009

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.

The Research Component

Oral Proficiency Interview (OPI)

- Cited in more than 131 journal articles
- Cited in 2 books
 - *Teaching world languages for social justice: A sourcebook of principles and practices* (T. A. Osborn, 2006)
 - *The Art of Non-Conversation, A reexamination of the ACTFL OPI* (Marysia Johnson, 2001)

Oral Proficiency Interview (OPI)

- Cited in more than 131 journal articles
 - 77 journal articles specifically mention the ACTFL OPI.
- Cited in 2 books
 - *Teaching world languages for social justice: A sourcebook of principles and practices* (T. A. Osborn, 2006)
 - *The Art of Non-Conversation, A reexamination of the ACTFL OPI* (Marysia Johnson, 2001)

Writing Proficiency Test (WPT)

- Cited in 1 journal and 1 book
 - Journals include
 - *Foreign Language Annals*
 - Books include
 - *Teaching world languages for social justice: A sourcebook of principles and practices*

Computerized Oral Proficiency Interview (OPIC)

- Cited in 2 journals and 1 proceedings
 - Journals include
 - *Language Testing*
 - *Language, Learning, and Technology*
 - *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2000*

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

Surface & Dierdorff, 2003

Study Design

- 5,881 interviews were rated by ACTFL-certified testers using the OPI
- Reliability analyses:
 - Pearson correlation
 - Spearman rank–order correlation
 - Kendall's tau
 - Goodman and Kruskal's gamma
 - Cohen's kappa
 - Raw percentages of agreement

Research Question 1

- What is the overall interrater consistency and agreement for all languages tested with the same ACTFL Revised Guidelines and rating protocol?
 - Overall interrater consistency across all rater pairs in all languages was significant ($p < .05$) for each of the test statistics.
 - Eighty percent of the ratings across all languages showed perfect agreement, compared to 18% that disagreed by one proficiency category.

Research Question 2

- Do interrater consistency and rater agreement levels vary across languages that are more commonly tested compared to those that are less commonly tested?
 - There were very little differences in both rater consistency and rater agreements levels between languages that are more commonly tested and those that are less frequently tested.

Research Question 3

- Do interrater consistency and rater agreement levels vary according to language difficulty?
 - The language difficulty classifications had practically no moderating effects on the magnitude of rater consistency.
 - The raw agreement percentages did not show any substantial discrepancies across the four language difficulty groups.

Research Question 4

- Do ratings of particular languages show more consistency or greater agreement than others?
 - No substantially large differences across the tested languages.

Research Question 5

- Does rater agreement vary across proficiency categories?
 - 41% of disagreement cases crossed a single major proficiency level boundary.
 - No disagreements were associated with crossing two major proficiency categories.

Research Question 6

- When the first and second raters disagree and a third rater must be utilized, is the third rater significantly more likely to resolve the disagreement in favor of one rater more often than the other?
 - Interrater reliability was clearly higher for second raters paired with third raters than it was for the first and third rater pairs.
 - This held across the testing density, language difficulty classifications, and specific languages as well.

Conclusions

- Overall, interrater reliability was found to be high.
- The estimates of interrater consistency found in this study were all above the recommendation for applied projects (.90).
- Consistency estimates found within the present study were similar to, but generally higher than, estimates found in previous OPI research.
- The reliability results presented for these languages ranged from .96 to .98.

The “Write” Stuff: Preliminary Assessment of a Writing Proficiency Test

Surface, Dierdorff, & Poncheri,
2006

Study Design

- 509 WPTs were administered and rated by ACTFL-certified testers.
 - 460 had OPI scores in addition to the WPT
- Five Languages:
 - French (n = 81)
 - German (n = 8)
 - Italian (n = 22)
 - Russian (n = 3)
 - Spanish (n = 395)

Research Question 1

- What is the reliability of the WPT across different languages?
 - All consistency estimates were statistically significant ($p < .05$) and well within desirable levels.
 - 80% of rater pairs provided identical proficiency judgments.

Research Question 2

- What is the reliability of the WPT for the most commonly tested language (Spanish)?
 - Consistency estimates were slightly higher for each test statistic when comparing the full sample and the Spanish-only sample
 - Raters judging the Spanish version of the WPT were in absolute agreement 77.7% of the time.
 - Advanced category displayed the highest level of absolute agreement (49.6%)
 - 50% of the identical judgment fell in the Advanced-Low and Advanced-Mid proficiency levels

Research Question 3

- Has the reliability of the WPT changed over time since its inception in 2002?
 - Corrected interrater reliabilities showed an upwardly trending pattern across the three years of WPT application.

Research Question 4

- What is the relationship between WPT and OPI scores?
 - The Pearson correlation computed between the WPT and OPI for all languages was .81
 - The Spearman rank–order correlation computed between the WPT and OPI for all languages was .81

Findings

- The results of this study offered strong evidence of favorable interrater reliability for judges scoring the ACTFL WPT.
 - Interrater reliability estimates were above the .90 level
 - Weighted Kappa coefficients were in the mid .80s

Findings

- A strong relationship between writing and speaking proficiency measures was found.
- Overall, the results of our study are positive for the WPT.

What Proficiency Testing is Telling Us about Teacher Certification Candidates

Hamlyn, Surface, & Swender,
2007

Study Design

- Proficiency of Teacher candidates
 - 3,198 of teacher candidates from 10 states
 - Languages Examined
 - Spanish, German, Mandarin, French, Italian
 - Other World Languages
 - Tests Included: OPI & WPT
 - State Requirements:
 - OPI
 - OPI and WPT

Key Findings

- Large Percentage of Teacher Certification Candidates met the minimal proficiency requirement for their State on the first attempt
 - 85% for State
 - 87% OPI
 - 94% WPT
 - Percentages varied by language

Key Findings

- The relationship between the OPI and WPT rating for Teacher Certification candidates varied across the proficiency scale
 - Below Advanced
 - Writing scores tended to be equal or higher than speaking scores
 - Advanced Low
 - Distribution of writing scores evened out in comparison to speaking scores
 - Above Advanced Low
 - Writing scores tended to be equal or lower than speaking scores

Key Findings

- OPI Retests
 - 2nd Attempt 49.8% successful
- WPT Retests
 - 2nd Attempt 84.4% successful
- Writing proficiency was more likely to improve during the 90-day waiting period in comparison to speaking proficiency.

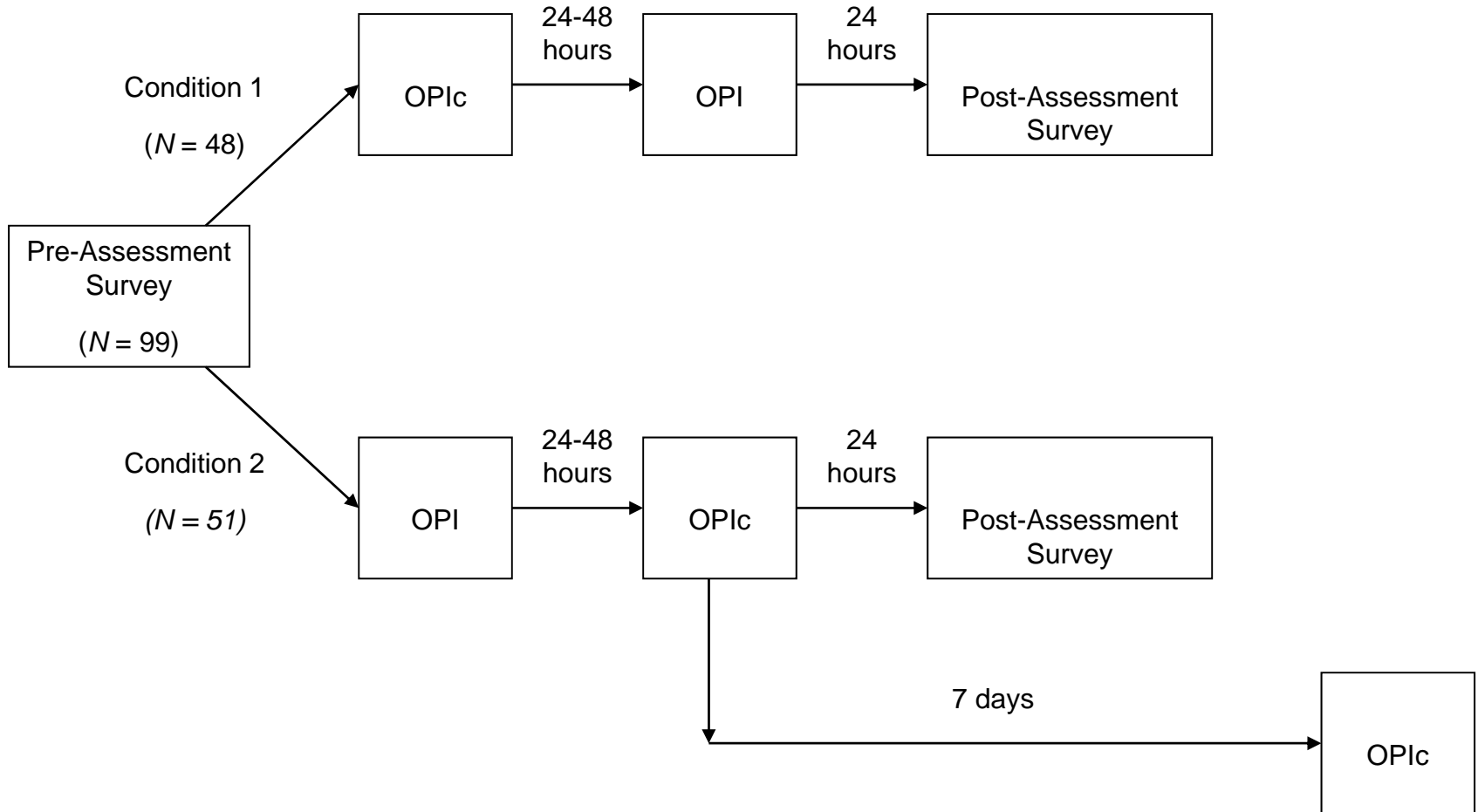
Key Findings

- NCATE OPI scores
 - 59.5% met minimal requirement (AL)
 - Up 9% from last study
 - 82% of State Teacher Candidate met AL or higher
 - NCATE OPI scores were more typical of the second language learner population.

Two Studies Investigating the Reliability and Validity of the English ACTFL OPIc[®] with Korean Test Takers

Surface, Poncheri, & Bhavsar,
2008

Study 1 Design



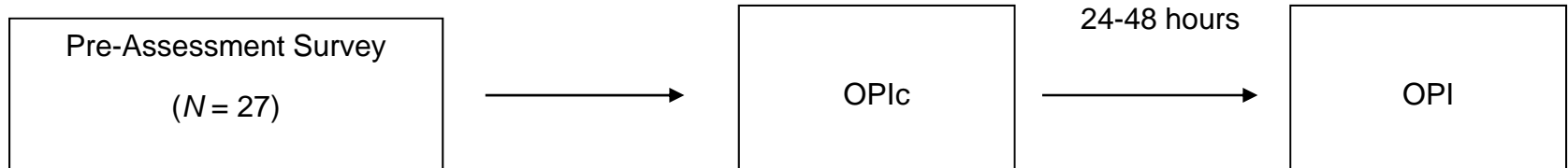
Study 1 Key Results

- The interrater reliability and agreement for the OPIc was consistent with the OPI
- There was a significant, strong positive relationship between the OPIc and the OPI ($r = .92$, $R = .91$)
- Results suggested that test takers who self-assessed at level one were being underestimated by the OPIc
- There was a significant, strong positive relationship between the 1st and 2nd administrations of the OPIc ($r = .94$, $R = .91$)

Study 1 Key Results

| | |
|--|-----|
| Absolute Agreement between final ratings of the OPI and OPIc | 63% |
| Agreement within major categories (novice, intermediate, advanced) | 85% |
| Agreement when the major category boundaries ignored, and agreement defined as an exact match or being off +/- one step (Adjusted Agreement) | 98% |

Study 2 Design



Study 2 Key Results

- The interrater reliability for the OPIc was consistent with the OPI
- There was a significant, strong positive relationship between the OPIc and the OPI ($r = .97$, $R = .95$)
- There was no evidence of underestimation of proficiency
- For 58% of the OPIc cases, there was exact agreement between Rater 1 and Rater 2

Study 2 Key Results

| | |
|--|------|
| Absolute Agreement between final ratings of the OPI and OPIc | 87% |
| Agreement when the major category boundaries ignored, and agreement defined as an exact match or being off +/- one step (Adjusted Agreement) | 100% |

Test-Retest Reliability and
Absolute Agreement Rates of
English ACTFL OPIc Proficiency
Ratings for Double and Single
Rated Tests within a Sample of
Korean Test Takers

SWA Consulting Inc., 2009

Overview

- Purpose: assess the stability and agreement of the final ratings obtained by individual test takers over the course of consecutive English OPIc administrations
- Sample: Korean test takers who completed multiple administrations of English OPIc
 - All completed two administrations within at least one 30-day period prior to 1/20/2009

Stability of All Final Ratings

- Final ratings of all OPIc tests were highly stable across the first two administrations (*r* values from .90-.93, *R* values from .90-.94)
- Absolute agreement analyses indicated high rates of agreement (85-92%) between the 1st and 2nd final ratings on the OPIc

Stability of Ratings Provided by Single Raters

- Test-retest reliability ($r = .98$) and absolute agreement (97%) both were very high
- Results indicated a high degree of stability in ratings across the first and second single-rater administrations

Investigating the Effectiveness of ACTFL OPI Tester Training

Surface, Swender, Brown, &
Dierdorff, 2008

Original Tester Training Format

Workshop Day 1: ACTFL Scale



Workshop Day 2: Assessment Criteria



Workshop Day 3: Interview Structure & Elicitation



Workshop Day 4: Implications & Applications of the OPI



Certification: *Practice Round*



Certification: *Certification Round*

Study Design

- Objectives:
 - Document the success of the 20+ year OPI Tester Training
 - Capture standardized participant feedback
 - Look for areas to improve the training & certification process
- Conducted at 4 Locations from July 2004 through November 2005
- PRE-POST without control group design
- Criteria:
 - Reactions
 - Learning
 - Certification

Key Findings – Reactions

- Positive reactions to training
 - Satisfaction ($M = 6.04$); Effectiveness ($M = 6.09$)
- Majority of participants (75%) said the workshop influenced them to consider changing the goals, objectives, methods, teaching practices, materials, etc. of their foreign language courses
- 42% of respondents made comments
 - Most common categories: Logistics, Instructor, and Environment

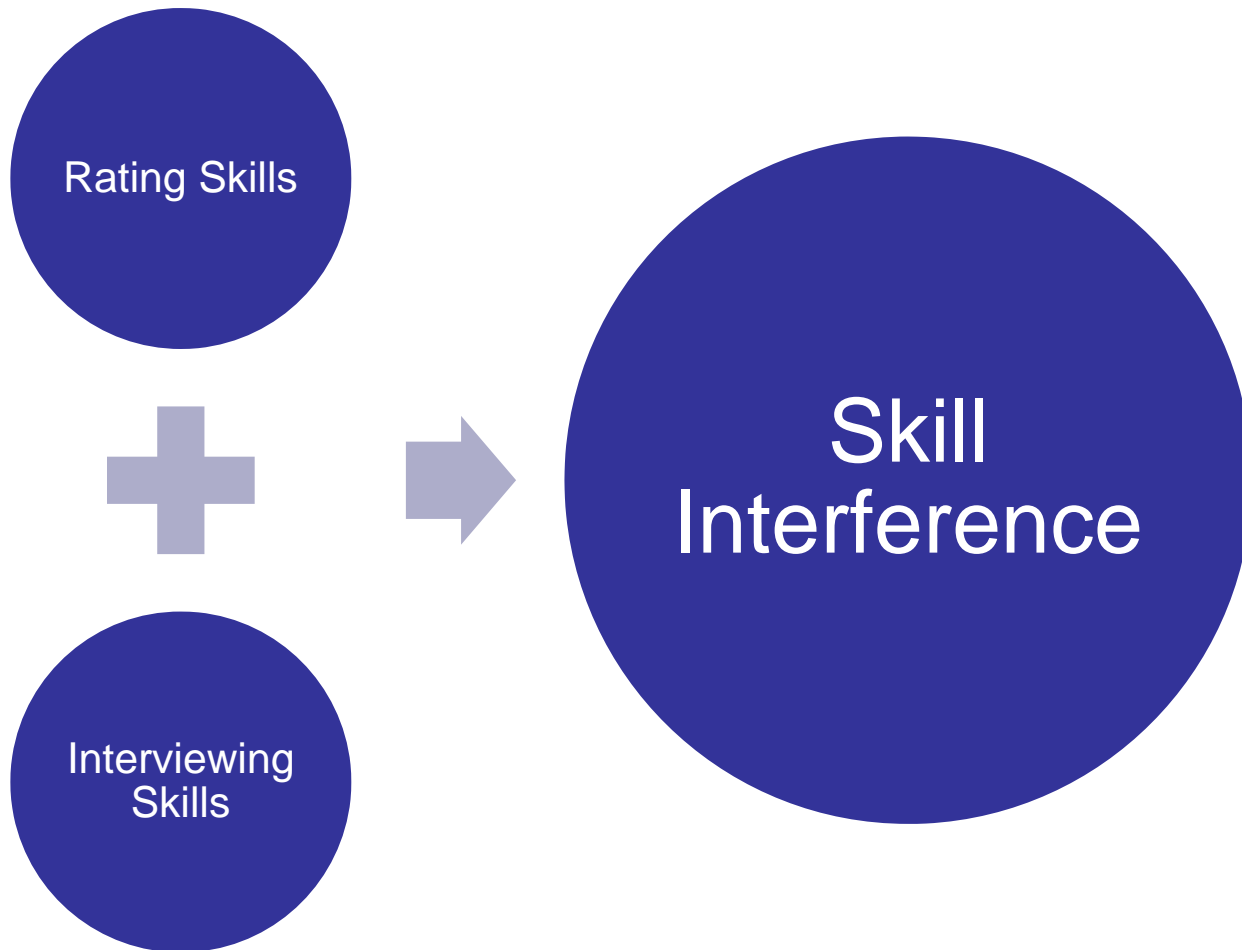
Key Findings – Learning

- Significant increase from pre-workshop ($M = 4.58$) to post-workshop ($M = 5.90$) for task self-efficacy
- Significant increase from pre-workshop ($M = 7.56$) to post-workshop ($M = 9.81$) on the knowledge test
- Significant increase in accurate ratings from early (44%) to late (52%) in training

Key Findings – Certification

- Intention to pursue certification unchanged by training
- Actual certifications: 30%
- Who gets certified?
 - High intent to certify ($r = .30, p < .01$)
 - High task self-efficacy ($r = .15, p < .01$)

Training Study Insights



Training Study Insights

- Certification process is a major bottleneck in creating a pool of certified OPI testers
- Mainly because certification candidates must find & interview people at each of the proficiency levels
 - Likely preventing many Workshop participants from applying for certification or completing process once begun
- The focus on both rating and interviewing skills in the initial certification process:
 - Increases barriers to entry
 - Can decrease motivation to become certified
 - Increases the time required to produce a certified tester
 - Increases the costs and complexity of the process

OPI Assessment Workshop Training Changes

- Days 1 and 2 focus on learning to listen and rate
 - Variety of standardized activities
 - Determining at level v. below level performance
 - Differentiating between Mid and Low sublevels
 - Differentiating between sustained and unsustained performance
- Days 3 and 4 focus on elicitation and rating
 - Practice testing does not begin until Day 3

Certification Process Changes

- New certification categories:
 - OPI Rater
 - OPI Tester
- New OPI Rater certification procedures:
 - Personal OPI
 - Online Listen and Rate Practice Round
 - Online Listen and Rate Certification Round
- New OPI Tester certification procedures:
 - Must first be a certified OPI Rater
 - Conduct Practice Round interviews and receive feedback
 - Conduct Certification Round interviews

On-Going Research Projects

On-Going Research

- OPI Assessment Workshop training evaluation
- OPIc training workshop evaluation
- Who goes on for OPI/OPIc certification? Who works as official rater/tester for ACTFL?
- OPIc Validation studies

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) or contact Dr. Eric A. Surface (esurface@swa-consulting.com) or Dr. Stephen J. Ward (sward@swa-consulting.com).