Foreign Language Annals
Published by the American Council on the Teaching of Foreign Languages

Comparing the OPI and the OPIc: The Effect of Test Method on Oral Proficiency Scores and Student Preference

Gregory L. Thompson
Brigham Young University

Troy L. Cox
Brigham Young University

Nieves Knapp
Brigham Young University

Abstract: *While studies have been done to rate the validity and reliability of the Oral Proficiency Interview (OPI) and Oral Proficiency Interview–Computer (OPIc) independently, a limited amount of research has analyzed the interexam reliability of these tests, and studies have yet to be conducted comparing the results of Spanish language learners who take both exams. For this study, 154 Spanish language learners of various proficiency levels were divided into two groups and administered both the OPI and OPIc within a 2-week period using a counterbalanced design. In addition, study participants took both a pre- and postsurvey that gathered data about their language learning background, familiarity with the OPI and OPIc, preparation and test-taking strategies, and evaluations of each exam. The researchers found that 54.5% of the participants received the same rating on the OPI and OPIc, with 13.6% of examinees scoring higher on the OPI and 31.8% scoring higher on the OPIc. While the results found that students scored significantly better on the OPIc, the overall effect size was quite small. The authors also found that the overwhelming majority of the participants preferred the OPI to the OPIc. This research begins to fill important gaps and provides empirical data to examine the comparability of the Spanish OPI and OPIc.*

Key words: *Spanish, language proficiency, oral proficiency, postsecondary*

Gregory L. Thompson (PhD, University of Arizona) is Associate Professor of Spanish Pedagogy, Brigham Young University, Provo, Utah.
Troy L. Cox (PhD, Brigham Young University) is Associate Director of Research and Assessment, Center for Language Studies, Brigham Young University, Provo, Utah.
Nieves Knapp (PhD, University of Oviedo, Spain) is Associate Teaching Professor of Spanish Pedagogy, Brigham Young University, Provo, Utah.

Foreign Language Annals, Vol. 49, Iss. 1, pp. 75–92.
DOI: 10.1111/flan.12178

Introduction

> Being self-taught is no disgrace; but being self-certified is another matter.
> —Hugh W. Nibley

In the modern globalized economy, many educational institutions need to demonstrate their students' developing levels of foreign language proficiency. These situations include, for example, obtaining U.S. teaching certifications, measuring language gains from study abroad programs, and reporting on course learning outcomes. In addition, organizations that seek employees who can use a non-English language in particular ways for particular purposes, such as those who conduct business internationally and government agencies, also find it essential to ensure that future employees' level of language proficiency is well suited to their expected tasks and responsibilities.

When framing conversations about proficiency, the ACTFL Proficiency Guidelines have served as a common, external point of reference that allows educators to define and discuss various levels of functional language ability. Based on these guidelines, the Oral Proficiency Interview (OPI) has long been considered the gold standard for measuring proficiency levels. Because cost and scheduling constraints can make the OPI impractical in some situations, a computer-based alternative, the Oral Proficiency Interview–Computer (OPIc), was developed. While some studies have examined the comparability of the two exams in English (Surface, Poncheri, & Bhavsar, 2008; SWA Consulting Inc., 2009), little research has been done to compare the two versions of the exam in other languages, and questions about the extent to which the OPI and OPIc are really commensurate remain. Specifically, one might ask to what extent both assessments result in the same rating for any particular test taker. Similarly, can it be assumed that both assessments measure exactly the same constructs and thus, does the OPIc truly represent a suitable, low-cost, easy-to-schedule substitute for the OPI? In addition, because the two versions of the assessment are delivered in different ways and support differing degrees of interpersonal participation, it is important to consider the extent to which participants prefer one test method over another and why. For example, investigating aspects of the OPI or OPIc that are particularly beneficial, enjoyable, or distasteful to test takers provides important insights into the nonmonetary benefits of one delivery system vs. the other.

Background

Development

The OPI evolved from work done on language proficiency testing in the 1950s by the Foreign Service Institute of the U.S. Department of State. In the early 1980s, the ACTFL adapted the government language standards, resulting in the first set of foreign language proficiency guidelines that were specifically geared toward educators (ACTFL, 2012). Based on those guidelines, the ACTFL created an oral proficiency interview protocol that could be used to assess speaking proficiency (Liskin-Gasparro, 2003). The resulting ACTFL OPI is a face-to-face or telephone interview between a trained interviewer and an examinee. The interviewer elicits speech functions that reflect the different proficiency levels, as outlined in the guidelines, and strategically adapts the topics and probes so that each examinee receives a personalized, but comparable, assessment of his or her skill level. Based on the speech sample that is elicited, the interviewer determines the proficiency level (Novice to Superior[1]) and, for Novice, Intermediate, and Advanced speakers, the sublevel (Low, Mid, and High) (Malone, 2003). A second certified rater subsequently evaluates the speech sample, and any discrepancy between assigned scores is moderated by a third certified rater.

The OPI has generally been found to be a reliable test of language proficiency

(Abadin, Fried, Good, & Surface, 2012; Surface & Dierdorff, 2003), and since the 1980s, tens of thousands of interviews have been performed. Although Norris (2001) raised some questions as to the validity of computerized language tests, especially those that attempt to measure oral proficiency, and although the assessment is not without other critics (see Chalhoub-Deville & Fulcher, 2003; Liskin-Gasparro, 2003), the interrater (e.g., raters agree with the relative ordering of a group of individuals from low to high) and test-retest reliability for either test modality (OPI or OPIc) have generally been found to be quite high. As a result, the OPI has become one of the most widely accepted and administered instruments for measuring second language speaking proficiency in the United States, and it is consistently used in high-stakes situations by government, businesses, and educational institutions.

Because the OPI is administered face to face or in person over the telephone by trained interviewers, the associated costs are often high. In order to mitigate those costs and thus make the test both less expensive and easier to schedule and administer, the OPIc was developed in the early 2000s for Novice through Advanced speakers. In part due to the increase in the number of heritage and native speakers as well as increasing proficiency levels in nonnative speakers, in 2012, the ACTFL adapted the OPIc to provide ratings through the Superior level. Unlike with the OPI, prior to beginning the OPIc, test takers complete a survey that requires them to provide demographic information, select specific topics of interest (e.g., sports, politics, literature), and indicate from a list the tasks and situations in which they can use the language. Based on responses to the survey, an individualized test that takes into consideration the test taker's self-reported proficiency level and interests is then created from among the pool of items in the question bank. In this manner, the OPIc is similar to the OPI in that each test is personalized, although obviously not to the same extent. Like the OPI, the OPIc yields a ratable speech sample that is subsequently double-blind-rated by at least two certified OPIc raters. The use of the computer not only facilitates the administration of the exam but also costs less than half of the traditional OPI (ACTFL, 2009a, 2009b).

Reliability: OPIc

While the increased convenience and decreased cost of administering the OPIc relative to the OPI are obvious, a limited amount of research has examined the OPIc's reliability. SWA Consulting (2009) analyzed a sample of 2,934 Korean ESL speakers who took the OPIc multiple times during a 30-day period. The report indicated high test-retest reliability with an *r* (i.e., Pearson Product Moment Correlation) above 0.90 (ranging from 0.90 to 0.93), an *R* (i.e., Spearman Rank Order Correlation) above 0.90 as well (ranging from 0.90 to 0.94), and rater agreement (different raters award the same rating to the same test taker) above 85% (85–92%) for the first two OPIc administrations. However, score stability did decrease slightly as the time gap between administrations increased (SWA Consulting Inc., 2009, p. 2). Abadin et al. (2012) conducted a similar study and reported high levels of interrater reliability across Spanish, English, and Arabic OPIc administrations (*R* ranging from 0.95 to 0.97) and over time (*R* ranging from 0.96 to 0.97). Levels of interrater reliability, as measured by absolute score agreement, were higher than 70% across all languages (English = 80%, Spanish = 80%, Arabic = 71%). At the sublevels (High, Mid, Low), adjacent agreement rates were more than 90% across all main levels (Novice, Intermediate, Advanced). The lowest absolute rates of interrater agreement (75% overall) were found in the Advanced category: While the highest rate was at Advanced Mid (83%), agreement rates at the Advanced Low and Advanced High sublevels were substantially lower (66 and 60%,

respectively) (pp. 2, 10–11). Taken together, these studies indicate that ratings of speech samples that are collected using the OPIc do seem to be reliable. However, because, as seen with the Arabic OPIc (Abadin et al., 2012), some languages may present lower rates of interrater agreement or other concerning trends, researchers still need to examine versions of the OPIc across a broader range of other languages. The curious drop in interrater agreement at the Advanced proficiency level also needs to be better understood.

Concurrent Validity: OPI and OPIc

The OPIc has often been presented and perceived as "[A] different modality of the ACTFL OPI and not a different assessment" (Surface et al., 2008, p. 8). Unfortunately, the literature supporting this claim has been sparse, and no definite conclusions can be drawn without more extensive, experimental studies; to our knowledge, only one report has been published specifically comparing participants' scores across the two exam formats (Surface et al., 2008). In order for the OPI and OPIc to be considered truly equivalent forms of the same assessment that measure the same underlying body of knowledge and skills, research should show high levels of correlation between examinees' scores on the OPI and the OPIc, as well as comparable levels of test-retest reliability, interrater reliability (e.g., raters agree with the relative ordering of a group of individuals from low to high), and rater agreement (different raters award the same rating to the same test taker).

Surface et al. (2008) published a report on two studies that were undertaken on behalf of Language Testing International. Study 1 was designed to test the reliability and validity of the OPIc. Study participants ($n = 99$) were employees at a Korean company and had varying levels of experience with English. Participants were randomly assigned to two groups and completed a pre- and postassessment survey, as well as an OPI and an OPIc (in different orders, depending on group assignment). Those assigned to the second group took the OPIc twice in order to allow for an examination of test-retest reliability (Surface et al., 2008). The authors found the following results:

- There was a significant correlation between final scores on the OPI and the first administration of the OPIc (Pearson's $r = 0.92$, Spearman's $R = 0.91$).
- Absolute agreement for final scores on the OPI and the OPIc (first administration) was at 63%. Agreement within major categories (Novice, Intermediate, etc.) was at 85%. When considering both absolute (e.g., the same rating) and adjacent agreement (e.g., a rating within one sublevel such as Advanced Low and Advanced Mid), final scores agreed 98% of the time.
- The order in which the test was administered had no significant effect on test scores; however, the self-assessment task did seem to affect final ratings, with participants who self-assessed at the lowest proficiency level tending to score lower on the OPIc than on the OPI.
- The test-retest reliability for the first and second administrations of the OPIc was high ($r = 0.94$, $R = 0.91$).
- Confirmatory factor analysis methods supported the reliability and validity of the OPI and OPIc. (p. 30)

Although many of the results strongly supported the comparability of the OPI and OPIc, the findings suggested some areas of obvious concern. First, the level of absolute agreement (63%) was much lower than expected (Surface et al., 2008). Although perfect agreement at the sublevel is rare, 63% seems too low to claim equivalence between test modalities. However, as noted previously, when adjacent agreement was included, rates did jump to 98%.

Study 2 took place after the ACTFL updated the OPIc in an effort to address the areas of concern suggested in Study 1. A sample of 27 participants from the same

Korean company was selected, all of whom took the OPI and the revised OPIc within a 48-hour period (Surface et al., 2008). These were the key results:

- Interrater reliability for the OPIc was found to be generally high ($r = 0.86$, $R = 0.85$, $G = 0.93$); however, interrater reliability for the OPIc was slightly less robust than for the OPI, and exact agreement between raters 1 and 2 was only 58%.
- The correlation between final scores on the OPI and revised OPIc was strong ($r = 0.97$, $R = 0.95$), and final ratings exactly agreed 87% of the time. If adjacent agreement within major categories was taken into account, agreement levels were at 100%. (Surface et al., 2008, pp. 36–37)

Once again, although some areas were strong (e.g., correlation between final scores, adjacent agreement levels, etc.), others (e.g., absolute agreement) were weaker than expected. Specifically, measures of interrater reliability were somewhat lower than in Study 1, and the small sample size prevented any definitive conclusions, especially regarding the effect of the self-assessment task.

Other Issues

In addition to questions about the extent to which the OPI and OPIc result in the same ratings both by different raters and across testing sessions and the extent to which they measure the same body of knowledge and skills albeit in different ways, other differences between the OPI and OPIc merit consideration. First, the effect of self-assessment on the selection of prompts that are offered on the OPIc and the impact of prompt selection on candidates' final proficiency rating remain unclear. For example, based on data from the Computerized Oral Proficiency Instrument (COPI) and the Simulated Oral Proficiency Interview (SOPI), Malabonga, Kenyon, and Carpenter (2005) found that although self-assessment scores were shown to correlate well with independent assessments and COPI/SOPI scores, there was some indication that participants who rated themselves at a higher level of proficiency than was warranted may have been negatively affected by starting their testing session at too high a level (Malabonga et al., 2005). This affected the ability of the raters to establish a floor indicating what the examinees were able to accomplish, as the prompts were too difficult.

What is more, issues of test-taker preference should be investigated. As is the case with research regarding the comparability of scores on the OPI and OPIc, the literature that deals with test examinees' preferences is also in its infant stages. Surface et al. (2008) found that although attitudes toward the OPIc were generally positive, test takers preferred the OPI to the OPIc and felt that the OPI offered a better opportunity to demonstrate their language abilities (Surface et al., 2008). Although no other investigation has been published comparing attitudes toward the OPI and the OPIc, research dealing with computerized vs. face-to-face versions of proficiency tests has supported Surface et al.'s findings: Kiddle and Kormos (2011) found that test takers preferred the face-to-face test to the computerized version, and they also described the face-to-face test as more "fair" than the computerized test. While test takers in Kenyon and Malabonga's (2001) study also felt that the OPI allowed for a better demonstration of ability, they reported that the OPI was not as fair as the COPI/SOPI. Contrary to some previous studies, Mousavi (2009) found that test takers had very positive perceptions of digitally delivered proficiency tests and that they actually felt more comfortable with the computerized test than the face-to-face modality. Research about test-taker preferences in assessment modality as well as the potential relationship between test-taker preferences and final ratings is needed.

In summary, while studies have generally shown high levels of interrater and

test-retest reliability when scoring the OPI or OPIc, research comparing the resulting scores, the impact of assessment modality, self-assessment and topic selection on those scores, and test-taker preferences is still in its infancy and is made more complex given the large number of languages in which the OPI and OPIc are offered. Surface et al. (2008) advocated for the need of additional studies to be conducted "to add to the foundation of evidence supporting the ACTFL OPIc testing modality" (p. 34). To better understand these issues, this study addressed the following research questions:

1. What is the effect of test method (OPI or OPIc) on the speaking proficiency scores of Spanish language learners?
2. Which test method do participants prefer, and why?

Methodology

Participants

One hundred fifty-four students (81 females and 73 males) at a large private university in the western United States participated in the study. The mean participant age was 23.4, with a standard deviation of 5.29. The sample included approximately 54 introductory students in lower-division classes as well as 100 students in upper-division courses including Spanish majors, Spanish translation majors, Spanish teaching majors and minors, and students from other majors who were participating in the university's language certificate program. Mean years of study were 3.48. Most of the study participants in the upper-division classes had significant experience with the language through living abroad or in formal study abroad programs. All of the students in these upper-division programs were required to score Advanced Low or higher on the OPI in order to qualify for their respective certificate, minor, or major. The university routinely covers the cost of the OPI for the first examination, and those who do not score at or above Advanced Low must take it again at their own expense. Although participation in the study was voluntary, several participation incentives were offered: All costs associated with taking the second test (either the OPI or the OPIc) would be paid for by a grant received by the researchers, students could choose the higher of their two test scores if there was a disparity between their OPI and OPIc final ratings, and participants would receive a nationally recognized certificate of their speaking proficiency.

Instruments

Four instruments were used: two oral proficiency tests (the telephone version of the OPI and the OPIc); a presurvey, which test takers completed before taking either assessment; and a postsurvey, which they filled out after having completed both assessments. The presurvey consisted of questions designed to elicit students' relevant background information, language experience, goals, and familiarity with the OPI and OPIc. Participants were also asked to rank their own proficiency from poor to excellent in the different language modalities. The postsurvey consisted of questions designed to elicit participants' experiences with, and attitudes toward, both testing methods and also asked students to rate what they thought their score would be based on the ACTFL Proficiency Guidelines. Participants also expressed their attitude toward each test modality on a nine-point Likert scale ("dislike extremely" to "like extremely") and were asked to explain which test format they preferred and the reasons why in response to a final open-ended question ("As a test taker, which method did you prefer and why?").

Procedures

In order to control for a potential order effect, students were randomly assigned to two groups. The first group (41 females and 36 males; average age = 23.2, $SD = 4.21$; mean years of study = 3.51, $SD = 1.92$) took the OPI before the OPIc; participants

in the second group (40 females and 37 males; average age = 23.6, SD = 6.23; mean years of study = 3.51, SD = 1.92) took the OPIc and then the OPI. Both tests were administered according to student availability but within a 2-week period so as to minimize the potential effect of additional instruction and other interactions on the students' oral proficiency. With only a few exceptions, the postsurvey was completed before students received their final ratings, thus avoiding a score-based preference toward either test. For those cases in which there was a two-sublevel difference in a participant's OPI and OPIc scores ($n = 5$), participants were contacted and asked to answer further questions about their testing experience.

Data Analysis

A mixed-methods approach was used. While many educational researchers use parametric statistics with test scores that may or may not have been verified to be interval data, using both parametric and nonparametric analysis to obtain information that responds to these research questions can ensure that the findings are not an artifact of a failure of data to meet the strict assumptions of parametric research.

First, in order to determine the effect of test method (OPI vs. OPIc) on the speaking proficiency scores of Spanish language learners, the fact that the scores were ordinal rather than interval or ratio in nature had to be accounted for. In other words, the test scale ranked test takers on a scale from 1 to 10 (Novice Low, Novice Mid, and so forth to Superior), without distinguishing or defining the relative distance between each unit on the scale. This meant that the amount of language ability required to move from 9 (Advanced High) to 10 (Superior) on the OPI/OPIc scale was exponentially greater than what was required for a learner to move from 1 (Novice Low) to 2 (Novice Mid). Given that the relative distances between each point on the scale were not equal, both nonparametric (e.g., the Wilcoxon Signed-Rank Test) and parametric (the repeated-measures ANOVA) statistics are reported here. The Wilcoxon Signed-Rank Test compared the relative rank of the same group of subjects on two measures. Using the repeated-measures ANOVA with a counterbalanced design controlled for an order effect (e.g., scoring higher on the second test) by balancing out that effect between the two groups. The dependent variables were test scores on OPI and OPIc, and these variables were analyzed. First, the within-subjects effect was analyzed by comparing individual participants' scores on the two tests. Second, the between-subjects effect was ascertained by comparing mean group scores and standard deviations for each test method and examining the confidence intervals.

To determine which test method (OPI vs. OPIc) participants preferred, students' Likert-scale ratings of the two test modalities on the postsurvey were compared using the Friedman rank order test. Examinees' open-ended responses were grouped into broader categories (e.g., more personal, more realistic, easier to understand, no embarrassment) for easier qualitative analysis.

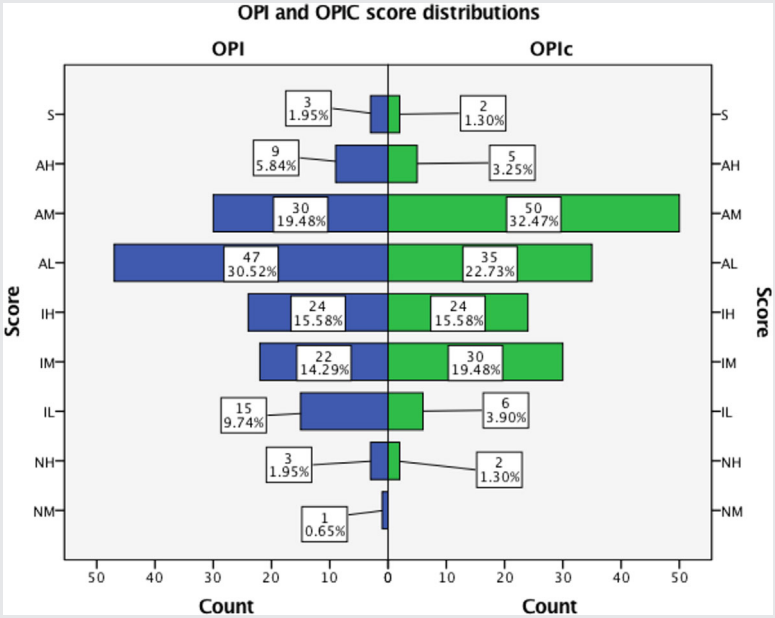
Findings

The Effect of Test Method

To understand the effect of test method on scores, the scores of the test takers' OPI were compared to their OPIc. The mean of the OPI (average = 6.52, $SD = 1.56$) was lower than the mean of the OPIc (average = 6.71, $SD = 1.41$). While the mean of the OPIc was higher, the distribution of OPIc scores was more skewed than the distribution of the OPI scores. In Figure 1, the population distribution is presented with a histogram distribution of the OPI scores presented vertically on the left side of the graph and a mirrored histogram distribution of the OPIc scores with the same examinees on the right. When comparing rank orders of the test takers on the two tests, one can see that 54.5% ($n = 84$) of examinees scored the same on both measures, 13.6% ($n = 21$)

FIGURE 1

Histogram Illustrating OPI and OPIc Score Distribution



scored higher on the OPI, and 31.8% ($n = 49$) scored higher on the OPIc. The Wilcoxon Signed-Rank Test found that the groups were statistically different in terms of mean rank ($z = -3.13$, $p = 0.002$); however, the adjacent agreement was 97%. The Wilcoxon statistic was unable to control for a potential test order effect, necessitating the need for further analysis.

Descriptive statistics showed that Group 1 (those who took the OPIc first) scored higher on both the OPIc and the OPI than Group 2 (see Table 1). However, both groups received higher scores on the OPIc than they did on the OPI, indicating that there was no ordering effect. To verify this observation, a repeated-measures ANOVA was conducted. The between-subject variable was Group (Group 1: OPIc First, Group 2: OPI First), and the within-subject variable was Test Method (OPIc or OPI) with the dependent variable being the numerical conversion of the OPI/OPIc score. The mean difference between groups was

TABLE 1

Descriptive Statistics of OPI and OPIc Results

	OPI First Group		OPIc First Group		Total	
	OPIc	OPI	OPIc	OPI	OPIc	OPI
N	77	77	77	77	154	154
Mean	6.57	6.36	6.86	6.69	6.71	6.53
SD	1.39	1.67	1.43	1.46	1.41	1.57

0.305, 95% CI [−0.16, 0.77], and it was not found to be significant (F (1,152) = 1.71, $p = 0.193$) with a negligible effect size (partial $\eta^2 = 0.011$). This indicated that the results were not confounded by test order. The within-subject variable showed that the mean difference between OPI/OPIc scores was 0.19, 95% CI [0.07, 0.30], and it was found to be significant (F (1,152) = 10.44, $p = 0.002$), although the effect size was still small (partial $\eta^2 = 0.06$).

Test Method Preference

To better understand students' test method preferences, the postsurvey results were analyzed. One hundred forty (91%) of the examinees completed the postsurvey, although some only answered some of the questions. The Likert scale ranged from 1 ("dislike extremely") to 9 ("like extremely"). For the question about the OPI, the mean was 6.86 ($n = 140$, $SD = 1.78$), which fell between "like slightly" and "like moderately" (see Figure 2). For the question on the OPIc, the mean was 5.27 ($n = 139$, $SD = 2.08$), which fell between "neither like nor dislike" and "like slightly" (see Figure 2). To see if the difference was statistically significant, the nonparametric Friedman rank order test was performed, as it is used with repeated measures. This test revealed a significant difference between preference χ^2 (1) = 40.01, $p < 0.001$, with examinees preferring the OPI overall.

The examinees were then divided into three groups: those who preferred the OPI, those who preferred the OPIc, and those with no preference. The groups were determined by looking at the preference rating for each test and subtracting the OPI score from the OPIc score. Those with positive scores were placed in the OPI preference group ($n = 94$), those with negative values were in the OPIc preference group ($n = 25$), and those with 0 were in the no-preference group ($n = 20$). Figure 3 displays a scatterplot of the Likert scale responses with OPI

FIGURE 2

Distribution of OPI vs. OPIc Preference Using Likert Scale

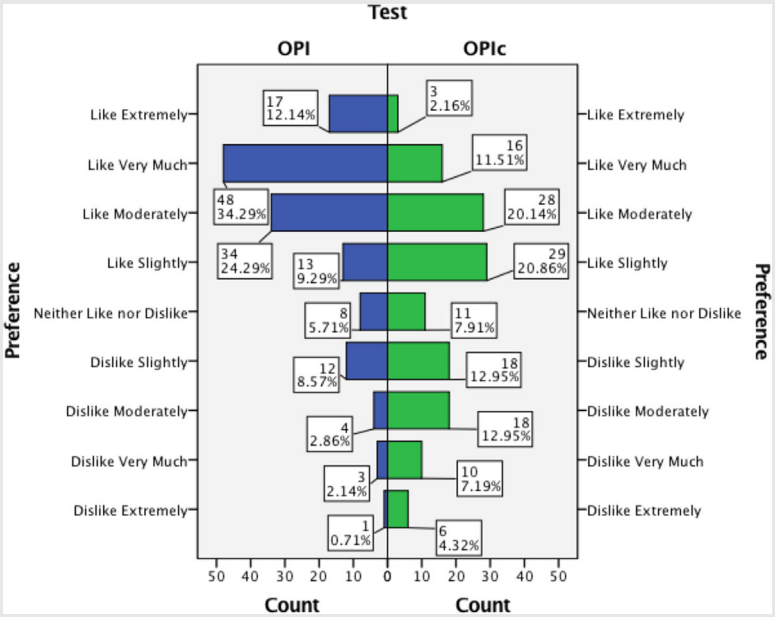
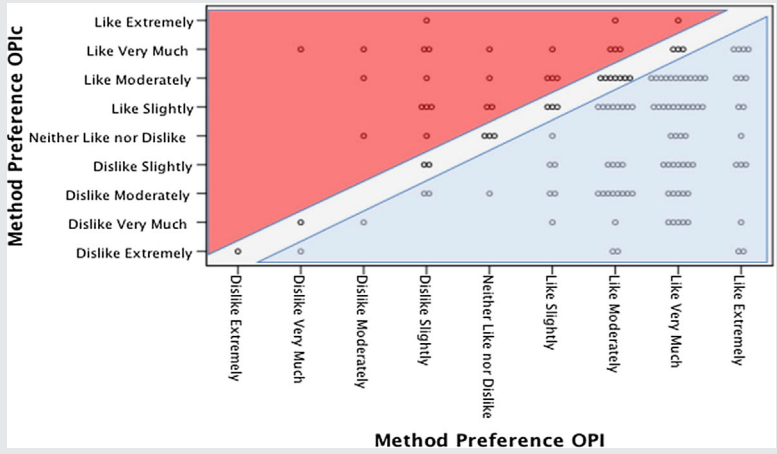


FIGURE 3



preference group in the lower right corner, the OPIc preference group in the upper right corner, and the no-preference group along the diagonal.

This figure shows some interesting trends. First, there were only 20 participants (14.3%) who were ambivalent about the version of the test. These participants varied from disliking both exams on the lower end to having no preference in the middle to those who were satisfied and liked both exam formats. This indicates that 119 participants (85%) liked one test format more than the other (see Table 2). Second, there were 39 participants (28%) who liked the OPI ("like slightly" to "like extremely") but disliked the OPIc to some degree, but there were only seven participants (5%) who liked the OPIc but disliked the OPI, indicating that there might be a stronger preference for in-person interviews.

To ascertain why examinees preferred one test to the other, the responses to the open-ended question were analyzed. A theme analysis was performed on the comments to determine important commonalities and differences. The responses were then coded based on the categories (themes) established and assigned to the appropriate category or categories, as some of the participants' comments addressed more than one established category.

A total of 132 students explained their level of satisfaction with the OPI and OPIc in response to the open-ended question "As a test taker, which method did you prefer and why?" The majority of the students (94 students; 71%) preferred the OPI to the OPIc. Only 27 (21%) students expressed a preference for the OPIc, and 11 (8%) had no preference about the form of the exam. One hundred twenty-eight of the 132 students also wrote comments explaining the reason for their ratings. Four (3%) of the participants did not provide a reason for their rating. Of those who preferred the OPI, a total of 90 comments were offered, which are presented in order of frequency below:

1. More natural ($n = 78$; 87%). Many students stated that the OPI felt more like a conversation, since they were talking in real time to an actual person. This "more natural" feeling apparently impacted their view of the OPI as a better measure of actual oral proficiency. Some of the participants felt that they were better able to understand a live person and sometimes struggled when the computer

TABLE 2

Matrix of OPIc vs. OPI Preferences

		OPIc									
		Dislike extremely	Dislike very much	Dislike moderately	Dislike slightly	Neither like nor dislike	Like slightly	Like moderately	Like very much	Like extremely	Total
OPI	Like extremely				1			1	1		3
	Like very much		1	1	2	1	1	3	3	4	16
	Like moderately			1	1	1	3	7	12	3	28
	Like slightly				3	2	3	8	11	2	29
	Neither like nor dislike			1	1	3	1		4	1	11
	Dislike slightly				2		2	4	7	3	18
	Dislike moderately				2	1	2	8	5		18
	Dislike very much		1	1			1	1	5	1	1
	Dislike extremely	1	1					2		2	6
	Total	1	3	4	12	8	13	34	48	16	139

was speaking. In addition, the participants felt that the interviewer could elaborate and thus make the questions more understandable. These were some of the comments:

- "*I much* preferred the OPI because it felt a lot more natural and organic having a real person asking you real questions that were relevant to my life, affirming my statements, commenting on what I said, and forming questions based off what I said." (emphasis in original)
- "I preferred the OPI because I feel that it was a better measure of my Spanish speaking skills since I will rarely find myself in a situation where I will need to use my Spanish in order to speak to a computer."
- "I liked speaking to a live person rather than speaking to a computer. It made it easier to talk and to understand the questions being asked."
- "I preferred the OPI method because I took the test more seriously when speaking to a real person. I was also able to ask for clarification on questions or phrases I did not understand."
- "I prefer the OPI. The conversation came more naturally. Plus, the fact that the interviewer was able to relate each question to my previous response allowed an easier transition from topic to topic. This also allowed me think for myself which vocabulary I wanted to use and which direction I'd like to take the conversation. The experience with the OPI seemed more natural and realistic, which I liked."

2. Feedback ($n = 40$; 44.5%). The students found that the immediate feedback from the interviewer made them feel more comfortable and helped them communicate more effectively. Even though ACTFL-trained OPI raters do not assist interviewees as they look for words and phrases, the students were still able to gauge whether or not they were being understood thanks to the feedback interviewers offered. The students wrote the following:

   - "Taking the OPIc I got a little tired of talking with no one listening, even though I knew that my responses would be graded later. I guess just not having a real person give any dynamic feedback in the moment was what caused that."
   - "I quickly grew tired of the OPIc because I didn't receive any feedback during the interview. For me, that feedback is important, even if it is just, 'Yes' or 'No.'"
   - "I preferred speaking to a live interviewer because she could respond with small vocal cues mid-sentence to confirm that she was understanding the thrust of my idea. These included things like, '*Sí*,' 'Mm-hmm,' '*Claro*,' and even simply the audible breathing cues that help us navigate when each person will speak."

3. Better topics ($n = 13$; 14%). The OPIc uses a presurvey to determine the speaker's topics of interest and then chooses questions from a database related to those topics. While this is done to tailor the exam to the students, some students still found that the OPI provided more interesting discussion topics. The students stated the following:

   - "I preferred the OPI because it was easier to guide the conversation to topics that I feel comfortable talking about. . .. Also, with the OPI, you can simply tell the tester that you don't like that question (they tell you that at the beginning), and they'll ask you something different."
   - "While taking the OPI, the interviewer was able to redirect me when I was not answering the question with enough detail, etc. I was also able to talk to the interviewer about things that are more personal to me (my job, my classes,

my hobbies) because there was a dynamic conversation going on rather than a question-and-answer session."
- "I liked the OPI because I could talk about the things that I wanted to talk about. If I didn't know much about a subject, we would change the subject."

4. More time ($n = 4$; 4%). The students felt that the artificial time limit set by the OPIc sometimes interrupted them as they were speaking and did not allow them to finish their ideas. Having the human interviewer allowed them to use the necessary time to complete their ideas.

    - "OPI. I wasn't just randomly cut off with time. On the OPIc that was really annoying. Not knowing the time limit on each question to get what I wanted to say across. The OPI allowed the interviewer to help me get a better idea."
    - "The OPI method felt more personal, even though I feel I did worse. I feel I could actively engage in the role-plays and discussions. In addition, the interviewer gave me time to finish my responses, whereas the OPIc did not always."
    - "The conversation also goes more smoothly—it's not such a brusque transition from topic to topic. The tester naturally moves on to the next topic instead of the tester running out of time like can happen on the OPIc—that happened to me on a few questions and I would panic, wondering if that was a bad thing and what would happen if I hadn't addressed all aspects of the question in the allotted time."

In spite of the fact that the majority of the students preferred the OPI, 27 students explained in the qualitative comments why they preferred the OPIc. Not only did fewer students prefer the OPIc, but the range of reasons also varied considerably and often fell in more than one category. The themes, listed in order of frequency, that emerged for the 27 participants who preferred the OPIc were as follows:

1. Less anxiety due to lack of real personal interaction ($n = 21$; 78%). The majority of people who favored the OPI commented that they preferred having a live person who was able to accommodate their style, questions, and pace. The opposite was true for the students who felt more comfortable talking to the computer due to the total absence of a live person. They attributed the reduced anxiety to the lack of pressure from a specific live person who would be perceived as consciously or unconsciously evaluating each phrase and sentence and noted that the avatar was easier to understand than a live person.

    - "Computer-generated questions are easier for me to calmly compose an answer and give my response. I relate to objects (things) more easily than people."
    - "I preferred the computer. The point of the OPI is to express yourself in Spanish, and talk as much as possible, and this was easier to do with the computer ... just to concentrate on speaking a lot and trying to express myself. I didn't have to worry about being interrupted or saying weird things, I could just blab on and on until the computer automatically stopped me."
    - "The OPIc. I don't know why people would be more uncomfortable speaking to an avatar than to a real person; it seems counterintuitive. Speaking with a real person is incredibly nerve-wracking, especially as it can be difficult to hear them through the phone. The OPIc was always very clear, and I didn't feel worried about what the other person would think of

my answers or how they would respond."
- "I preferred the OPIc because I didn't get as nervous while speaking, helping me to maintain my train of thought. It was easier for me to keep going after I made a mistake."

2. Repeat questions ($n = 11$; 41%). On the OPIc, students could ask the avatar to repeat the questions in a totally nonjudgmental assessment context. Some sample comments included the following:

   - "I actually really liked the OPIc. I didn't feel uncomfortable at all. I thought it was nice that you could repeat the question, especially because some of the questions were really long and involved. And it was nice to be able to move on when you wanted to."
   - "The OPIc actually helped me remain more calm and think through my answers more. I wasn't as nervous, and I could repeat the question with the push of a button."
   - "I preferred the OPIc because there is no chance of the recording questioning your points or ideas. It is also more comfortable to repeat the question as many times as you want to make sure you understand."

3. Flexibility ($n = 7$; 26%). Some students felt less restricted by the OPIc and felt that they had more freedom to express their own ideas and thoughts because no one else was guiding the conversations. This is due in part to the fact that the OPIc uses a presurvey to determine the topics of interest and then selects from a database of questions related to those topics when choosing questions. A few students found that this system benefited them, and they felt more comfortable with these questions:

   - "The OPIc, because I could expound on what I knew how to talk about, rather than trying to answer questions they asked me that I didn't necessarily have a vocabulary for."
   - "I like the OPIc because I felt like I could go off on tangents a bit more because I was talking to a computer. I just talked as much as I could until my time was up. The OPI had good aspects to it as well, but I think I felt less nervous talking to an avatar."
   - "I felt more comfortable speaking to the computer because I got to influence the topics that were discussed instead of having them chosen for me by the interviewer. I liked that my responses were being recorded because then I could express my thoughts and not have an awkward pause if I made a mistake. When I messed up in the computerized test I could fix it if I caught it without disrupting my flow of expression."

A small percentage of students preferred neither the OPI nor the OPIc and rated them the same. Of the 132 participants, 11 (or 8%) either found positive aspects of both exams or found both of them to be problematic:

- "Taking the OPI was easier in the fact [*sic*] to the test being a conversation, which moved you from one question to the next, but the OPIc was easier in the fact that you are not actually talking to someone, so you don't get as nervous."
- "I preferred the flexibility of the OPIc, but I preferred talking to a real person on the OPI. Both had their advantages."
- "I liked both of them. It was really nice to have responses from a real person and be able to have a conversation, but it was nice to be able to talk as much as I wanted in the OPIc. It lessened the pressure to not have someone respond right away. Both were actually pretty fun."
- "I like that the OPI interviewer could reword the question or stop me if I misunderstood him. It felt more like a

conversation, but I hate talking on the phone with people so I get really flustered when I mess up. The OPIc was better for me in that I could talk without feeling self-conscious and I didn't panic as bad. I may have answered questions wrong, but I could talk more fluently without feeling embarrassed."

- "Neither. I hate tests."

The quantitative data from this study reveal that the students performed better on the OPIc but that, in contrast, they demonstrated a strong preference for the OPI. The qualitative data give valuable insights into the why of the students' preferences. The quantitative data and the qualitative data together reveal that different types of learners prefer different assessment experiences and thus favor either the OPI or OPIc.

Discussion

First, the results indicate that 55% of the students received exactly the same rating; 32% scored higher on the OPIc and 13% scored higher on the OPI. Adjacent category agreement was 97%. However, the students who took the OPIc received ratings that were statistically significantly higher on average although the overall effect size was quite small, indicating that even though differences did exist between students' scores on the two versions of the assessment, the practical significance of this difference was minimal. This confirms that both assessments are reliable and valid measurements of oral proficiency as defined by the ACTFL Proficiency Guidelines in spite of small variations in performance across the two exams.

That said, it is important to point out that differences of one sublevel within the major sublevels can be a concern especially since receiving a score of Intermediate High instead of Advanced Low has implications for teacher certification in many states, though in this study only 4 of the 154 examinees (2.6%) would have been affected by this rating difference. Similarly, for dual immersion programs, the difference between a rating of Advanced Low and the required level, Advanced Mid, is personally, if not statistically, meaningful. Thus, the choice of exam format may facilitate achieving the credentials that are required for certain career paths.

It is also interesting to point out that a centralizing tendency with the OPIc was observed; that is, the data suggest that students moved more toward the middle of major levels (Intermediate Mid, Advanced Mid) and away from the lower and higher ends of these major levels. This was especially true of students at the Advanced level. Twenty-three of the 48 (or 48%) students who scored Advanced Low on the OPI scored at the Advanced Mid level on the OPIc. In addition, one of the concerns raised in this study was whether students could be pushed to perform at the Superior level by the OPIc or if a trained OPI rater would be required to encourage candidates to produce the level of speech and language usage that are needed to demonstrate very high levels of proficiency. Of the 13 students who scored Advanced High or Superior with the OPI, 8 (62%) received lower scores on the OPIc, with 2 of those individuals dropping from Advanced High to Advanced Low. Because three participants tested into the Superior level on the OPIc, it is clear that students can place into this level on this version of the assessment; however, a larger percentage of students produced speech samples that were rated Advanced High or Superior when interacting with a live interviewer, when taking the OPI. Thus, it will be important to investigate ways to push Advanced High/Superior border-level speakers to produce more complex speech that exemplifies higher-level functions on the OPIc without the prompting of a human interviewer.

It is also interesting that students overwhelmingly preferred the OPI in spite of the fact that, on average, they scored either the same or slightly higher on the OPIc. In the majority of cases, students in this study found the OPI to be more natural, more enjoyable, and more like a real speaking

exam; they appreciated interacting with a real person who could answer their questions and challenge them to produce the highest level of speech that they were capable of maintaining. Finally, they felt that the OPI provided was a better measure of their proficiency, reinforcing the face validity of the OPI over the OPIc. Interestingly, those who preferred the OPIc, while fewer in number, cited the exact opposite reasons: Talking to a real person increased their level of anxiety and made them feel that they were being judged for their mistakes. While some students found that talking to a computer was completely unnatural, others felt that the computer's nonjudgmental presence and the opportunity to have questions repeated multiple times allowed them to be more relaxed and thus better able to demonstrate their actual speaking ability.

These results clearly indicate that individual students' personal characteristics and preferred interpersonal style must be taken into consideration when selecting an assessment format. Because students' preferences varied widely, preparing students to take oral proficiency exams is critical. The students in this study did not receive any background information or specific instructions about the assessments prior to taking the OPI or the OPIc, other than that one was a telephone interview and the other was computer-based. It is thus unclear if prior knowledge about the structure of the exam and the kinds of speech tasks that were required at each progressive level on the proficiency scale may have worked to students' benefit; for example, perhaps students who had been expressly informed that they must be able to support and defend an opinion rather than simply share personal stories to receive a rating at the higher proficiency levels may have pushed harder to demonstrate the required skills. At the very least, informing students about the opportunity to select questions, that questions could be repeated, and that their speech samples would be limited to a certain number of minutes for each question may have increased their satisfaction with the OPIc.

Finally, the setting in which these learners acquired their second language seemed to have a direct impact on their exam preference. While the majority of the participants preferred the OPI to the OPIc, those participants who had extensive immersion experiences showed an even higher preference for the OPI. Because many of these participants had spent 18–24 months in a foreign country or in a region of the United States where they could be fully immersed in the Spanish language, these participants were more comfortable with and accustomed to speaking in a more naturalistic setting and felt that the truly interpersonal nature of the OPI allowed them to provide a much better representation of their abilities. Interestingly, even though these students preferred the OPI, they tended to be rated slightly higher on the OPIc.

Conclusion

The results from this study inform language educators' understanding of potential differences between the OPI and the OPIc as well as the extent to which the OPIc can serve as a viable substitute for the OPI. Understanding and explaining to students the differences in assessment formats and expectations at each proficiency level, as well as taking test takers' preferences into consideration, will help universities, businesses, and governmental agencies support individuals when selecting the test administration format that best meets their needs, financial means, and personal preferences.

This study also raises many questions as well as possible areas for future research, particularly about the effectiveness of the OPIc for students who are between major levels. In addition, since more than 31% of the participants scored better on the OPIc than on the OPI and more than 13% scored better on the OPI than the OPIc, it will be important to examine in even greater detail the impact of students' personal characteristics, environmental factors, depth of knowledge about and

experience with different exam formats, awareness of the kinds of language that must be demonstrated at each proficiency level, type of prior language acquisition experiences, proficiency level, and attitudes toward interpersonal or technology-based exams to tease out the causes of the slight differences in scores between the two formats. Since this research shows that more than 70% of the participants in this study preferred the more costly version (the OPI), programs can decide whether they have the resources to support individuals who choose the OPI, if they can justify requiring the more expensive version of the assessment, and if they have time to prepare students to complete either form by overtly teaching students at all levels to provide the richest speech sample and fully demonstrate the tasks and use language in the contexts that particularly characterize higher levels of proficiency.

In conclusion, since many important factors play a role in assessing an individual's oral proficiency and since these exams have important implications for candidates' future success in business, education, and government, it is important to consider the accuracy, incurred costs, and satisfaction of those who take them. While this research indicates that both assessments are equivalent measures of oral proficiency, many aspects of the exams and the examinees require further study.

Note

1. Note that while the 2012 Proficiency Guidelines added Distinguished as the highest category, the interview protocol only tests through Superior.

Acknowledgments

We would like to express our gratitude to our research assistants Doug Porter and Steve Stokes for their invaluable help in organizing and analyzing the data for this study, and we would also like to thank the three anonymous reviewers and Anne Nerenz for their valuable comments and insights on the different drafts of this article.

References

Abadin, H., Fried, D., Good, G., & Surface, E. (2012). *Reliability study of the ACTFL OPI in Chinese, Portuguese, Russian, Spanish, German, and English for the ACE review*. Raleigh, NC: SWA Consulting Inc.

ACTFL. (2009a). *Oral Proficiency Interview (OPI): Testing for proficiency* [Electronic version]. Retrieved July 23, 2015, from http://www.actfl.org/professional-development/certified-proficiency-testing-program/testing-proficiency?pageid=3348

ACTFL. (2009b). *Oral Proficiency Interview (OPI): ACTFL Oral Proficiency Interview–computer (OPIc)* [Electronic version]. Retrieved July 23, 2015, from http://www.languagetesting. com/oral-proficiency-interview-opi-2

ACTFL. (2012). *ACTFL proficiency guidelines 2012* [Electronic version]. Retrieved August 15, 2015, from http://www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012

Chalhoub-Deville, M., & Fulcher, G. (2003). The Oral Proficiency Interview: A research agenda. *Foreign Language Annals*, *36*, 498–506.

Kenyon, D., & Malabonga, V. (2001). Comparing examinee attitudes toward computer assisted and other oral proficiency assessments. *Language Learning & Technology*, *5*, 60–83.

Kiddle, T., & Kormos, J. (2011). The effect of mode of response on a semidirect test of oral proficiency. *Language Assessment Quarterly*, *8*, 342–360.

Liskin-Gasparro, J. (2003). The ACTFL proficiency guidelines and the Oral Proficiency Interview: A brief history and analysis of their survival. *Foreign Language Annals*, *36*, 483–490.

Malabonga, V., Kenyon, D. M., & Carpenter, H. (2005). Self-assessment, preparation and response time on a computerized oral proficiency test. *Language Testing*, *22*, 59–92.

Malone, M. E. (2003). Research on the Oral Proficiency Interview: Analysis, synthesis, and future directions. *Foreign Language Annals*, *36*, 491–497.

Mousavi, S. A. (2009). Multimedia as a test method facet in oral proficiency tests.

International Journal of Pedagogies and Learning, *5*, 37–48.

Norris, J. (2001). Concerns with computerized adaptive oral proficiency assessment. *Language Learning & Technology*, *5*, 99–105.

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

Surface, E., Poncheri, R., & Bhavsar, K. (2008). Two studies investigating the reliability and validity of the English ACTFL OPIc with Korean test takers: The ACTFL OPIc validation project technical report. Retrieved August 15, 2015, from http://www.languagetesting.com/wp-content/uploads/2013/08/ACTFL-OPIc-English-Validation-2008.pdf

SWA Consulting Inc. (2009). *Brief reliability report 5: 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*. Raleigh, NC: Author.

Submitted October 7, 2015

Accepted November 30, 2015