ACTFL
Language Connects

Examination Evaluation of the ACTFL OPIc

in Arabic, English, and Spanish for the ACE Review

Stephen Cubbellotti, Ph.D.
Independent Psychometric Consultant

Publication No. AAR/OPIc/ACE/R—2015-001

EXECUTIVE SUMMARY

This report documents the American Council on the Teaching of Foreign Languages (ACTFL) Oral Proficiency Interview – Computer (OPIc®) from 2012 to 2014 to satisfy a review requirement of the American Council of Education College Credit Recommendation Service (CREDIT) program. The ACTFL OPIc® is an on-demand, internet or phone-delivered proficiency test of spoken language ability. An avatar provides a customized series of recorded prompts based on the interests and experience of the test-taker derived from answers to a Background Survey and Self-Assessment at the beginning of the test. The candidate responses are recorded and digitally archived on a secure data base. Completed OPIc®s are evaluated by ACTFL certified raters who assign a holistic score based on the descriptions contained in the *ACTFL Proficiency Guidelines 2012 – Speaking* and the ACTFL Rating Scale. The computerized nature of the OPIc® permits valid and reliable oral proficiency testing on a large scale.

The structure of this document is outlined to address several areas including: general test information, item/test content development, reliability information, validity information, scaling and item response theory procedures, validity of computer administration, and cut-score information.

METHOD

ACTFL and LTI have an extensive collection of resources available publically that document the rigor of defining language competency as well as the precision in their assessments. All documentation cited is publically available and citations for these resources are given in the bibliography at the end of this document. The reliability information section is the only section which contains uniquely generated statistics for the purposes of this study. An outline of the results can be found below.

Given the ordinal nature of the ACTFL proficiency scale and ACTFL OPIc® scores, inter-rater reliability was measured by the Spearman's *rho* correlation, which is a coefficient of reliability appropriate for ordinal data. Inter-rater agreement was measured by the extent to which ratings exhibited absolute (i.e., exact) and/or adjacent (i.e., +/- one level) agreement. The combination of Spearman's *rho* and absolute/adjacent agreement results provides sufficient information about reliability.

Comparisons of ACTFL OPIc® inter-rater reliability and agreement were made across three languages: Arabic, English, and Spanish. Comparisons were also made across language categories (i.e., language difficulty) and interview years (i.e., 2012 to 2014 in this sample). For inter-rater agreement, rater concordance was further investigated by major proficiency level and sub-level.

FINDINGS

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

Overall, the findings support the reliability of the ACTFL OPIc® as an assessment of speaking proficiency. Areas for continued improvement include increasing rater agreement at the Advanced Mid sub-level and the Novice High-Intermediate Low border. Findings are presented in more detail in the report.

Table of Contents

General Test Information

Rationale and Purpose of the test

The American Council on the Teaching of Foreign Languages (ACTFL) Oral Proficiency Interview - computer (OPIc®) is a semi-direct test of functional spoken language proficiency in an internet-delivered interview format. An embodied agent (avatar) is the virtual interviewer. The ACTFL OPIc® is designed to elicit a 20 to 40 minute sample of ratable speech. A Background Survey allows the test taker to select topics from areas of interest within his/her experience. A Self-Assessment survey allows the test taker to select a range of linguistic levels. Based on the selections made by the test taker, a unique and individualized assessment, tailored to linguistic ability, work experience, academic background, and interests, is generated by the computer from an item bank of pre-recorded prompts.

The goal of the instrument is the same as the goal of the ACTFL Oral Proficiency Interview (OPI): to obtain a ratable sample of speech which a rater can evaluate and compare to the *ACTFL Proficiency Guidelines 2012 – Speaking* in order to assign a rating. The current version of the OPIc® measures oral proficiency up to the Superior level on the ACTFL scale. An ACTFL OPIc® is assigned one of the following ratings: Superior, Advanced High, Advanced Mid, Advanced Low, Intermediate High, Intermediate Mid, Intermediate Low, Novice High, Novice Mid, or Novice Low.

The ACTFL OPIc® is appropriate for both small group and large scale testing. Thousands of test candidates can take the test online at the same time. The recording of their responses is made available via a secure "Rater Site" to Certified OPIc® Raters and therefore, can be evaluated by raters within a short period of time. Because of the availability of access to the test, proctors can schedule and administer the OPIc® to test candidates easily, anywhere in the world.

Name(s) and institutional affiliations of the principle author(s) or consultant(s)

- Kathy Akiyama, Ph.D., Mt. Angel Seminary
- Mahdi Alosh, Ph.D, (Ret) Ohio State University
- Bill Prince, Ph.D, Furman University
- Robert Vicars, Ph.D, (Emeritus) Milliken University
- Karen Breiner-Sanders, Ph.D, (Emerita) Georgetown University
- Mildred Rivera Martinez, Ph.D,
- Cindy Martin, Ph.D, University of Maryland
- Irina Dolgova, Ph.D., Yale University
- Ping Xu, Ph.D, Baruch College
  Mei Kong, Ph, D., University of Maryland
- Erwin Tschirner, PH, D, University of Leipzig

Types of scores reported for examinees

Examinees' scores are reported as a major level and sublevel according to the *ACTFL Proficiency Guidelines 2012 - Speaking*. The ACTFL Guidelines describe the tasks that speakers can handle at each level, as well as the content, context, accuracy, and discourse types associated with tasks at each level. The description of each major level is representative of a specific range of abilities. They also present the limits that speakers encounter when attempting to function at the next higher major level. While the *ACTFL Proficiency Guidelines* are comprised of five major levels of proficiency – Novice, Intermediate,

Advanced, Superior, and Distinguished – the current exam only tests through Superior. Together these levels form a hierarchy in which each level subsumes all lower levels. The major levels of Advanced, Intermediate, and Novice are divided into High, Mid, and Low sublevels. ACTFL publically shares their Guidelines for defining the levels of proficiency, describing what examinees have displayed during their examination.

A rating at any major level is reached by confirming the sustained performance across ALL the criteria of the level. The sublevel is determined by the quality of the performance at that level and the proximity to the next higher major level. The assessment criteria used to evaluate the ACTFL OPIc® is provided in the chart below:

Proficiency Level*	Global Tasks and Functions	Accuracy	Text Type
Superior	Support opinions, hypothesize, and deal with topics abstractly.	Errors virtually never interfere with communication or distract the native speaker from the message.	Extended discourse
Advanced	Narrate and describe in major time frames and deal effectively with an unanticipated complication.	Understood without difficulty by speakers unaccustomed to dealing with non-native speakers.	Paragraphs
Intermediate	Create with language, initiate, maintain, and bring to a close simple conversations by asking and responding to simple questions.	Understood with some repetition by speakers accustomed to dealing with non-native speakers.	Sentences
Novice	Communicate minimally with formulaic and rote utterances, lists and phrases.	May be difficult to understand, even for speakers accustomed to dealing with non-native speakers.	Individual words and phrases

Directions for scoring and procedures and keys

Once the OPIc® test is completed, the speech sample is uploaded and saved automatically on a secure Internet site. An ACTFL Certified OPIc® Rater listens to the sample and evaluates the sample according to the Assessment Criteria. Once a preliminary rating is reached, the rater compares the sample to the descriptions in the *ACTFL Proficiency Guidelines 2012 – Speaking* and selects the best match between the sample and description for the rating. The rater enters the rating into the system. The OPIc® is then blindly second rated by another certified OPIc® rater, following the same protocol. If the two ratings agree exactly, the rating is finalized; if the two ratings differ, the OPIc® is assigned to a third rater for a blind arbitration.

ACTFL Certified OPIc® Raters are highly specialized language professionals who have completed a rigorous training process that concludes with a rater's demonstrated ability to consistently rate samples with a high degree of reliability.

Certified OPIc® Raters are expected to respect and follow OPIc® rating protocol. Confidentiality and exclusivity are important practices for all OPIc® Raters. Every rater agrees to respect the rules and regulations regarding OPIc® rating, and the exclusivity of the OPIc® as ACTFL property. Work with the OPIc® rating process must be done exclusively through Language Testing International, the ACTFL Testing Office. Raters are required to follow all OPIc® procedures and guidelines, as well as any other information received on behalf of LTI and ACTFL.

Item/Test Content Development

Specifications that define the domain(s) of content, skills, and abilities that the test samples

The ACTFL OPIc® utilizes a Background Survey. This survey is a questionnaire which elicits information about the test taker's work, school, home, personal activities and interests. The test taker completes the survey and the answers determine the pool of topics from which the computer will randomly select questions. The test taker also completes a linguistic Self-Assessment, comparing his/her perceived ability with samples that are provided in both written and spoken form. The test taker then selects a Self-Assessment level that best reflects his/her ability. Based on the variety of topics and the linguistic level selected by the test taker, a computer algorithm generates appropriate questions that target functions across two contiguous levels and a variety of topics (simulating the iterative process of the ACTFL OPI). The range of possible combinations the computer can generate allow for individually designed interviews. Even if two test takers selected the same combination of Background Survey and Self-Assessment responses, the resulting test would not be the same due to the size of the item bank and the selection algorithm.

The *ACTFL Proficiency Guidelines* describe the tasks that speakers can handle at each level, as well as the content, context, accuracy, and discourse types associated with tasks at each level. They also present the limits that speakers encounter when attempting to function at the next higher major level. Further descriptions of each level are available online.

Statement of test's emphasis on each of the content, skills, and ability areas

The tested content, skills and ability areas are based on the Assessment Criteria for Speaking and the descriptions contained in the *ACTFL Proficiency Guidelines - Speaking*. The ACTFL OPIc® measures how well a person spontaneously speaks language in response to carefully constructed prompts dealing with practical, social, and professional topics that are encountered in true-to-life informal and formal contexts. These tasks range from creating with language, asking questions, story-telling, providing detailed descriptions, producing paragraph-length narrations and descriptions in major time frames, dealing abstractly with current issues of general interest, supporting one's opinion and hypothesizing with extended discourse.

Rationale for the kinds of tasks (items) that make up the test

The tasks of the ACTFL OPIc® reflect the linguistic functions of each of the major levels of proficiency as described in the *ACTFL Proficiency Guidelines 2012 – Speaking*. Test takers are presented with questions that span two or more major levels across a variety of content areas. In this way, the sample that is produced provides sufficient evidence of a speaker's patterns of linguistic strengths (their "floor performance") and weaknesses (their "ceiling").

Information about the Adequacy of the items on the test as a sample from the domain(s)

The *ACTFL Proficiency Guidelines – 2012 – Speaking* describe the range of contents and contexts a speaker at each major level should be able to handle. This was main driver behind the topics generated for each level. Additionally, candidates fill out a Background Survey which elicits information about the test taker's work, school, home, personal activities, and interests. The survey answers determine the pool of prompts from which the computer will randomly select topics for prompts. The variety of topics, the types of questions, and the range of possible computer-generated combinations allows for individually designed assessments. Even if two test takers select the same combination of Background Survey responses, the resulting tests will be different. Based on the Background Survey, questions are pulled that reflect the background and interests of the candidate.

Information on the currency and representativeness of the test's items

The representativeness of the items in a test is guaranteed by providing a diversity of topics, subtopics, genres, domains and rhetorical organization so that the test can provide ample evidence of the proficiency of the test-taker across a broad spectrum of target language use domains.

Some of the topics from which the test-taker may choose include: home, school, free-time activities, sports, work, family, music, travel, etc. New topics are always being developed and old ones revised as they become less current.

Description of the item sensitivity panel review

The use of a Background Survey allows the test taker to avoid the selection of items which may be insensitive or irrelevant for the test taker. In an effort to ensure that test-takers are not offended or made uneasy while taking an OPIc®, item writers are instructed to avoid sensitive topics (e.g., immigration, national origin, sexual preference, religion, marital status, racism, etc.) when developing OPIc® prompts.

Whether and/or how the items pre-tested (field tested) before inclusion in the final form

Since each OPIc® is generated based on the test taker's responses to the Background Survey and Self- Assessment, there is no standard "final form." However, items are pre-tested before they are added to the item pool; items that do not elicit the expected level of response are modified or removed.

Item analysis results (e.g. item difficulty, discrimination, item fit statistics, correlation with external criteria

All OPIc® items target the linguistic tasks, contexts and content areas as described in the *ACTFL Proficiency Guidelines 2012 – Speaking.*

Reliability Information

Previous studies have provided psychometric support for the use of speaking proficiency measures developed according to the *ACTFL Proficiency Guidelines*.

Thompson (1996) presented results from Russian speaking, reading, listening and writing proficiency assessments. The study used two samples of students: one from the University of Iowa and one from the Middlebury Russian Summer program. The inter-rater reliabilities for both the Iowa and the Middlebury samples were statistically significant, Pearson's r = .91 and .72, respectively. Surface, Dierdorff, and Poncheri (2008) found strong support for favorable inter-rater reliability for the OPIc® English version with Korean test takers. Further, the majority of rater pairs were making identical proficiency level judgments when scoring the OPIc®.

SWA consulting (2012) found Spearman Rs exceeded the standard for use, ranging from 0.95 to 0.97 across languages and years analyzed for the OPIc®. In addition, overall inter-rater agreement was higher than 70% for all languages and lowest for Novice High. These results were consistent across languages and highest for Novice-Mid and Superior.

To start, a concordance analysis is seen below. It cannot be used to judge the correctness of measuring or rating techniques; rather, it shows the degree to which different measuring or rating techniques agree with each other.

Note that category names were shortened to fit into the tables below. They follow the following abbreviation:

NL="Novice Low", NM="Novice Mid", NH="Novice High, IL="Intermediate Low", IM="Intermediate Mid", IH="Intermediate High", AL="Advanced Low", AM="Advanced Mid", AH="Advanced High", S="Superior"

Table 1 Concordance Table for Arabic OPIc® from 2012 to 2014

		NL	NM	NH	IL	IM	IH	AL	AM	AH	S
	NL	11	0	0	0	0	0	0	0	0	0
	NM	0	57	12	0	0	0	0	0	0	0
	NH	0	8	83	15	0	0	0	0	0	0
	IL	0	1	24	94	12	1	0	0	0	0
	IM	0	0	3	12	69	9	0	0	0	0
	IH	0	0	0	2	17	51	6	0	0	0
	AL	0	0	0	0	1	7	51	6	0	0
	AM	0	0	0	0	0	0	9	22	4	0
	AH	0	0	0	0	0	0	2	6	23	8
	S	0	0	0	0	0	0	0	0	5	55

Table 2 Concordance Table for English OPIc® from 2012 to 2014

		NL	NM	NH	IL	IM	IH	AL	AM	AH	S
	NL	28	0	0	0	0	0	0	0	0	0
	NM	2	74	7	0	0	0	0	0	0	0
	NH	0	6	114	12	2	0	0	0	0	0
	IL	0	0	8	209	35	0	0	0	0	0
	IM	0	0	1	40	1194	89	17	1	0	0
	IH	0	0	0	1	65	730	82	24	1	0
	AL	0	0	0	0	3	80	339	64	33	2
	AM	0	0	0	0	0	23	50	175	95	21
	AH	0	0	0	0	0	3	20	86	301	89
	S	0	0	0	0	0	0	2	29	70	388

Table 3 Concordance Table for Spanish OPIc® from 2012 to 2014

		NL	NM	NH	IL	IM	IH	AL	AM	AH	S
	NL	31	9	0	1	0	0	0	0	0	0
	NM	10	87	26	1	1	0	0	0	0	0
	NH	0	15	132	39	4	0	0	0	0	0
	IL	0	1	41	289	121	2	0	0	0	0
	IM	0	0	1	139	1125	201	6	0	0	0
	IH	0	0	0	1	232	1554	207	15	0	0
	AL	0	0	0	1	11	293	1459	124	6	0
	AM	0	0	0	0	0	13	147	728	115	9
	AH	0	0	0	0	0	0	6	146	371	46
	S	0	0	0	0	0	0	0	7	92	151

The concordance tables illustrate generally good agreement between the raters as there are no ratings that are strikingly different than one another as seen by the large quantity of 0s in the upper right and bottom left of the rater matrix.

Internal consistency reliability

There are two types of inter-rater reliability evidence for rater-based assessments—inter-rater reliability coefficients and inter-rater agreement (concordance of ratings). Although there are many types of reliability analyses, the choice of a specific technique should be governed by the nature and purpose of the assessment and its data.

Spearman's rank-order correlation (R) is a commonly used correlation for assessing inter-rater reliabilities, and correlations should be at or above .70 to be considered sufficient for test development and .80 for operational use (e.g., LeBreton et al., 2003). Spearman's R is the most appropriate statistic for evaluation of the ACTFL OPIc® data because the proficiency categories used for ACTFL OPIc® ratings are ordinal in nature.

Spearman's rank-order correlation is another commonly used correlation for assessing inter-rater reliability, particularly in situations involving ordinal variables. Spearman rank-order correlation (ρ, rho) has an interpretation similar to Pearson's r; the primary difference between the two correlations is computational, as ρ is calculated from ranks and r is based on interval data. This statistic is appropriate for the OPIc® data in that the proficiency categories are ordinal in nature.

Table 4 Spearman's Correlations by Language from 2012-2014

Language	N	ρ	p
Arabic	686	0.968	<0.001
English	4607	0.958	<0.001
Spanish	8017	0.940	<0.001

Table 5 Spearman's Correlations by Language and Year

Language	Year	N	ρ	p
Arabic	2012	198	.971	<0.001
	2013	229	.952	<0.001
	2014	259	.972	<0.001
English	2012	1643	.947	<0.001
	2013	1584	.950	<0.001
	2014	1389	.959	<0.001
Spanish	2012	1936	.919	<0.001
	2013	2485	.936	<0.001
	2014	3596	.952	<0.001

Overall, the ACTFL OPIc® exceeded inter-rater reliability minimum standards and was quite high. All three OPIc® language exams have a high (around 0.95) Spearman rho correlation. This indicates that there is a strong relationship between the ratings of the two raters. These results are consistent with

previous years' results (Thompson, 1995; Surface & Dierdorff, 2003; SWA Consulting, 2012) providing evidence of acceptable inter-rater agreement for operational use over time.

Evidence for equivalence of forms of the test

The ACTFL OPIc® utilizes a Background Survey. This survey is a questionnaire which elicits information about the test taker's work, school, home, personal activities and interests. The test taker completes the survey and the answers determine the pool of topics from which the computer will randomly select questions. The variety of topics, the types of questions, and the range of possible combinations the computer can generate allow for individually designed interviews. Even if two test takers selected the same combination of Background Survey responses, the resulting test would not be the same. The equivalence of the forms comes with the rating assigned to the elicited speech sample by reflecting the descriptors contained in the *ACTFL Proficiency Guidelines 2012 – Speaking.*

Scorer reliability for extended response items

Another common approach to examining reliability, in addition to Spearman's rho (ρ), is to use measures of inter-rater agreement. Whereas inter-rater reliability assesses how consistently the raters rank-order test-takers, inter-rater agreement assesses the extent to which raters give the same score for a particular test-taker. Since rating protocol assigns final test scores based on agreement (concordance) between raters rather than rank-order consistency, it is important to assess the degree of interchangeability in ratings for the same test taker. Inter-rater reliability can be high when inter-rater agreement is low, so it is important to take both into account when assessing a test.

Inter-rater agreement can be assessed by computing absolute agreement between rater pairs (i.e., whether both raters provide exactly the same rating). Standards for absolute agreement vary depending on the number of raters involved in the rating process. When two raters are utilized, there should be absolute agreement between raters more than 80% of the time, with a minimum of 70% for operational use (Feldt & Brennan, 1989). Absolute agreement closer to 100% is desired, but difficult to attain. Each additional rater employed in the process decreases the minimum acceptable agreement percentage.

This accounts for the fact that agreement between more than two raters is increasingly difficult. Adjacent agreement is also assessed in this reliability study. Adjacent agreement occurs when raters are within one rating level in terms of their agreement (e.g., rater one gives a test taker a rating of Intermediate Mid and rater two gives a rating of Intermediate Low). In the ACTFL process, when there is not absolute agreement, an arbitrating third rater will provide a rating that resolves the discrepancy.

Table 6 Absolute/Adjacent Agreement by Language

Language	N	Absolute Agreement (exact)	Adjacent Agreement (+/- 1)	None (+/- 2)
Arabic	686	75%	23%	2%
Spanish	4607	77%	19%	4%
English	8017	74%	25%	1%

Table 7 Absolute/Adjacent Agreement by Language and Year

Language	Year	N	Absolute Agreement (exact)	Adjacent Agreement (+/- 1)	None (+/- 2)
Arabic	2012	198	66%	31%	3%
	2013	229	75%	24%	1%
	2014	259	82%	17%	1%
Spanish	2012	1643	79%	17%	4%
	2013	1584	73%	22%	5%
	2014	1389	79%	18%	3%
English	2012	1936	73%	26%	1%
	2013	2485	76%	22%	2%
	2014	3596	73%	26%	1%

Table 8 Absolute/Adjacent Agreement by Language and Sublevel Proficiency from 2012-2014

Language	Rating	N	Absolute Agreement (exact)	Adjacent Agreement (+/- 1)	None (+/- 2)
Arabic	Novice Low	11	100%	0%	0%
	Novice Mid	69	86%	12%	2%
	Novice High	106	68%	30%	2%
	Intermediate Low	132	76%	22%	2%
	Intermediate Mid	93	70%	29%	1%
	Intermediate High	76	75%	24%	1%
	Advanced Low	65	75%	22%	3%
	Advanced Mid	35	65%	35%	0%
	Advanced High	39	72%	28%	0%
	Superior	60	87%	13%	0%
English	Novice Low	28	93%	7%	0%
	Novice Mid	83	93%	8%	0%
	Novice High	134	88%	12%	1%
	Intermediate Low	252	80%	20%	0%

	Intermediate Mid	1342	92%	8%	0%
	Intermediate High	903	79%	18%	3%
	Advanced Low	521	66%	26%	8%
	Advanced Mid	364	46%	40%	14%
	Advanced High	499	60%	33%	7%
	Superior	489	78%	18%	5%
Spanish	Novice Low	41	76%	24%	0%
	Novice Mid	125	78%	21%	1%
	Novice High	190	66%	34%	0%
	Intermediate Low	454	61%	38%	1%
	Intermediate Mid	1472	75%	24%	1%
	Intermediate High	2009	75%	24%	1%
	Advanced Low	1894	80%	19%	1%
	Advanced Mid	1012	71%	26%	2%
	Advanced High	569	64%	35%	1%
	Superior	250	73%	22%	4%

Absolute agreement was higher than 70% for all comparisons within a major level. Absolute agreement and adjacent agreement all summed to at least 95%. Absolute agreement was similar across interview language and language category. Absolute agreement deviated in the extreme scores and near the Novice High-Intermediate Low border more so than in other sublevels. Comparisons made by Language and Sublevel Proficiency should be viewed with caution as sample sizes can be limited and thus they should be used as a tool to help improve rater training.

Overall, the findings support the reliability of the ACTFL OPIc® as an assessment of speaking proficiency. Based on a small sample size, the areas for continuous improvement include increasing rater agreement at the Novice High-Intermediate Low border, particularly for Arabic and Spanish (68% and 66% absolute agreement at Novice High, respectively). Although review of the limited sample would indicate that the NH/IL border is an area for continued improvement in interrater reliability, this has less of an impact on ACE Credit recommendations as the number of credits recommended by ACE for the ratings of Novice High and Intermediate Low is the same. Current ACE credit recommendations for ACTFL OPIc® ratings are listed in the chart below:

Official ACTFL OPIc Rating	ACE Credit Recommendation
AH/S	6 (LD) + 8 UD)
AM	6 (LD) + 3 (UD)
IH/AL	6 (LD) + 1(UD)
IM	6 (LD)
NH/IL	3 (LD)

Errors of classification percentage for the minimum score for granting college credit (cut score)

The minimum score for granting college credit for an ACTFL OPIc® rating is Novice High. ACE determines the number of credits to be conferred based on the recommendations of expert reviewers, foreign language faculty who are familiar with language proficiency and the skills that students are expected to attain after various sequences of college language study.

Validity Information

Content-related validity

Content validity addresses the alignment between the test prompts and the content area they are intended to assess. There are two types of content-related validity, face validity and curricular validity. Face validity refers to the extent to which a test or the questions on a test *appear* to measure a particular construct. While curricular validity is the extent to which the content of the test matches the objectives of a specific curriculum. Both types of validity are evaluated by groups of content experts. Content validity evidence for the OPIc® (similarly to the OPI®) is represented by the degree to which the content of the test relates to the construct of speaking proficiency as defined by *the ACTFL Proficiency Guidelines 2012 – Speaking* (ACTFL 2012).

Criterion-related validity

Similar to content-related validity, criterion-related validity also has two types. One type of criterion-related validity is predictive validity which refers to the power or usefulness of test scores to predict future performance. Concurrent validity, the other type of criterion-validity, focuses on the power of the test to *predict* outcomes on another test with similar content-related validity.

The OPIc® is an integrative test addressing a number of abilities simultaneously and looking at them from a global perspective rather than from the point of view of the presence or absence of any given linguistic feature. Linguistic components are viewed from the wider perspective of their contribution to overall speaking performance. In evaluating a speech sample, the following criteria are considered:

- functions and global tasks the speaker is able to sustain
- accuracy or precision with which these tasks are accomplished and understood
- type of oral text or discourse the speaker is capable of producing.

The goal of the instrument is the same as the OPI: to obtain a ratable sample of speech which a rater can evaluate and compare to the 10 levels described in the *ACTFL Proficiency Guidelines 2012 – Speaking* in order to assign a rating.

Surface, Poncheri, and Bhavsar (2008) performed a study investigating the reliability and validity of the ACTFL OPI and OPIc® English Language exams on Korean test takers. The researchers concluded that both assessments measure the same construct, have similar reliabilities, and provide similar inferences. The findings from the two studies provide sufficient evidence to justify the initial use of the ACTFL OPIc® for commercial testing. However, ACTFL should maintain its commitment to using research to inform the test development and validation process as it extends the computerized interview format to other languages and test takers.

Construct validity (if appropriate)

Construct validity refers to the degree to which a test or other measure assess the underlying theoretical construct it is supposed to measure. Within construct validity there are two types: convergent validity and discriminant validity. Convergent validity consists of providing evidence that two tests are believed to measure closely related skills and addresses the reciprocity/correlation between measures that share the same content-related validity. Conversely, discriminant validity consists of evidence that two tests do not measure closely related skills.

Dandonoli and Henning (1990) reported on the results of research conducted by ACTFL on the construct validity of the *ACTFL Proficiency Guidelines* and the oral interview procedure which mainly focused on the speaking, writing, listening and reading sections of the French and English language examinations. The researchers found strong support for the use of the Guidelines as a foundation for the development of proficiency tests and for the reliability and validity of the OPI. Given the strong relationship between the OPI® and OPIc®, the findings from this study can likely be generalized to the OPIc®.

Tschirner and Bärenfänger (2012) performed a study to link the ACTFL OPI and OPIc® to the Common European Framework for Reference for Languages: Learning, Teaching, Assessment (Council of Europe, 2001; CEFR) by following the benchmarking protocol established by the Council of Europe (Figueras, et. al., 2009). The researchers concluded that all measures investigated indicated a strong correspondence between CEFR and the ACTFL ratings. While the study only involved German samples, Tschirner and Barenfanger purport that since the study used very experienced tester trainers and testers for The European Language Certificates (TELC), the results can be generalized to the TELC suite of languages including English, Spanish, French, Portuguese, Italian, Russian, Czech, and Turkish.

Possible test bias of the total test score

Bias exists when a test makes systematic errors in measure or prediction (Murphy & Davidshofer, 2005, p.317). An example of this would occur when a test yields higher or lower scores on average when it is administered to specific criterion groups such as people of a particular race or sex than when administered

to an average population sample. Negative bias is said to occur when the criterion group scores lower than average and positive bias when they score higher.

Bias is typically identified at the item level. Since this test's content is routed based on the ability and interests of the test taker, no two interviews are the same, thus a test of item bias would not be appropriate A bias analysis of total test score may be appropriate; however, demographic information is not tracked and thus it is not possible.

Evidence that time limits are appropriate and that the exam is not unduly speeded

The OPIc® contains 12-15 timed prompts that are aimed at two contiguous levels based on the Self-Assessment (Novice/Intermediate, Intermediate/Advanced and Advanced/Superior). The candidate has 30 seconds to respond to Novice-level prompts, 1 minute to respond to Intermediate-level prompts, 2 minutes to respond to Advanced-level prompts, and 2:30 minutes to respond to Superior-level prompts. The number and topical variety of prompts within a limited linguistic range of the test, as well as the length of the allowed response time, give test candidates many, repeated opportunities to show their language ability.

Provisions for standardizing administration of the examination

Before beginning the OPIc®, test takers receive a complete explanation of OPIc® test procedures and instructions including a sample test question. These instructions are delivered in the test taker's native language. Each test taker then completes a Background Survey and a Self-Assessment.

The Background Survey is a questionnaire which elicits information about the test taker's work, school, home, personal activities and interests. The test taker completes the survey and the answers determine the pool of topics from which the computer will randomly select questions. The variety of topics, the types of questions, and the range of possible combinations the computer can generate allow for individually designed interviews. Even if two test takers selected the same combination of Background Survey responses, the resulting test would not be the same.

The Self-Assessment provides six different descriptions of how well a person can speak a language. Test takers select the description that they feel most accurately describes their language ability. Samples of speech accompany each descriptor, so test takers can also listen to samples to help select the most appropriate description. The Self-Assessment choice determines which one of five OPIc® test forms (Form 1, Form 2, Form 3, Form 4, or Form 5) is generated for the specific individual. The choices made by the test taker in response to the Background Survey and the Self-Assessment assure that each test taker receives an adaptive and unique test.

The OPIc® provides detailed test instructions and directions on how to listen to the questions and record answers. In order to ensure that the test taker understands these instructions, a sample question is provided for the test taker to practice the functionality of the OPIc®. The test taker has the opportunity to re-review the instructions and sample question before beginning the test. The test taker then begins the OPIc® test.

Ava is an avatar figure that personifies the OPIc® tester and speaks the prompts in the language that is being assessed. Test takers listen to the avatar's questions and respond. Having the picture of Ava on the

screen helps to engage the test takers in conversation and mimics a one-on-one conversation with a native speaker of the target language.

The OPIc® structure is based on one of five test forms:

Form 1 - targets proficiency levels Novice Low through Novice High, though any rating from Novice Low through Intermediate Low can be assigned to a sample that is elicited using Form 1.

Form 2 - targets proficiency levels Novice High through Intermediate Mid, though any rating from Novice Low through Intermediate High can be assigned to a sample that is elicited using Form 2.

Form 3 - targets proficiency levels Intermediate Mid through Advanced (Low), though any rating from Novice Low through baseline Advanced can be assigned to a sample that is elicited using Form 3.

Form 4 - targets proficiency levels Advanced Low through Advanced Mid, though any rating from Intermediate High through Advanced High can be assigned to a sample that is elicited using Form 4.

Form 5 – targets proficiency levels Advanced High through Superior, though Advanced Mid can also be assigned to a sample that is elicited using Form 5.

The elicited sample is digitally recorded and archived in a secure data base.

Provisions for exam security

Official OPIc®s are administered in proctored environments. All proctors must read and review proctor instructions and sign an official proctor agreement before being given access to any logins for assessments.

When the OPIc® is administered to an academic institution, educational organization, or corporate clients, the following personnel qualify as potential proctor candidates:

K-12 Schools and School Districts

A proctor at a K-12 school or school district may only be a Principal, Assistant Principal, Dean, Administrative Assistant to the Principal or Dean, School District HR personnel, or Academic Chair. No other administrators or staff are permitted to act as proctors. All must submit a signed proctor agreement.

University or College

A proctor at a college may be a Professor, Department Chair, Department Administrative Assistant or Department Coordinator. No other administrators or staff are permitted to act as proctors. All must submit a signed proctor agreement.

Corporate clients

A proctor at a corporate site must be a managerial-level Human Resource staff member, or executive staff member, or, for branch offices without an on-site human resource representative, a senior-level manager may act as proctor. All must submit a signed proctor agreement.

Security Measures

Each test candidate is required to fill out a personal survey before the start of the OPIc®. Responses to the survey trigger the random selection of a set of test prompts (9-15 depending on the level) from a test prompt pool of over 3200 prompts. All official OPIc®'s are proctored to ensure that candidates do not record the prompts they receive. Logins for assessments are only valid for use for two weeks and once a candidate has logged into an assessment, they must complete that assessment in one sitting within an hour. If a test candidate tries to access another website while logged into the assessment, the OPIc® will close and only a proctor can log the candidate back in.

Raters also listen for suspicious behavior: the sound of someone helping the candidate, a change in the candidate's voice, etc. Raters are instructed to assign the score of UR for "unratable" and to notify LTI test administration of "suspicious behavior" which is then investigated by the Director of Test Administration.

Scaling and Item Response Theory Procedures

Types of IRT scaling model(s) used

Item Response Theory (IRT) models are not used in the calibration or scoring model for this exam. Test-takers are scored based on meeting criteria fitting the description of a major level which is representative of a specific range of abilities. Written descriptions of language abilities that a test taker must exhibit can be found in the *ACTFL Proficiency Guidelines 2012 - Speaking*.

Evidence of the fit of the model(s) used

The primary goal of the OPIc® is to produce a ratable sample of speech. To be ratable, a speech sample must clearly demonstrate the highest sustained level of performance of the speaker (known as the "floor") and the level at which the speaker can no longer sustain the performance (known as the "ceiling"), over a variety of topics. To this end, the tester follows a specific protocol, with four mandatory phases, in order to elicit a ratable sample.

Evidence that new items/tests fit the current scale used

The ACTFL Proficiency Guidelines and the Assessment Criteria for Speaking describe the range of content and contexts a speaker at each major level should be able to handle. For example, at the Intermediate level, topics of personal interest and related to one's immediate environment are selected; at the Advanced level, topics move beyond the autobiographical to topics of general community, national, and international interest; at the Superior level, topics are presented as issues to be discussed from abstract and/or hypothetical perspectives.

Validity of Computer Administration

Size of the operational test item pool for test

Each test candidate is required to fill out a Background Survey before the start of the OPIc®. Responses to the survey trigger the random selection of a set of test prompts (9-15 depending on the level) from a test prompt pool of over 3200 prompts. Prompts are rotated on a regular basis; new prompts are created and implemented while existing prompts are disabled.

Exposure rate of items when examinees can retake the test

The somewhat adaptive nature of the OPIc® allows for some level of exposure control as the questions are adapted to elicit ratable samples from the test taker. The variety of topics, the types of questions, and the range of possible combinations the computer can generate allow for individually designed interviews. Even if two test takers selected the same combination of Background Survey responses, the resulting test would not be the same. Records of retests are maintained to ensure that candidates receive alternative prompts, regardless of the number of re-tests an individual may take. Additionally, ACTFL controls for testing effects by limiting future retests to be 90 days from the most recent testing attempt.

Cut-score information

Rationale for the particular cut-score recommended

Once a ratable sample of speech has been elicited, that sample is evaluated according to the Assessment Criteria of the rating scale. A rating at any major level is determined by identifying the speaker's floor and ceiling. The floor represents the speaker's highest sustained performance across ALL of the criteria of the level all of the time in the Level Checks for that particular level; the ceiling is evidenced by linguistic breakdown when the speaker is attempting to address the tasks presented in the Probes. An appropriate sublevel can then be determined, and one of ten possible ratings is assigned by comparing the sample to the descriptions in the *ACTFL Proficiency Guidelines 2012 – Speaking* and identifying the rating that best matches the sample.

Evidence for the reasonableness and appropriateness of the cut-score recommended

The *ACTFL Proficiency Guidelines* are descriptions of what individuals can do with language in terms of speaking, writing, listening, and reading in real-world situations in a spontaneous and non-rehearsed context. For each skill, these guidelines identify five major levels of proficiency: Distinguished, Superior, Advanced, Intermediate, and Novice. The major levels Advanced, Intermediate, and Novice are subdivided into High, Mid, and Low sublevels. The levels of the *ACTFL Proficiency Guidelines* describe the continuum of proficiency from that of the highly articulate, well-educated language user to a level of little or no functional ability.

These Guidelines present the levels of proficiency as ranges, and describe what an individual can and cannot do with language at each level, regardless of where, when, or how the language was acquired. Together these levels form a hierarchy in which each level subsumes all lower levels. The Guidelines are not based on any particular theory, pedagogical method, or educational curriculum. They neither describe how an individual learns a language nor prescribe how an individual should learn a language, and they

should not be used for such purposes. They are an instrument for the evaluation of functional language ability.

The *ACTFL Proficiency Guidelines* were first published in 1986 as an adaptation for the academic community of the U.S. Government's Interagency Language Roundtable (ILR) Skill Level Descriptions. The third edition of the *ACTFL Proficiency Guidelines* includes the first revisions of Listening and Reading since their original publication in 1986, and a second revision of the ACTFL Speaking and Writing Guidelines, which were revised to reflect real-world assessment needs in 1999 and 2001 respectively. New for the 2012 edition are: the addition of the major level of Distinguished to the Speaking and Writing Guidelines; the division of the Advanced level into the three sublevels of High, Mid, and Low for the Listening and Reading Guidelines, and; the addition of a general level description at the Advanced, Intermediate, and Novice levels for all skills.

Another new feature of the 2012 Guidelines is their publication online, supported with glossed terminology and annotated, multimedia samples of performance at each level for Speaking and Writing, and examples of oral and written texts and tasks associated with each level for Reading and Listening.

The direct application of the *ACTFL Proficiency Guidelines* is for the evaluation of functional language ability. The Guidelines are intended to be used for global assessment in academic and workplace settings. However, the Guidelines do have instructional implications. The *ACTFL Proficiency Guidelines* underlie the development of the *ACTFL Performance Guidelines for K-12 Learners* (1998) and the *ACTFL Performance Descriptors for Language Learners (2012)* and are used in conjunction with the National Standards for Foreign Language Learning (1996, 1998, 2006, 2014) to describe how well students meet content standards. For the past 25 years, the *ACTFL Proficiency Guidelines* have had an increasingly profound impact on language teaching and learning in the United States.

Procedures recommended to users for establishing their own cut scores (e.g. granting college credit)

The summary of the Official ACTFL credit recommendations can be found on the Language Testing International (LTI) website, the ACTFL testing office. Depending on the rating level achieved, ACE recommends anywhere from three lower division baccalaureate/ associate degree category credits for the achievement of Novice High/Intermediate Low, up to six lower division baccalaureate /associate degree category credits and eight upper division baccalaureate / associate degree category credits for the achievement of Advanced High/Superior language proficiency.

Bibliography

ACTFL (2012). ACTFL Proficiency Guidelines 2012. Retrieved October 1, 2015 (http://www.actfl.org/publications/guidelines-andmanuals/actfl-proficiency-guidelines-2012)

Breiner-Sanders, K.E., Lowe, Jr., P., Miles, J., Swender, E. (2000). ACTFL proficiency guidelines – Speaking revised 1999. *Foreign Language Annals*. 33(1). 13-18.

Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). Council of Europe, Language Policy Unit, Strasbourg (2001) http://www.coe.int/t/dg4/linguistic/source/framework_en.pdf

Dandonoli, P., Henning, G. (1990). An investigation of the construct validity of the ACTFL proficiency guidelines and oral interview procedure. *Foreign Language Annals. 23(1). 11-19.*

Feldt, L. S., & Brennan, R. (1989). Reliability. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 105–146). New York: Macmillan.

Murphy, K.R., Davidshofer, C.O. (2005). *Psychological testing: Principles and Applications.* New Jersey, USA: Pearson Prentice Hall.

Surface, E.A., Dierdorff, E.C. (2003). Reliability and the ACTFL oral proficiency interview: Reporting indices of interrater consistency and agreement for 19 languages. *Foreign Language Annals. 36(4). 507-519.*

Surface, E.A., Poncheri, R.M., Bhasvsar, K.S. (2008). Two studies investigating the reliability and validity of the English ACTFL OPIc® with Korean Test Takers. *Technical Report.* Available online at: http://www.languagetesting.com/wp-content/uploads/2013/08/ACTFL-OPIc®-English-Validation-2008.pdf

SWA Consulting. (2009). 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. Available online at: http://www.languagetesting.com/wp-content/uploads/2013/12/ACTFL-OPIc®-Retest-Reliability-Study-2009.pdf

SWA Consulting. (2012). Reliability study of the ACTFL OPI ® in Chinese, Portuguese, Russian, Spanish, German, and English for the ACE review. *Technical Report.* Available online at: http://www.languagetesting.com/wp-content/uploads/2013/08/ACTFL-OPI-Reliability-2012.pdf

Thompson, I. (1996). A study of interrater reliability of the ACTFL oral proficiency interview in five European languages: Data from ESL, French, German, Russian, and Spanish. *Foreign Language Annals. 28(3). 407-422.*

Tschiner, E., Bärenfänger, O. (2012). Assessing evidence of validity of assigning CEFR ratings to the ACTFL oral proficiency interview (OPI) and the oral proficiency interview by computer (OPIc®). *Technical Report*. Available online at: http://www.languagetesting.com/wp-content/uploads/2014/02/OPIc®-CEFR-Study-Final-Report.pdf