



# **ACTFL ORAL PROFICIENCY INTERVIEW - COMPUTER**

Part A: General Test Information

Part B: Statistical Analysis &  
Evidence of Validity



# ACTFL ORAL PROFICIENCY INTERVIEW - COMPUTER

Part A: General Test  
Information

**Stephanie Dhonau, Ph.D.**  
Department of World Languages  
University of Arkansas

## Table of Contents

Rationale and purpose of the OPIC .....	1
Proficiency rating and score reporting .....	2
Directions for scoring procedures and keys .....	4
Cut scores .....	4
Procedures recommended to users for establishing their own cut scores .....	4
Equivalence of forms .....	5
Information on norms and normative groups (if appropriate) .....	5
Item/Test content development .....	6
<i>Specifications that define the domain(s) of content, skills, and abilities that the test samples</i> .....	6
<i>Statement of test's emphasis on each of the content, skills, and ability areas</i> .....	7
<i>Rationale for the kinds of tasks (items) that make up the test</i> .....	7
Information about the adequacy of the items on the test as a sample from the domain(s) .....	7
Information on the currency and representativeness of the test's items .....	8
Description of the item sensitivity panel review .....	8
Whether and/or how the items pre-tested (field tested) before inclusion in the final form .....	8
Name(s) and institutional affiliations of the principle author(s) or consultant(s) .....	9
Item analysis results (e.g. item difficulty, discrimination, item fit statistics, correlation with external criteria) .....	10
References .....	11

## Table of Figures

Figure 1: OPIC assessment criteria chart .....	3
Figure 2: Time as a critical component for developing language performance .....	5

## Rationale and purpose of the OPlc

The American Council on the Teaching of Foreign Languages (ACTFL) Oral Proficiency Interview - Computer (OPlc®) is a semi-direct test of functional spoken language ability which is delivered online. The test uses an interview format to elicit speech from a test taker on a variety of topics over the course of 20 to 40 minutes in order to capture a sample of *unrehearsed* speech as evidence of what an individual *can do* using the spoken language. In line with its proficiency-oriented framework, it is not an assessment of what the test taker knows about a language (e.g. its rules and structures), but rather an assessment of real-life oral skills. Ava, an avatar, serves as the virtual interviewer posing sets of questions organized by theme or topic.

The ACTFL OPlc® is appropriate for both an individual test taker who desires an official proficiency rating and also for large-scale testing in a variety of educational, commercial, and governmental settings. Since it is delivered online, the OPlc® can accommodate thousands of individual test takers at any given administration. Because of this accessibility, proctors can schedule and administer the OPlc® to test candidates easily, anywhere in the world. After the test has been recorded, the sample is made available via a secure “Rater Site” to Certified OPlc® Raters, facilitating prompt evaluation and reporting of the official certified rating.

Each test taker completes both a Background Survey indicating personal interests and experiences as well as a Self-Assessment of perceived language proficiency. The Background Survey allows the test taker to indicate experience with a wide variety of topics. The results of the Background Survey determine the set of questions or prompts that are explored during the test. In the Self-Assessment, the test taker is asked to indicate what their perceived proficiency level may be based on descriptions of the major levels of Novice, Intermediate, Advanced and Superior. Based on these two surveys, a unique and individualized test tailored to the most likely range of linguistic ability of the test taker is generated. Included topics relate to the test taker’s work experience, academic background, and interests. Prompts are selected from an item bank of pre-recorded prompts that are organized into testlets by topic and proficiency level. Each OPlc® explores 4-5 topics, depending on the form. Prior to the administration of the test, the test taker receives a full overview and explanation of OPlc® procedures, including a sample test question. These instructions are delivered in the test taker’s first language (L1) to ensure instructions are clear.

The ACTFL OPlc® is designed to garner a spontaneous, *unrehearsed* sample of speech that is uploaded to a rating website from which a certified OPlc® rater can access the speech sample and evaluate it according to the criteria and descriptors provided in the *ACTFL Oral Proficiency Guidelines - Speaking 2012*. As the ACTFL OPlc® is a criterion-referenced assessment of an individual's ability to communicate in a target language, the rater compares the test taker’s language in each task to the *Guidelines*. By listening holistically to the entire speech sample, a rater evaluates the evidence and assigns a final rating by referring to the descriptors found in the *Guidelines*. Currently, possible ratings that may be assigned are Novice Low, Novice Mid,

Novice High, Intermediate Low, Intermediate High, Advanced Low, Advanced Mid, Advanced High, and Superior. Test forms and rating parameters are discussed below.

The *ACTFL Proficiency Guidelines* (2012) are a set of descriptors of functional language ability. Each description represents a range of ability, and each level subsumes all levels below it. For example, an Advanced-level speaker can perform all the functions associated with the Novice and Intermediate levels as well as functions at the Advanced level. Advanced, Intermediate, and Novice levels are divided into sublevels (Low, Mid and High) that indicate the range of ability in terms of both quality and quantity of the language produced at the major level. The Superior level is not divided into sublevels.

For purposes of understanding how proficiency ratings may be linked to real-life usage and application, ACTFL (2015) has published a recommended minimal level of proficiency for several common professions based on job-related language use.<sup>1</sup> These minimal levels of proficiency often depend on local, state, or even federal requirements. For purposes of illustration, a few examples of recommended proficiency levels required to perform successfully in various workplace positions include, but are not limited to, a receptionist or a cashier at IM; a tour guide or firefighter at IH; a K-12 language teacher or police officer at AL, a human resource benefits specialist at AM; a physician or financial advisor at AH; and a court interpreter or university language professor at S. Novice level proficiency, while important to acknowledge, is not included in these recommendations, as speakers are not able to demonstrate enough functional ability for the workplace at this level. Note that these are only *suggested* proficiency levels for the workplace.

### **Proficiency rating and score reporting**

An examinee receives a score that confirms the major level of proficiency sustained across the test, as well as a sublevel indicating the quality and quantity of language produced when performing the functions of that major level. A High sublevel score indicates that the test taker can consistently produce language at the major level with excellent quantity and quality, and also demonstrates substantial ability to perform the functions of the next major level most of the time. Conversely, a Low sublevel score notes the test taker's *minimal* ability to sustain the required functions at the major level. A Mid sublevel signifies that the test taker can comfortably provide not only quantity of language but also solid quality performance of the major level functions consistently throughout the test.

To reiterate, a rating at any major level confirms the sustained performance across ALL of 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 *ACTFL Proficiency Guidelines* describe the

---

<sup>1</sup> ACTFL (2015). *Oral Proficiency in the Workplace Poster*.

[https://www.actfl.org/sites/default/files/pdfs/TLE\\_pdf/OralProficiencyWorkplacePoster.pdf](https://www.actfl.org/sites/default/files/pdfs/TLE_pdf/OralProficiencyWorkplacePoster.pdf)

tasks that speakers can handle at each major level, as well as the content, context, accuracy, and discourse types associated with the ability to perform those tasks. More specifically, content and context refers to the variety of topics and situations found in the testlets of the exam; accuracy refers not only to grammatical accuracy but also to features that affect the test taker's comprehensibility to the listener, such as pronunciation, tones, for example; and finally discourse types refer to what type of language the speaker produces such as words or phrases at Novice, sentences at Intermediate, paragraphs at Advanced, and extended discourse at Superior. Further descriptions of each level are available online:

<https://www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012> .

As per the ACTFL Guidelines, the general assessment criteria used to evaluate the language performance on an ACTFL OPIc® is provided in the chart below:

Proficiency Level	Global Tasks and Functions	Context/Content	Text Type	Accuracy
<b>Superior</b>	Discuss familiar and unfamiliar topics Supports opinions, hypothesize, and deal with topics abstractly	Most informal and formal settings  <i>Wide range of public interest topics and some special fields of interest and expertise</i>	Extended discourse	No pattern of error in basic structures Errors virtually never interfere with communication or distract from the message
<b>Advanced</b>	Narrate and describe in major time frames and deal effectively with an unanticipated complication.	Most informal and some formal settings  <i>Topics of personal and general interest</i>	Oral Paragraphs, Connected Discourse	Understood without difficulty by speakers unaccustomed to dealing with language learners (non-sympathetic listener)
<b>Intermediate</b>	Create with language, initiate, maintain, and bring to a close, simple conversations by asking and responding to simple questions	<i>Some informal settings and a limited number of transactional situations</i>  <i>Predictable familiar topics related to daily activities and personal environment</i>	Sentences	Understood with some repetition by speakers accustomed to dealing with language learners (sympathetic listener)
<b>Novice</b>	Communicate minimally and formulaic and rote utterances, lists, and phrases	<i>Most common informal settings</i>  <i>Most common aspects of daily life</i>	Individual words, phrases, and lists	May be difficult to understand even for speakers accustomed to dealing with language learners

*Figure 1: OPIc assessment criteria chart*

For more detailed sublevel information, please refer to the [ACTFL Proficiency Guidelines - Speaking 2012](#).

## Directions for scoring procedures and keys

Once the OPlc® is completed and submitted, the speech sample is uploaded and saved automatically on a secure Internet site. An ACTFL Certified OPlc® Rater listens to the sample holistically 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 the descriptors and enters the rating into the online rating system. The OPlc® is then blindly second rated by another certified OPlc® rater, following the same protocol. If the two ratings agree exactly, the rating is finalized; if the two ratings differ, the OPlc® is assigned to a third rater for a blind arbitration. This protocol helps to maintain interrater reliability that is monitored closely by ACTFL Quality Assurance.

ACTFL Certified OPlc® Raters are highly specialized language professionals who have completed a rigorous preparation process that concludes with a rater's demonstrated ability to consistently rate samples with a high degree of reliability. Raters are required to participate in regular calibration activities to maintain their individual rating reliability and inter-rater reliability.

## Cut scores

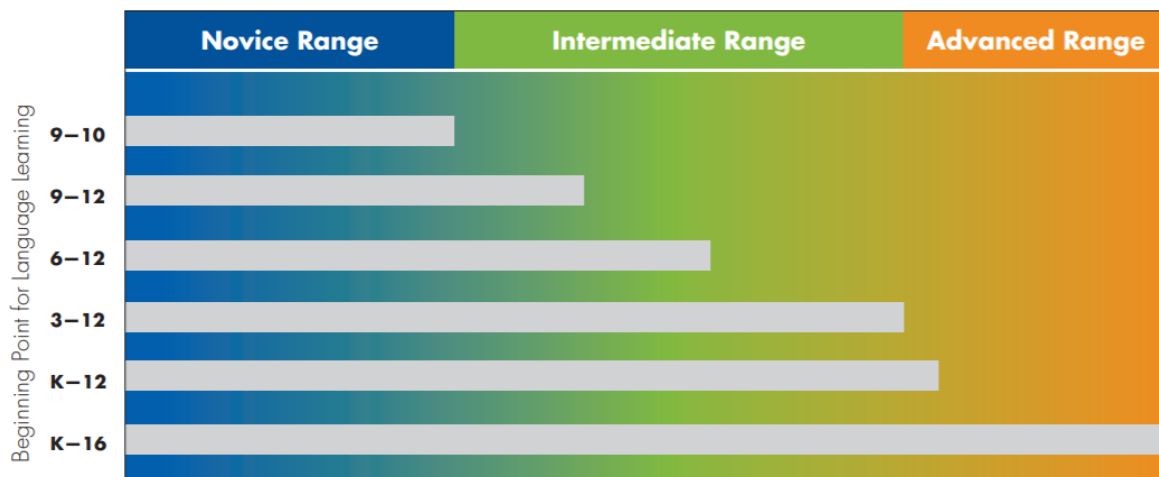
The OPlc® does not have numeric cut scores. The OPlc is an assessment of language proficiency that is rated holistically according to the *ACTFL Proficiency Guidelines (2012)*.

## Procedures recommended to users for establishing their own cut scores

As previously referenced, the ACTFL OPlc is a proficiency-oriented assessment with no recommended cut scores. That is, the OPlc should result in a description of the test taker's spontaneous, unrehearsed language abilities. As such, the 2015 – 2019 ACE credit recommendations relate proficiency levels to credit recommendations.

ACTFL RATING	OPI/OPlc
Novice High/Intermediate Low	3LD
Intermediate Mid	6LD
Intermediate High/Advanced Low	9LD
Advanced Mid	6LD + 3UD
Advanced High/Superior	6LD + 6UD

For any language program, the proficiency levels can be mapped to course and program goals by analyzing the descriptors and comparing them to course and/or program objectives in addition to factors such as time.



*Figure 2: Time as a critical component for developing language performance*

ACTFL suggests that the credit recommendations and proficiency targets above are in line with the number of courses and years of study that an undergraduate student of typical aptitude might achieve (see Figure 2).

### Equivalence of forms

The OPIc is made up of a 2,000-item pool; as referenced above, items are selected from the pool based on an algorithm which builds a test form based on the test taker's responses to the background survey and the self-assessment. Each examinee should receive a unique set of items in many instances.

The OPIc is based on the Oral Proficiency Interview (OPI), meaning that test prompts are function-based as outlined in the *ACTFL Proficiency Guidelines*. This allows for a standardized approach to test development such that the content of a prompt along with tasks used to convey the functions differ from item to item and examinee to examinee; however, the functions for which test takers must demonstrate a sustained ability to communicate remain consistent. Prompt writer's adherence to the function-informed and rating scale-normed item writing protocol along with adherence to the process of awarding ratings according to the ACTFL Proficiency Descriptors allow for equivalence between forms.

### Information on norms and normative groups (if appropriate)

The OPIc® is a criterion-referenced test. No norm-referenced information is reported.



## Item/Test content development

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

As mentioned above, the ACTFL OPIc® utilizes a Background Survey to elicit information about the test taker's work, school, home, personal activities and interests to ensure that the test taker has the best opportunity to use their language proficiency in topics of interest and relevance. The results of the survey determine the pool of topics from which the computer randomly selects questions in order to build out the testlets. The test taker also completes a linguistic Self-Assessment, comparing their perceived ability to can-do statements. These statements map to ranges in the *ACTFL Proficiency Guidelines – Speaking 2012* (though the terms Novice, Intermediate, Advanced, Superior are not used). These statements are accompanied by samples of the language that are generally representative of each proficiency range. The test taker's identification of their approximate level determines which test form is delivered.

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 major levels and a variety of topics (simulating the iterative process of the ACTFL OPI). The topics are organized into testlets, a set of 2-3 questions related to the same topic that may all be designed to elicit language at one major level or may spiral a given topic from one major level to the next. This design is intended to match what an OPI tester does when developing a topic in a live, real-time conversational format. The range of possible combinations the computer can generate allows for individually designed assessments. Even if two test takers select 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.

It is important to note that based on the Self-Assessment and the Background Survey, there are five different forms that may be generated for the test taker, and each form has specific rating parameters. Form 1 targets Novice level proficiency from NL-NH and may be rated from NL to IL; Form 2 has a targeted range from NH-IM and may be rated from NL to IH, and Form 3 targets IM-AL and may be rated from NL to AL. Forms 4 and 5 target the higher levels of proficiency. Form 4 targets IH-AM with a highest possible rating of AH and is considered not ratable if the candidate falls below IH; Form 5 targets AH-S with a highest possible rating of S and is considered not ratable if the candidate falls below AL.

In the event that a test taker substantially underassesses or over assesses their proficiency, these rating limits may result in a rating that does not reflect the test taker's genuine proficiency level. When a rater believes this to be the case, a standard notation is made at the time of evaluation and can be conveyed to the examinee or client. It is relevant to note that in some instances, the form given to the test taker is determined by a requesting client and may not be generated from a Self-Assessment survey.

### *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 can spontaneously speak in response to carefully constructed prompts dealing with practical, social, and professional topics that are encountered in true-to-life informal and formal contexts and situations. These language tasks range from creating meaning with language to engage in simple conversations, asking questions, telling stories, providing detailed descriptions, narrating and describing in major time frames in cohesive paragraph-length discourse, dealing abstractly with current issues of general interest, to supporting one's opinion and hypothesizing using extended discourse. Depending on the form, each OPIc® elicits these functions in four to six topic areas so as to provide a sample that exhibits how well a test taker can communicate about a variety of topic areas. Throughout the test, the test taker must demonstrate the consistent ability to maintain these tasks across a variety of topics and contexts so that the holistic rating assigned certifies the speaker has demonstrated general proficiency and not proficiency in one or two narrow contexts related to specific personal or professional interests.

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

The rationale for the types of tasks required on the test are based upon the functional requirements of communicative tasks associated with the major proficiency levels. Within individual testlets consisting of two or three prompts, the speaker responds to at least one question regarded as a level check (a prompt that elicits functions at the “floor” or major level sustained by the speaker), followed by either additional level checks or a probe (a prompt that targets a function at the next major level or the “ceiling” where the test taker cannot sustain performance). For example, within a testlet, the test taker may be prompted to respond to an Intermediate prompt targeting an Intermediate function such as asking questions in a role play situation, followed by two Advanced level prompts targeting Advanced level functions such as solving a problem related to the role play situation and then telling a personal narration of a similar situation that occurred in the past. In this way, over the course of the test, test takers are able to provide evidence of sustained functional performance of one major level, and breakdown from the next major level. Over the course of a number of testlets on a variety of topics, the sample that is produced provides sufficient evidence of a speaker's patterns of linguistic strengths (their “floor performance”) and weaknesses (their “ceiling”). When rating the sample, the rater evaluates those patterns of strength and weakness and, in conjunction with the proficiency level descriptions of the *ACTFL Proficiency Guidelines*, decides on the rating that most closely matches the language produced throughout the test.

### **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 content and contexts a speaker at each major level should be able to handle. The OPIc® test pool covers

numerous personal activities, work related situations, and a wide variety of other topics and interests. Based on the incorporation of responses on both the Background Survey and Self-Assessment, the items (testlets) generated by the test algorithm provide adequate test item types to produce a robust sample of test taker abilities. Even if two test takers choose the same combination of Background Survey and Self-Assessment responses, the final form generated for each will be different.

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

Given the individualized nature of each test and the ample pool of possible topics that form testlets, along with the range and diversity of topics, subtopics, genres and functional rhetorical structures the prompts target, each test includes a variety of test items that gives the test taker the opportunity to provide sufficient evidence of target language use over the course of the 20 to 40 minute recorded sample.

Topic currency is directly tied to common domains taken from everyday modern life and may include issues related to school, home/housing, work, free-time activities, technology use, travel, social activities, sports, etc. New topics are developed frequently, and less current or no longer relevant prompts are retired from the item bank.

### **Description of the item sensitivity panel review**

Just as in the case of the ACTFL OPI, sensitive and/or controversial topics are avoided in the design of the test items for the OPIc®. The Background Survey helps to limit the inclusion of unknown or irrelevant topics being generated on an individualized test. During the writing and revising of test prompts, item writers are instructed to avoid sensitive and/or controversial topics (e.g. nationality, ethnicity, religion, sexual preference, immigration, racism, gun control, etc.). With test takers choosing topics based on personal experiences, the resulting individualized test should prevent the inclusion of items that might lead to feelings of uncertainty or unease.

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

Because each OPIc® is generated based on the test taker's responses to the Background Survey and Self-Assessment, there is no standard OPIc® "final form." However, items are pre-tested before they are added to the item pool. After pretesting items, those that do not elicit the targeted functional criteria or fail to produce expected responses for other reasons are revised or removed from the item pool. Quality assurance reviews of test items take place periodically to evaluate whether certain test items are no longer effective.

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

The principal authors for the original OPlc include:

- Kathy Akiyama, Ph.D., Mt. Angel Seminary
- Mahdi Alish, 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

Subsequent item refreshes have taken place and include item writers such as:

- Mika Hoffman, Ph.D., Excelsior College, NY
- Reuben Vyn, Ph. D., University of Iowa
- Cynthia Martin, Ph.D., University of Maryland
- Quyen Ngo (Vietnamese), Independent Consultant
- Kim Le (Vietnamese), Independent Consultant

On-going prompt development includes item writers such as:

- Mark Darhower (English), Ph.D., North Carolina State University
- Stephanie Dhonau (English), Ph.D., University of Arkansas
- Kathy Akiyama (English), Mount Angel Seminary
- Jennifer Swender (English), Independent Consultant
- Martina Lindseth (German/English), Ph.D., University of Wisconsin-Eau Claire
- Sahie Kang (Korean), Ph.D., Asian School
- Bill Prince (Spanish/English), Ph.D., Furman University
- Mei Kong (Chinese), Ph.D., University of Maryland
- Nawal Moussa (Arabic), Canadian Defense Academy
- Mildred Rivera-Martinez (Spanish/English), Ph.D., Emeritus, Peace Corp
- Mindy Lindgren (English/Spanish), Ashland Middle School

**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*. Please refer to Alpine Testing Solutions (2020b) for a statistical analysis of the ACTFL Oral Proficiency Interview - Computer.

## References

ACTFL (2012). *ACTFL proficiency guidelines– speaking*.

<https://www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012/english/speaking>

Alpine Testing Solutions. (2020b). [\*Examination of the ACTFL Oral Proficiency Interview - Computer® \(OPIc\) in Arabic, English, and Spanish for the ACE Review - Part B: Statistical analysis & evidence of validity\*](#). Orem, UT: Alpine Testing Solutions.

Cubbellotti, S. (2015b). [\*Examination evaluation of the ACTFL OPIc® in Arabic, English, and Spanish for the ACE Review\*](#) (ACTFL Publication No. AAR/OPIc/ACE/R—2015-001). Alexandria, VA: ACTFL.

ACTFL (2015). *Oral proficiency in the workplace poster*.

[https://www.pinterest.com/pin/698409854701284664/?nic\\_v2=1a2rdZ0z4](https://www.pinterest.com/pin/698409854701284664/?nic_v2=1a2rdZ0z4)  
[https://my.actfl.org/portal/ItemDetail?iProductCode=POSTER\\_WORKPL](https://my.actfl.org/portal/ItemDetail?iProductCode=POSTER_WORKPL)

# **Examination Evaluation of the ACTFL Oral Proficiency Interview - Computer® (OPIc) in Arabic, English, and Spanish for the ACE Review**

## **Part B: Statistical Analysis & Evidence of Validity**

Submitted to the American Council on the Teaching of Foreign Languages  
(ACTFL)

Submitted May 29, 2020



# Table of Contents

Executive Summary.....	3
Statistical Performance.....	4
Item Analysis Results (e.g., Item Difficulty, Discrimination, Correlation with External Criteria) .....	4
Reliability Information, Scorer Reliability for Essay Items, Errors of Classification When Single or Multiple Cut Scores are Used.....	4
Score Stability Over Time.....	10
Evidence of Validity.....	12
Content Related .....	12
Criterion Related .....	12
Construct Related .....	13
Possible Test Bias .....	13
Evidence that Time Limits are Appropriate and that the Exam is not Unduly Speeded ..	13
Provisions for Standardizing Administration of the Examination.....	13
Irrelevant Sources of Difficulty Affecting Test Scores.....	14
Provisions for Exam Security.....	14
Interpretations and Conclusions.....	15
References .....	17





## Executive Summary

This document is structured to parallel the ACE Examination Checklist, which addresses the following topics: general test information, item/test content development, statistical performance, and validity evidence.

This report documents the American Council on the Teaching of Foreign Languages (ACTFL) Oral Proficiency Interview – Computer (OPIc®) from 2016 to 2020 to satisfy a review requirement of the American Council of Education (ACE) College Credit Recommendation Service (CREDIT) program. The ACTFL OPIc® is an assessment of functional speaking proficiency in a foreign language that is evaluated by trained and certified experts in a computerized interview format across numerous languages.

Inter-rater reliability and rater agreement were analyzed for three languages of the ACTFL OPIc: Arabic, English, and Spanish. Additionally, comparisons were analyzed across language proficiency levels, as well as for testing years (i.e. 2016-2020, in this sample).

Results show that the ACTFL OPIc surpassed the minimum inter-rater reliability and agreement requirements. Scores were in agreement within one sublevel of each other over 92% of the time, and within two sublevels 99% of the time. Additionally, the findings of the Spearman's *R* Correlation analyses demonstrate that the correlations of the ratings are almost always positive and strong, ranging from 0.74 - 0.93 across languages. Areas for improvement include a focus to the absolute agreement between raters within the Advanced Mid and Superior borders for English and Spanish, and within the Intermediate levels for English and Arabic. These findings are expanded upon and discussed in detail below.

Please refer to Part A for general test information.

## Statistical Performance

### Item Analysis Results (e.g., Item Difficulty, Discrimination, Correlation with External Criteria)

Examinees are scored at the level that “represents a range in which speakers demonstrate sustained functional ability of the linguistic functions associated with that level,” which means a single holistic score is assigned for the whole exam (see *ACTFL OPIc Examinee Handbook*, page 15). Individual item (prompt) data is not collected.

### Reliability Information, Scorer Reliability for Essay Items, Errors of Classification When Single or Multiple Cut Scores are Used

An inter-rater agreement analysis was conducted for each language from 2016 to 2020. In this analysis, the number of times Rating 1 and Rating 2 agreed exactly, within one category (proficiency level), within two categories, or beyond two categories was counted. When two ratings did not agree, a third rating contributed to the score. If there was still disagreement, a fourth rating contributed to the decision. It is noteworthy that Ratings 1, 2, 3, and 4 does not mean a specific *Rater 1, 2, 3, and 4*. Instead, Rating 1 refers to the rating assigned by “Rater 1”, where Rater 1 was selected from a pool of trained raters. An individual assigned as “Rater 1” for one candidate, may be Rater 2, 3, or 4 for another candidate. In other words, the rating number is not consistently connected to a specific individual.

The exam is initially scored by two raters (i.e., Rating 1 and Rating 2). If these two raters do not agree, a third rater is brought in for rater arbitration. If the third rater agrees with either of the first two raters, then the rating is finalized. However, if the third rater disagrees with both of the first two raters, a fourth rater is brought in. This process is followed for nearly all scores; however, there are cases in which scores are finalized after conversations with the involved raters.

Table 1 lists the number of examinees analyzed by year. Table 2 lists the percent of examinees that had exactly two, exactly three, or four ratings for their exam. Overall, the percentage of the number of ratings was fairly consistent across the three languages.

**Table 1. Number of Examinees by Year**

	2016	2017	2018	2019	2020	Total
Arabic	163	391	235	269	68	1126
English	1309	977	871	1226	200	4583
Spanish	3797	5045	5172	4305	1081	19400

\*Arabic data collected through March 11, 2020; English and Spanish data collected through March 30, 2020.

**Table 2. Percent of Examinees with 2, 3, or 4 Ratings from 2016 to 2020**

	N	2 Ratings	3 Ratings	4 Ratings
Arabic	1126	54%	46%	<1%
English	4583	54%	45%	1%
Spanish	19400	57%	43%	<1%

Tables 3-5 list the agreement of Rating 1 and Rating 2 by category. Table 6 summarizes the percent of exact agreement, adjacent agreement (within one category), and agreement within two categories.

**Table 3. Arabic OPIC: Rating 1 and Rating 2 Agreement from 2016-2020 (N = 1126)**

		Rating 1*									
		NL	NM	NH	IL	IM	IH	AL	AM	AH	S
Rating 2*	NL	1	2	0	0	0	0	0	0	0	0
	NM	1	18	15	2	0	0	0	0	0	0
	NH	1	14	63	52	13	2	0	0	0	0
	IL	0	0	21	67	72	18	2	0	0	0
	IM	0	0	0	21	81	61	25	1	0	0
	IH	0	0	0	2	32	58	69	9	2	0
	AL	0	0	0	0	2	14	72	28	5	0
	AM	0	0	0	0	0	1	7	26	12	1
	AH	0	0	0	0	0	1	1	13	74	12
	S	0	0	0	0	0	0	0	0	36	96

\*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 4. English OPIC: Rating 1 and Rating 2 Agreement from 2016-2020 (N = 4583)**

		Rating 1*									
		NL	NM	NH	IL	IM	IH	AL	AM	AH	S
Rating 2*	NL	1	1	0	0	0	0	0	0	0	0
	NM	1	7	1	0	0	0	0	0	0	0
	NH	0	5	10	8	4	2	0	0	0	0
	IL	0	0	9	16	24	1	1	0	0	0
	IM	0	0	2	30	173	84	10	0	0	0
	IH	0	0	1	7	104	185	84	30	4	0
	AL	0	0	0	0	13	106	228	183	46	4
	AM	0	0	0	0	0	39	228	554	292	23
	AH	0	0	0	1	1	3	55	326	651	154
	S	0	0	0	0	0	1	6	38	180	646

\*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 5. Spanish OPIC: Rating 1 and Rating 2 Agreement from 2016-2020 (N = 19,332)\***

		Rating 1**									
		NL	NM	NH	IL	IM	IH	AL	AM	AH	S
Rating 2**	NL	140	34	7	1	0	0	0	0	0	0
	NM	44	184	122	18	3	0	0	0	0	0
	NH	8	124	309	152	44	1	0	0	0	0
	IL	1	16	190	485	364	28	1	1	0	0
	IM	1	1	53	386	1443	610	41	2	0	0
	IH	0	0	2	35	501	2386	687	34	0	0
	AL	0	0	0	3	29	803	2429	813	71	11
	AM	0	0	0	0	3	44	954	2329	651	54
	AH	0	0	0	0	0	4	122	927	764	189
	S	0	0	0	0	0	0	6	99	283	285

\*68 examinees received a Rating 2 of "A" and could not be classified in this chart.

\*\*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

As shown in Table 6, Rating 1 and Rating 2 had exact agreement 49% of the time for the Arabic exam, 54% for the English exam, and 56% for the Spanish exam. All three were within one category of each other over 92% of the time. Tables 7-9 expand on these values by listing the percentage (and number) of exact agreements, adjacent agreements (within one category) and agreements within two categories, respectively.

**Table 6. Agreement between Rating 1 and Rating 2**

	N	Exact Agreement	Adjacent Agreement (within 1 category)	Agreement within 2 Categories
<b>Arabic</b>	1126	49.4%	92.2%	99.3%
<b>English</b>	4583	53.9%	93.6%	99.5%
<b>Spanish</b>	4037	55.6%	96.2%	99.8%

**Table 7. Percent (N) of Exact Agreement**

Language	Rating	Rating		
		2	3	4
<b>Arabic</b>	<b>1</b>	49.4% (556)	36.1% (189)	100.0% (1)
	<b>2</b>	---	45.7% (239)	0.0% (0)
	<b>3</b>	---	---	0.0% (0)
<b>English</b>	<b>1</b>	53.9% (2471)	38.5% (807)	43.5% (10)
	<b>2</b>	---	42.0% (880)	13.0% (3)
	<b>3</b>	---	---	30.4% (7)
<b>Spanish</b>	<b>1</b>	55.6% (10754)	48.6% (4077)	20.0% (2)
	<b>2</b>	---	40.5% (3365)	30.0% (3)
	<b>3</b>	---	---	60.0% (6)

**Table 8. Percent (N) of Adjacent Agreement (within 1 Category)**

Language	Rating	Rating		
		2	3	4
Arabic	1	92.2% (1038)	86.2% (451)	100.0% (1)
	2	---	93.7% (490)	100.0% (1)
	3	---	---	100.0% (1)
English	1	93.6% (4291)	88.1% (1847)	78.3% (18)
	2	---	89.1% (1867)	82.6% (19)
	3	---	---	78.3% (18)
Spanish	1	96.2% (18588)	95.4% (8001)	80.0% (8)
	2	---	93.4% (7767)	90.0% (9)
	3	---	---	90.0% (9)

**Table 9. Percent (N) of Agreement within 2 Categories**

Language	Rating	Rating		
		2	3	4
Arabic	1	99.3% (1118)	98.9% (517)	100.0% (1)
	2	---	99.6% (521)	100.0% (1)
	3	---	---	100.0% (1)
English	1	99.5% (4559)	99.1% (2077)	100.0% (23)
	2	---	98.5% (2064)	95.7% (22)
	3	---	---	87.0% (20)
Spanish	1	99.8% (19291)	99.8% (8370)	100.0% (10)
	2	---	99.7% (8293)	100.0% (10)
	3	---	---	100.0% (10)

The Spearman rank-order correlation ( $\rho$ ) was computed between each pair of Ratings. This correlation is a non-parametric measure of the strength and direction associated with the two variables of interest, in this case, two independent Ratings. The range of possible values is – 1.00 to +1.00. This correlation is computed by first ranking the items for one variable (in this case, one of the Ratings) and then correlating it to the ranking of the items for the other variable (in this case, another Rating). A statistical significance test of the correlation determines whether the correlation is statistically significant.

The Spearman rank-order correlation is similar to a Pearson correlation, except the Pearson correlation involves interval level data while the Spearman rank-order correlation involves ordinal level data. Similar to the Pearson correlation, positive values would indicate a positive correlation between the two Ratings and negative values would indicate an inverse relationship between the two Ratings. For this dataset, a positive correlation is expected (i.e., as the rating increases for one Rating, it is expected that the rating would also increase for the other Rating).

The strength of the correlation is determined by the magnitude of the correlation. Correlations with absolute values of at least 0.70 generally indicate a strong correlation.

Table 10 displays the Spearman rank-order correlation results for each pair of Ratings. Ratings involving Rating 4 are not shown due to the small sample sizes. All correlations were strong, positive, and statistically significant.

Table 11 breaks down the correlations by year. All correlations were strong, positive, and statistically significant. Again, ratings involving Rating 4 are not shown due to the small sample sizes.

**Table 10. Spearman Rank-Order Correlations by Language from 2016-2020**

<b>Ratings Compared</b>	<b>Language</b>	<b>N</b>	<b><math>\rho</math></b>	<b><i>p</i>-value</b>
1 and 2	Arabic	1126	0.927	< 0.001
1 and 2	English	4583	0.833	< 0.001
1 and 2	Spanish	19332	0.907	< 0.001
1 and 3	Arabic	523	0.862	< 0.001
1 and 3	English	2096	0.737	< 0.001
1 and 3	Spanish	8386	0.895	< 0.001
2 and 3	Arabic	523	0.892	< 0.001
2 and 3	English	2096	0.738	< 0.001
2 and 3	Spanish	8386	0.882	< 0.001

**Table 11. Spearman's Correlations by Year**

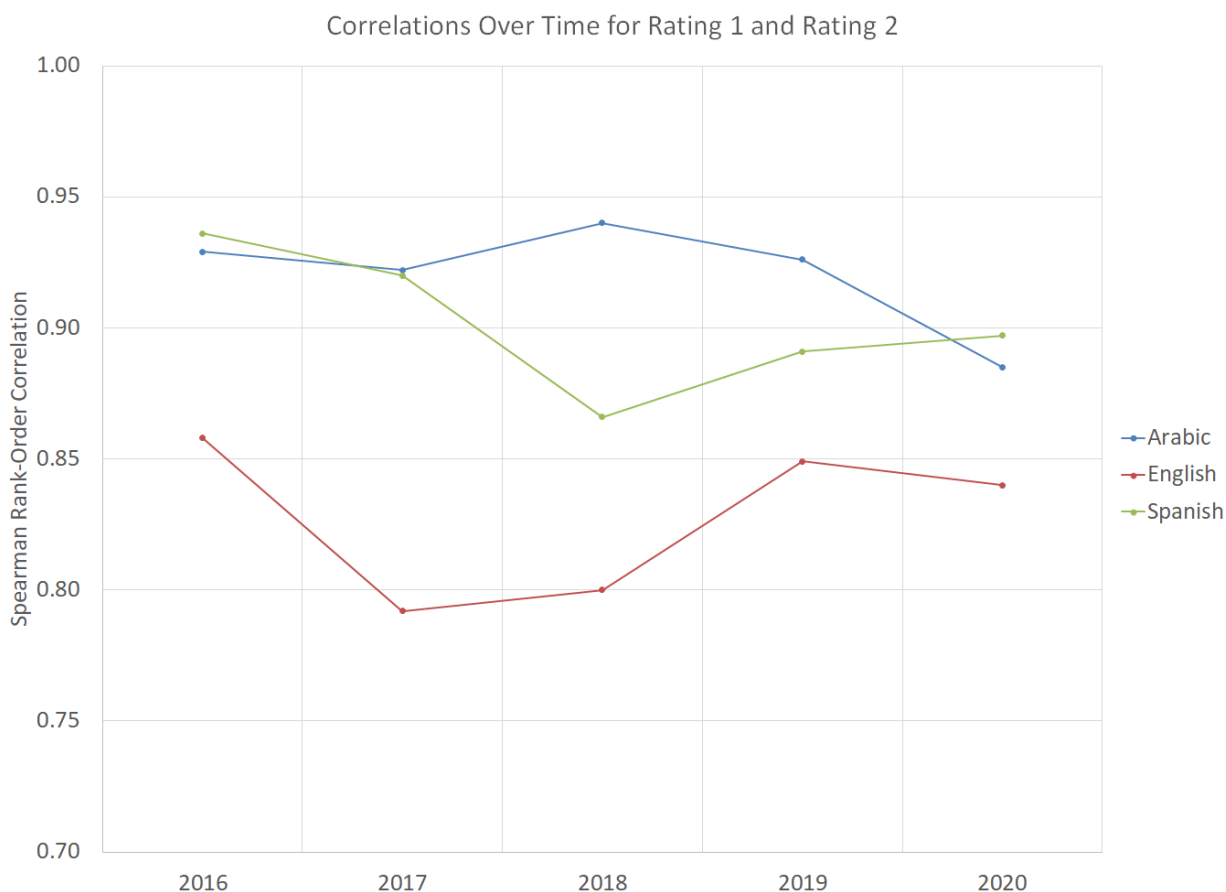
<b>Language</b>	<b>Ratings</b>	<b>Year</b>	<b>N</b>	<b><math>\rho</math></b>	<b><i>p</i>-value</b>
Arabic	1 and 2	2016	163	0.929	< 0.001
	1 and 2	2017	391	0.922	< 0.001
	1 and 2	2018	235	0.940	< 0.001
	1 and 2	2019	269	0.926	< 0.001
	1 and 2	2020	68	0.885	< 0.001
English	1 and 2	2016	1309	0.858	< 0.001
	1 and 2	2017	977	0.792	< 0.001
	1 and 2	2018	871	0.800	< 0.001
	1 and 2	2019	1226	0.849	< 0.001
	1 and 2	2020	200	0.840	< 0.001
Spanish	1 and 2	2016	3783	0.936	< 0.001
	1 and 2	2017	5029	0.920	< 0.001
	1 and 2	2018	5151	0.866	< 0.001
	1 and 2	2019	4289	0.891	< 0.001
	1 and 2	2020	1080	0.897	< 0.001

**Table 11. Spearman's Correlations by Year**

Language	Ratings	Year	N	$\rho$	$p$ -value
Arabic	1 and 3	2016	70	0.895	< 0.001
	1 and 3	2017	212	0.867	< 0.001
	1 and 3	2018	99	0.888	< 0.001
	1 and 3	2019	113	0.855	< 0.001
	1 and 3	2020	29	0.556	< 0.001
English	1 and 3	2016	604	0.780	< 0.001
	1 and 3	2017	442	0.700	< 0.001
	1 and 3	2018	405	0.672	< 0.001
	1 and 3	2019	560	0.755	< 0.001
	1 and 3	2020	85	0.668	< 0.001
Spanish	1 and 3	2016	1503	0.908	< 0.001
	1 and 3	2017	2247	0.926	< 0.001
	1 and 3	2018	2415	0.862	< 0.001
	1 and 3	2019	1784	0.869	< 0.001
	1 and 3	2020	437	0.871	< 0.001
Arabic	2 and 3	2016	70	0.940	< 0.001
	2 and 3	2017	212	0.872	< 0.001
	2 and 3	2018	99	0.882	< 0.001
	2 and 3	2019	113	0.869	< 0.001
	2 and 3	2020	29	0.755	< 0.001
English	2 and 3	2016	604	0.749	< 0.001
	2 and 3	2017	442	0.728	< 0.001
	2 and 3	2018	405	0.667	< 0.001
	2 and 3	2019	560	0.781	< 0.001
	2 and 3	2020	85	0.674	< 0.001
Spanish	2 and 3	2016	1503	0.904	< 0.001
	2 and 3	2017	2247	0.912	< 0.001
	2 and 3	2018	2415	0.835	< 0.001
	2 and 3	2019	1784	0.865	< 0.001
	2 and 3	2020	437	0.860	< 0.001

Overall, the results of this analysis suggest that the Ratings are reasonably in agreement with each other and the correlations of the ratings are almost always positive and strong. In the summary of the Rating 1 and Rating 2 correlations over time, Figure 1 shows that the correlations of the first two Ratings of the exams have a correlation at or above 0.885 for Arabic and Spanish and at or above 0.792 for English. The Ratings for the English exam were lower than that of the Arabic and Spanish; however, the difference in the correlations between English and the other two languages were less pronounced in 2020 compared to previous years. This observation may be explained by restriction of range for 2020, in that data was only

available into the month of March for that year. It is possible that examinees in the early part of the year are not representative of the full year.

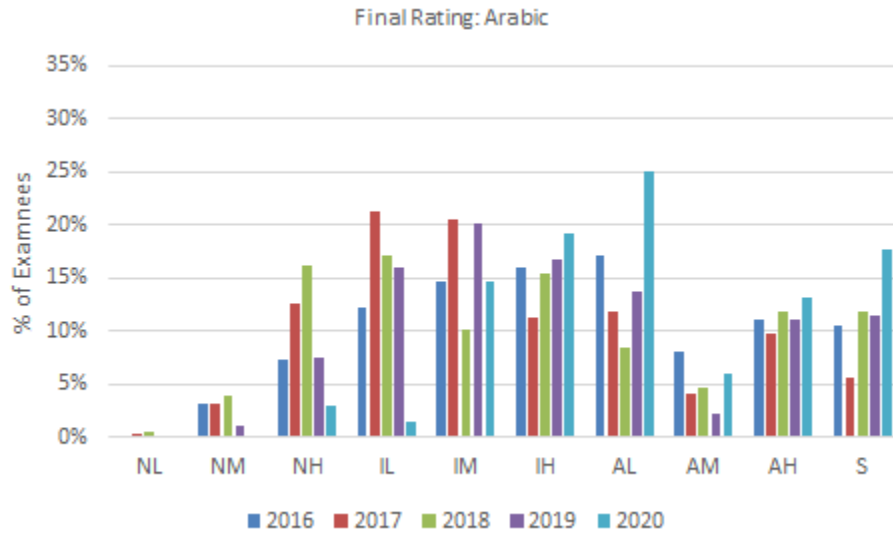


**Figure 1. Spearman-rank correlations of Rating 1 and Rating 2 from 2016 to 2020**

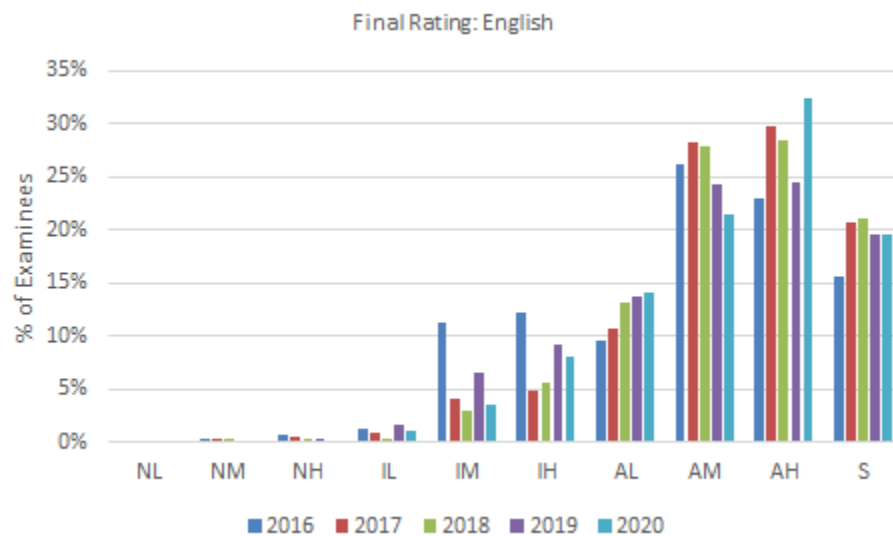
## Score Stability Over Time

An analysis was conducted to analyze the percent of each final rating over time. Figures 2-4 show the results graphically. For the Arabic exam, the distribution of the final exam ratings were similar (within approximately 10%) over time for most final ratings; however, the 2020 ratings were higher for the categories of “AL”, and “S” than in any previous year and lower in the categories of “NH” and “IL”. The distribution of final ratings for the English and Spanish exams were reasonably similar over time.





**Figure 2. Final ratings from 2016 to 2020 for the Arabic OPIc**



**Figure 3. Final ratings from 2016 to 2020 for the English OPIc**

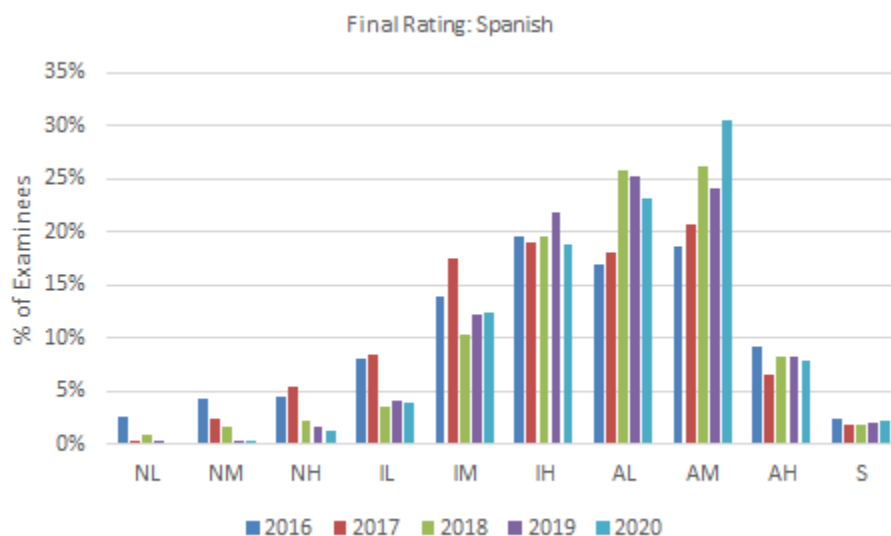


Figure 4. Final ratings from 2016 to 2020 for the Spanish OPIc

## Evidence of Validity


### Content Related

OPIc prompts are representative of the domain for which it is designed to measure – language proficiency (speaking) as per the *ACTFL Proficiency Guidelines*. To assess a test taker’s performance via a ratable sample of the language, the OPIc is programmed to establish a speaker’s level of consistent functional ability (patterns of strength) as well as the upper limits of that ability (patterns of weakness) through standardized assessment criteria (function, context/content, accuracy, text type). These function-related tasks are derived directly from the *ACTFL Proficiency Guidelines for Speaking*, and prompt writers are trained to adhere to the Guidelines for function-related guidance.

That is, the rationale for the types of tasks included in the ACTFL OPIc is rooted in the fact that it is a criterion-referenced assessment. In this case, OPIc elicitation tasks are derived from the *ACTFL Proficiency Guidelines*. At each level, the Guidelines identify the types of things the test taker can and cannot do with the language. As such, the OPIc tasks are necessarily based on the functions identified in the criteria of reference-- the *ACTFL Proficiency Guidelines*.

### Criterion Related

Scores from the current OPIc® have not been compared with any related measures of language performance that would allow for criterion-related validity evidence. Given that the exam scores are used for a variety of purposes including language fluency certification, employment selection, placement, and college credit; standardized measures of later performance would be



difficult to obtain. In addition, the OPIC® is not meant for use as a predictor of performance, but rather as a global assessment of functional speaking ability in a language that can indicate readiness for a given purpose. Since the intended use of the exam is not to predict levels of performance, traditional criterion-related validity evidence is not directly applicable.

## Construct Related

Traditional construct-related evidence typically involves correlation of one measure of a trait with other measures of the same or similar traits. It is not unusual for researchers to gather such data with, for example, psychological measures where the trait is tested indirectly (e.g., depression inventories). Scores from the current OPIC® have not been compared with any related tests of language ability largely because the OPIC® is a direct measure of language ability, and high correlations with similar direct measures of language ability would add little to the validity argument.

## Possible Test Bias

The use of a Background Survey allows the test taker to avoid the selection of items that might be insensitive or irrelevant for the test taker. In an effort to ensure that test takers are not offended or made uneasy while taking a OPIC®, item writers are instructed to avoid sensitive topics (e.g., immigration, national origin, sexual preference, religion, marital status, racism, political viewpoint) when developing OPIC® prompts. However, no demographic data is collected on the examinees that would allow for measurement of bias or adverse impact.

## Evidence that Time Limits are Appropriate and that the Exam is not Unduly Speeded

The OPIC is comprised of 12-15 prompts that are timed and aimed at adjacent levels (Novice/Intermediate, Intermediate/Advanced, and Advanced/Superior) based on the results of the Self-Assessment. The candidate is given 30 seconds to respond to Novice-level prompts, 60 seconds to respond to Intermediate-level prompts, 2 minutes to respond to Advanced-level prompts, and 2 minutes and 30 seconds to respond to Superior-level prompts. The amount of varied and topical prompts set within the limited linguistic range of the test, coupled with the allotted length of the response, gives the test candidate several repeated opportunities to demonstrate their language ability.

## Provisions for Standardizing Administration of the Examination

Administration procedures for the OPIC® are described on pages 5-10 of the *ACTFL OPIC Examinee Handbook*. The OPIC® is administered in live proctor or remote proctor settings.

## Irrelevant Sources of Difficulty Affecting Test Scores

A formal study of construct irrelevant variance for the OPIc® has not been undertaken. However, some likely sources of construct irrelevant variance are addressed through ACTFL's exam policies and procedures. Rater training is extensive, and scoring is done against a standardized rubric (see the *ACTFL OPIc® Rater Training Manual*, pages 15-21. The use of the background survey to select prompts most likely to be familiar to the examinee might help to minimize context effects (see the *ACTFL OPIc® Rater Training Manual*, page 9). As described above, administration procedures are standardized to ensure the examinee testing experience varies as little as possible.

## Provisions for Exam Security

Per *ACTFL's Assessment Integrity Policy*, "A test taker's language must be representative of their own language abilities (speaking, writing, listening, or reading) at the time of the test." Measures have been put into place in order to protect both test content but also the proficiency-based framework for this assessment.

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 must 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 must be a Professor, Department Chair, Department Administrative Assistant, or Department Coordinator. No other administrators or staff members 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. For branch offices without an on-site human resource representative, a senior-level manager may act as proctor.



In addition, educational or business proctors must have a work e-mail address; the e-mail address must contain the proctor's name and the organization's name. Personal e-mail addresses (e.g., AOL, Hotmail, Comcast, Verizon) are not accepted for proctors.

In addition to face to face proctoring, ACTFL also offers remote (virtual) proctoring which make use of a test taker's webcam to identify the test taker and monitor the computer screen and testing environment.

### **Security Measures**

Each test candidate must 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 3,200 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 2 weeks. Once a candidate has logged into an assessment, the candidate must complete that assessment in one sitting within 1 hour. If a test candidate tries to access another website while logged into the assessment, the OPIc® will close; only a proctor can log the candidate back in.


Raters also listen for suspicious behavior: for example, the sound of someone helping the candidate or a change in the candidate's voice. Raters are instructed to assign the score of UR for "unratable" and to notify Language Testing International (LTI) test administration of "suspicious behavior" which is then investigated by the Director of Test Administration. The interview is digitally recorded by the tester within the LTI Test Management System (TMS) and uploaded instantaneously to LTI's secure database. The record is stored under a test identification number which may be looked up on the certificate verification site.

Finally, ACTFL's retest policy prohibits re-taking the OPIc within ninety days of a testing period. This prevents unnecessary exposure of test items and reinforces the proficiency-based framework of the assessment.

## **Interpretations and Conclusions**

To conclude, the ACTFL OPIc met the minimum inter-rater reliability and agreement requirements. Ratings were in agreement within one sublevel of each other over 92% of the time, and within two sublevels 99% of the time. Additionally, the findings of the Spearman's *R* Correlation analyses demonstrate that the correlations of the ratings are almost always positive and strong, ranging from 0.74- 0.93 across languages. Across all three languages, the highest absolute agreement was Superior, and the lowest absolute agreement was found at Intermediate Low.

The findings of the Spearman's *R* Correlation analyses demonstrate that the correlations of the ratings are almost always positive and strong, ranging from 0.91-0.97. The results also suggest that the ratings are fairly in agreement with one another. Suggested areas of improvement



based on the analyses include a focus to the absolute agreement between raters within the Advanced Mid and Superior borders for English and Spanish, and within the Intermediate levels for English and Arabic. The results of this analysis confirm the reliability of the ACTFL OPIC as an assessment of oral proficiency.

## References

ACTFL (2020, May). *Quality Control of ACTFL Assessments: Assessment Integrity Policy*.

<https://www.actfl.org/center-assessment-research-and-development/actfl-assessments/quality-control-actfl-assessments>

ACTFL (2020). *Methods of Safeguarding ACTFL Assessments within Test Administration Processes and Procedures*. (ACTFL Publication No. AAR/SOP/P--2020-001). Alexandria, VA: ACTFL Assessment Program.

ACTFL (2020). *ACTFL Oral Proficiency Interview-- Computer Familiarization Guide*.

<https://www.actfl.org/sites/default/files/assessments/OPIc%20Familiarization%20Guide%2020.pdf>

ACTFL (2020, May). *Quality Control of ACTFL Assessments: Record Retention Policy*.

<https://www.actfl.org/center-assessment-research-and-development/actfl-assessments/quality-control-actfl-assessments>

ACTFL (2020, May). *Quality Control of ACTFL Assessments: Re Test Policy*.

<https://www.actfl.org/center-assessment-research-and-development/actfl-assessments/quality-control-actfl-assessments>

ACTFL and Language Testing International (2019). *ACTFL OPIc Examinee Handbook*.

<https://www.languagetesting.com/pub/media/wysiwyg/manuals/opic-examinee-handbook.pdf>

American Council on the Teaching of Foreign Languages (2012). *ACTFL Proficiency Guidelines*.

<https://www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012>

Swender, E. & Vicars, R. (eds.) (2012). *ACTFL Oral Proficiency Interview Tester Training Manual*. Alexandria, VA: ACTFL.