ACTFL
AMERICAN COUNCIL ON THE TEACHING OF FOREIGN LANGUAGES

AMERICAN COUNCIL ON THE TEACHING OF FOREIGN LANGUAGES
1001 North Fairfax Street, Suite 200 | Alexandria, VA 22314 | P 703-894-2900 | F 703-894-2905
445 Hamilton Avenue, Suite 1104 | White Plains, NY 10601-1832 | P 914-963-8830 | F 914-963-1275
www.actfl.org | www.leadwithlanguages.org | facebook.com/actfl | @actfl

Examination Evaluation of the ACTFL OPI® in French, Korean, Mandarin for the ACE Review

Prepared for:
American Council on the Teaching of Foreign Languages (ACTFL)
White Plains, NY

Principle Investigator: Stephen Cubbellotti, Ph.D.,
Independent Psychometric Consultant
ACTFL Consultant: Troy Cox, Ph.D.,
Brigham Young University

EXECUTIVE SUMMARY

This report documents the American Council on the Teaching of Foreign Languages (ACTFL) Oral Proficiency Interview (OPI®) from 2012 to 2014 to satisfy a review requirement of the American Council of Education College Credit Recommendation Service (CREDIT) program. The ACTFL OPI® is a live interview conducted telephonically or face-to-face between an ACTFL Certified OPI® Tester and the individual whose language proficiency is being assessed.

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 OPI® scores, inter-rater reliability was measured by the Spearman's *R* 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 *R* and absolute/adjacent agreement results provides sufficient information about reliability.

Comparisons of ACTFL OPI® inter-rater reliability and agreement were made across three languages: French, Korean, and Mandarin. 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.

The ACTFL OPI® 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 OPI® process yields relatively stable reliability results over time.

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

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.

Table of Contents

General Test Information

Rationale and Purpose of the test

The ACTFL OPI® is a live interview conducted telephonically or face-to-face between an ACTFL Certified OPI® Tester and the individual whose language proficiency is being assessed. The interview lasts between 20 and 30 minutes depending on the proficiency level of the test taker.

The primary goal of the OPI® is the efficient elicitation of a ratable sample. 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. A ratable sample is elicited through a series of personalized questions that adhere strictly to a standardized elicitation protocol. While the OPI® may resemble a conversation between the tester and the test taker, in fact, the tester follows strict elicitation protocol and structures the interview to respect that protocol.

The "floor" and "ceiling" are determined by the test taker's ability to meet the assessment criteria of the major levels of the ACTFL Proficiency Guidelines and ACTFL Rating Scale. Each major level is defined as a confluence of function (task), text type, content and accuracy expectations that grow exponentially as the scale progresses. Instead of defining oral language proficiency as a single, unitary construct, each major level is defined as its own construct with the subsumption principle that as the scale progresses, the relationship between the levels is hierarchal. A ratable sample is one in which the speaker demonstrates a base level or "floor" of sustained performance of the tasks (functions) for that level and a "ceiling" level at which the examinee cannot sustain performance of the tasks of the next higher level. Hence, each OPI® Interview contains the performance of tasks from at least two contiguous major levels.

The difference between the sublevels, then, is based on the quality and quantity of the examinee's language when s/he is engaged in at-level tasks. The *low sublevel* is indicative of a speaker who just barely demonstrates competence when performing the tasks for the major level. The *mid sublevel* indicates that the speaker fulfills all the requirements of the major level with quantity and quality of language across the assessment criteria. There is no doubt that the examinee can perform the functions of that major level; the quality and quantity of language is much more substantial than that of speakers at the low sublevel. The *high sublevel* rating is different in that, not only does it indicate a speaker's robust ability to meet the criteria for the major level, but provides information related to what happens to the speaker's language when s/he is attempting to meet the expectations for the next higher (adjacent) major level. A rating at the high sublevel is indicative of performance at the next higher level most of the time, that is to say, the speaker is unable to sustain all the criteria all the time.

The OPI® assesses language proficiency in terms of the ability to use spoken language effectively and appropriately in real-life situations. It does not address when, where, why, or the way in which a speaker has acquired his/her language. The OPI® is not an achievement test assessing a speaker's acquisition of specific aspects of course and curriculum content, nor is it tied to any specific method of instruction. The OPI® does not compare one individual's performance to others, but each individual performance to the assessment criteria.

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

No authorship has ever been ascribed to the ACTFL OPI®. The OPI® was originally based on the Speaking Test of the Foreign Service Institute (FSI) created in the mid-1950s by Claudia Wilds of FSI in consultation with John B. Carroll of Harvard University. The FSI Speaking test was then adopted by the Interagency Language Roundtable (ILR) for use at the Central Intelligence Agency (CIA) and the Defense Language Institute (DLI). These speaking tests were designed to elicit speech samples that would align with the government's Interagency Language Roundtable (ILR) proficiency scale. Instead of creating a set of detailed test specifications with specific item specifications, a stringent interview protocol was created that would elicit the speech that would demonstrate what was being assessed.

In 1982, the government's OPI® was adapted for use outside of the government context. The authors of the first ACTFL Manual with its subsequent descriptions of the ACTFL proficiency scale were Dr. Pardee Lowe, Jr. and Dr. Judith E. Liskin-Gasparro. Pardee Lowe, Jr. (Ph.D., University of California, Berkeley) was former chair of the ILR Testing Committee and Chief of Testing at the Central Intelligence Agency's Language School and consulted widely on language testing. As a member of the ACTFL Guidelines Project, he focused on the guidelines' commensurability with the government's scales so that a national standard might evolve. Since the original training, experienced OPI® testers have been recruited to conduct training and to ensure the protocols practiced by current testers retain their alignment with the ACTFL Proficiency Guidelines.

Types of scores reported to examinees

Examinee scores are reported as the major level and sublevel according to the *ACTFL Proficiency Guidelines 2012 - Speaking.* 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. The description of each major level is representative of a specific range of abilities.

Directions for scoring and procedures and keys

A [1]Certified ACTFL tester elicits a sample of speech by asking questions that target the functions of the test-taker's floor level (the level at which the speaker is able to sustain all of the criteria for the level) and ceiling level (the level at which the speaker is no longer able to sustain the criteria for the level) across a variety of topics. The speech sample is digitally recorded and stored on a secure Internet-based archive. Patterns of strength and weakness in accomplishing the assigned tasks are established by the tester. The speech is first placed within a major range and then matched to the sub-level description in the *ACTFL Proficiency Guidelines 2012 – Speaking.* A first rating is assigned by the tester after which the sample is then independently second rated by a second certified tester who is able to access the data base. The two ratings must agree exactly. Any rating discrepancy is blindly arbitrated by a third rater and an official ACTFL rating is assigned when two ratings agree exactly.

[1] For a more in depth discussion of the rating process, please refer to the section on *Scorer Reliability for Extended Response Items*

ACTFL tests are integrative tests, i.e., they address a number of abilities simultaneously and look 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 how they contribute to the overall sample. The test taker is evaluated through the lens of proficiency. Though holistically rated, there are four major categories of assessment criteria on which ratings are focused: the global tasks/functions performed with the language, the social contexts and content areas in which the language can be used, the accuracy features which define how well the speaker performs the task pertinent to those contexts and content areas, and the oral text type (from individual words to extended discourse) produced. The assessment criteria used to evaluate speaking are summarized in the chart below:

Proficiency Level*	Global Tasks and Functions	Context / *Content*	Accuracy	Text Type
Superior	Discuss topics extensively, support opinions, and hypothesize. Deal with a linguistically unfamiliar situation.	Most formal and informal settings. *Wide range of general interest topics and some special fields of interest and expertise.*	No pattern of errors in basic structures. 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.	Most informal and some formal settings. *Topics of personal and general interest.*	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.	Some informal settings and a limited number of transactional situations. *Predictable, familiar topics related to daily activities.*	Understood, with some repetition, by speakers accustomed to dealing with non-native speakers.	Discrete sentences
Novice	Communicate minimally with formulaic and rote utterances, lists, and phrases.	Most common informal settings. *Most common aspects of daily life.*	May be difficult to understand, even for speakers accustomed to dealing with non-native speakers.	Individual words and phrases

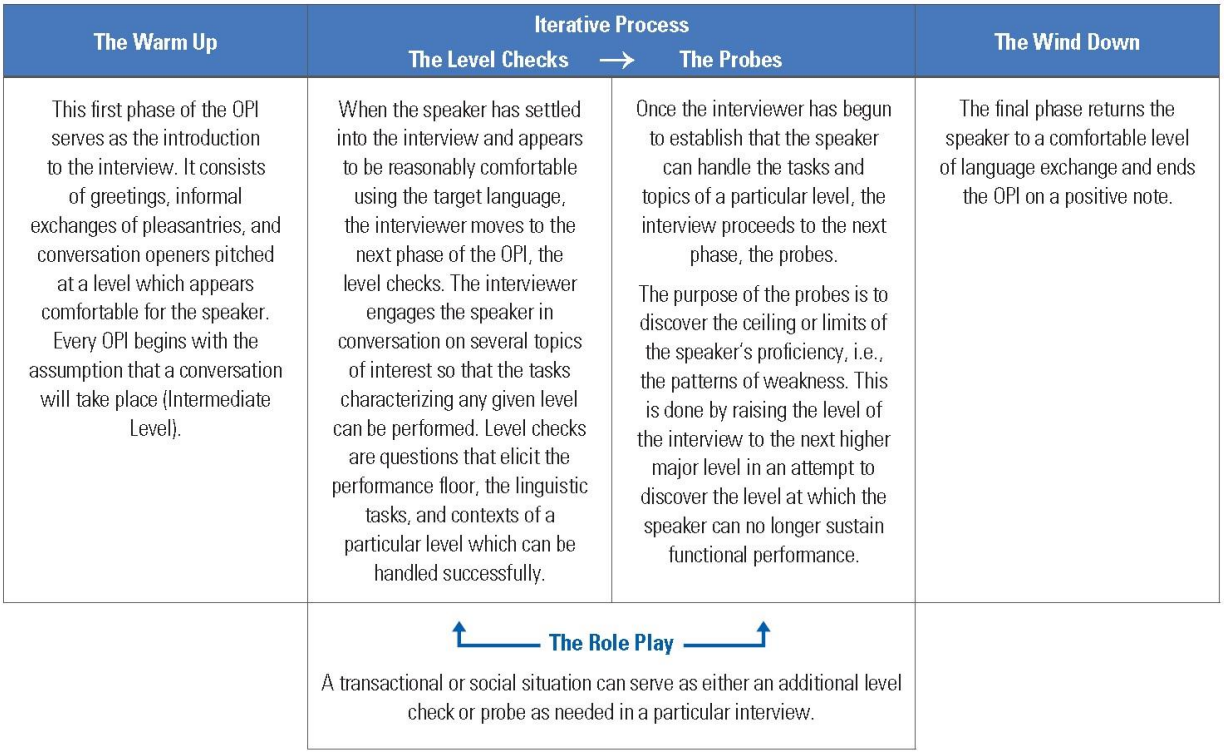
Item/Test Content Development

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

The ACTFL OPI® is an interactive and adaptive interview protocol that results in a unique and ratable sample of speech. No two interviews are exactly the same. The interview is interactive in that the questions elicited by the tester are based on the responses that the test taker provides. The interview is adaptive in that content areas are based on the interests and experiences of the examinee, and the major levels that are targeted are based on the linguistic range demonstrated by the examinee. There are required item types used to target each level of proficiency. For example, the request for the telling of a story from beginning to end is an effective item "type" or request "type" when eliciting a past narration at the Advanced level. Testers are trained in the types of requests that are most effective for elicitation at all

levels. Each certified tester is, in effect, an item writer, creating the prompts that elicit demonstration of the major functions and other assessment criteria according to the ACTFL Scale.

Following the "Warm-Up" phase, during which a preliminary range of ability of the test taker is determined, the tester seeks evidence of the "floor" through a series of "level checks" (questions that target the functions of the floor level). Level checks are followed by "probes" (questions that target the functions of the next higher level). The interview continues in an iterative process of alternating "level checks" and "probes", thereby clearly demonstrating the test taker's strengths and weaknesses across two contiguous major levels. Once the tester is satisfied that a ratable sample has been produced, a "Wind-Down" phase brings closure to the interview. The standardized elicitation protocol of the OPI® can be seen in the chart below:



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 OPI® measures how well a person spontaneously speaks language during in live interpersonal communication 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 OPI 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 contiguous 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. The OPI elicitation protocol is based on producing a sample of spoken language that can be rated according to the Guidelines.

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.

Common topics at the Intermediate level are self, home, family, daily routine, interests; at the Advanced level the topics expand to those of general and community interest; at the Superior level, topics expand to the issues level and are treated from an abstract perspective. While certain topics may be associated with a specific level, a topic can be treated at any level. Certified OPI testers are able to explore and develop topics across levels. For example, the topic of school can be developed at the Intermediate level by asking about daily class activities; at the Advanced level by asking for an anecdote about one's first day in a new school; at the Superior level by asking for a discussion about the value of higher education.

Description of the item sensitivity panel review

OPI topics are drawn from the interests and experiences of the test taker as provided during the interview. Testers are also trained to recognize when a seemingly neutral topic may be perceived as sensitive by the test taker and can change topics. Because of the interactive nature of the OPI, a test taker may decline a topic at any time. OPI Testers are instructed to avoid sensitive topics (e.g., immigration, national origin, sexual preference, religion, marital status, racism, etc.) when administering the OPI.

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

There is no "final form" for the ACTFL OPI as each OPI is based on the interests, experiences, and linguistic ability of the test taker.

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

OPI elicitation protocol targets 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. In a recent conference report, Surface, Dierdorff, and Poncheri (2006) found strong support for favorable inter-rater reliability for the OPI®, especially for the Spanish version. Further, the majority of rater pairs were making identical proficiency level judgments when scoring the OPI®.

SWA consulting (2012) found Spearman Rs exceeded the standard for use, ranging from 0.92 to 0.98 across languages and years analyzed. 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 abbreviations:

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 French OPI® from 2012 to 2014

		Rater 1									
		NL	NM	NH	IL	IM	IH	AL	AM	AH	S
Rater 2	NL	18	7	0	0	0	0	0	0	0	0
	NM	15	220	0	0	0	0	0	0	0	0
	NH	0	0	8	30	3	0	0	0	0	0
	IL	0	0	44	1024	142	1	0	0	0	0
	IM	0	0	1	257	1274	43	0	2	0	0
	IH	0	0	0	1	48	726	73	5	0	0
	AL	0	0	0	0	4	120	329	162	8	0
	AM	0	0	0	0	0	3	90	477	32	3
	AH	0	0	0	0	0	0	3	29	135	85
	S	0	0	0	0	0	0	0	1	91	332

Table 2 Concordance Table for Korean OPI® from 2012 to 2014

		Rater 1									
		NL	NM	NH	IL	IM	IH	AL	AM	AH	S
Rater 2	NL	27	7	0	0	0	0	0	0	0	0
	NM	11	96	10	0	0	0	0	0	0	0
	NH	0	9	45	23	2	0	0	0	0	0
	IL	0	2	22	319	35	1	0	0	0	0
	IM	0	0	2	53	113	5	0	0	0	0
	IH	0	0	0	0	14	334	20	1	0	0
	AL	0	0	0	0	2	24	44	35	2	0
	AM	0	0	0	0	1	2	42	278	11	0
	AH	0	0	0	0	0	0	4	17	36	69
	S	0	0	0	0	0	0	0	1	64	574

Table 3 Concordance Table for Mandarin OPI® from 2012 to 2014

		Rater 1									
		NL	NM	NH	IL	IM	IH	AL	AM	AH	S
Rater 2	NL	6	11	1	0	0	0	0	0	0	0
	NM	16	124	9	1	0	0	0	0	0	0
	NH	0	12	157	27	2	0	0	0	0	0
	IL	0	0	24	322	54	0	0	0	0	0
	IM	0	0	3	119	599	27	7	0	0	0
	IH	0	0	0	5	63	232	61	13	0	0
	AL	0	0	0	0	14	77	588	106	0	1
	AM	0	0	0	0	1	14	112	784	44	9
	AH	0	0	0	0	0	0	2	56	247	69
	S	0	0	0	0	0	0	0	6	51	1732

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 OPI® data because the proficiency categories used for ACTFL OPI® 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 OPI® data in that the proficiency categories are ordinal in nature.

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

Language	N	ρ	95% CI LL	95% CI UL
French	5848	0.961	0.959	0.965
Korean	2358	0.979	0.977	0.982
Mandarin	5811	0.979	0.977	0.981

Table 5 Spearman's Correlations by Year

Language	Year	N	ρ
French	2012	1801	0.966
	2013	1937	0.961
	2014	2110	0.957
Korean	2012	770	0.980
	2013	785	0.977
	2014	803	0.979
Mandarin	2012	2015	0.979
	2013	2038	0.977
	2014	1757	0.982

Overall, the ACTFL OPI® exceeded inter-rater reliability minimum standards and was quite high. The Spearman's R correlation was .961 for French, .979 for Korean, and .979 for Mandarin. Inter-rater reliability was high across language categories and interview years. These results are consistent with previous year's 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 OPI® is a protocol designed to elicit a ratable sample of speech following a standardized interview format. Due to the interactive nature of the interview, test forms in the traditional sense are not used. Rather equivalency is created by the certified tester based on the response patterns of the examinee. The tester provides the candidate with multiple opportunities to demonstrate his/her ability. This establishes the candidate's "floor," indicating his/her ability to perform the communicative functions associated with a given ACTFL level. Once a "floor" is established, the tester targets the functions of the next higher level on content areas that come from the interests and experiences of the examinee. No two interviews are exactly the same in content, but are the same as far as the specific communicative functions over a range of topics familiar to the candidate which the tester is required to elicit from the test taker across two contiguous major levels of proficiency. The equivalence of the forms comes from the ability of trained raters to respect the OPI elicitation protocol and structure and to consistently assign the same rating to a speech sample by reflecting on the criteria for each level descriptor contained in the *ACTFL Proficiency Guidelines 2012 – Speaking.*

Scorer reliability for extended response items

ACTFL Certified OPI® Testers are highly specialized language professionals who have completed a rigorous training process. They complete a minimum of 32 face-to-face hours of training followed by over 100 hours of independent training in which they receive formative feedback on both rating speech samples and conducting interviews. The certification process concludes successfully with a tester's demonstrated ability to consistently elicit ratable speech samples and consistently rate samples with a high degree of reliability. (OPI® Tester Training Manual). Current testers participate in regular norming sessions to ensure the ratings they award are on standard and must renew their certification every four years. Raters employed through LTI uphold the highest professional and ethical standards in test administration and rating. This level of training ensures that the ratings are reliable.

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 the 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 from 2012-2014

Language	N	Absolute Agreement (exact)	Adjacent Agreement (+/- 1 sublevel)
French	5848	78%	100%
Korean	2358	79%	99%
Mandarin	5811	82%	98%

Table 7 Absolute/Adjacent Agreement by Language and Year

Language	Year	N	Absolute Agreement (exact)	Adjacent Agreement (+/- 1 sublevel)
French	2012	1801	75%	99%
	2013	1937	79%	99%
	2014	2110	78%	99%
Korean	2012	770	79%	99%
	2013	785	79%	99%
	2014	803	79%	99%
Mandarin	2012	2015	81%	98%
	2013	2038	84%	99%
	2014	1757	83%	100%

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

Language	Rating	N	Absolute Agreement (exact)	Adjacent Agreement (+/- 1)
French	Novice Low	27	55%	100%
	Novice Mid	233	97%	100%
	Novice High	34	15%	98%
	Intermediate Low	1235	78%	100%
	Intermediate Mid	1566	87%	100%
	Intermediate High	34	81%	99%
	Advanced Low	555	66%	99%
	Advanced Mid	647	71%	99%
	Advanced High	282	51%	97%
	Superior	398	55%	100%
Korean	Novice Low	35	71%	100%
	Novice Mid	119	84%	98%
	Novice High	74	57%	98%
	Intermediate Low	410	81%	100%

	Intermediate Mid	153	68%	97%
	Intermediate High	367	91%	99%
	Advanced Low	78	40%	96%
	Advanced Mid	359	84%	100%
	Advanced High	122	32%	98%
	Superior	641	89%	100%
Mandarin	Novice Low	13	27%	100%
	Novice Mid	158	84%	100%
	Novice High	193	81%	98%
	Intermediate Low	436	68%	99%
	Intermediate Mid	744	82%	98%
	Intermediate High	381	66%	96%
	Advanced Low	771	76%	98%
	Advanced Mid	964	81%	98%
	Advanced High	359	72%	100%
	Superior	1791	27%	100%

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 OPI® as an assessment of speaking proficiency. There is some concern for the lack of agreement in extreme sublevel categories and in some languages (e.g., Novice Low and Superior in Mandarin at 27%). Areas for continuous improvement include increasing rater agreement within the Novice High-Intermediate Low border. This is especially true for French and Korean (15% and 57% absolute agreement at Novice High, respectively) and Mandarin at 68% absolute agreement at Intermediate Low. Although the research indicates 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 OPI® ratings are listed in the chart below:

Official ACTFL OPI® 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)

Currently, the minimum score for granting college credit for an ACTFL OPI® rating is Novice High. The higher the OPI® rating, the greater the number of recommended credits. 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 distribution between the test prompts and the content area they are intended to assess and is determined by content experts. The content experts in this case are the certified ACTFL testers and raters involved in the interview and rating processes. Evidence of content-related validity comes from obtaining a ratable sample of speech. A ratable sample entails requiring the examinee to perform functions associated with two adjacent levels across a range of topics/content domains. Since there is high inter-rater reliability, there is evidence that the testers are eliciting the needed functions at the appropriate levels and that other experts are able to identify those functions and provide a similar second rating.

Criterion-related validity

Criterion-related validity has been defined as the degree to which an exam measures the criteria it purports to measure. There is a preponderance of evidence of criterion-related validity because the goal of the OPI® is to elicit a sample of language that clearly reflects the assessment criteria associated with the ACTFL scale and the *ACTFL Proficiency Guidelines 2012 - Speaking*. 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.

Since 1993, the descriptions of spoken language ability contained in the *ACTFL Proficiency Guidelines*, as well as samples of speakers at different levels of proficiency, have been used in "language needs assessments" (LNA), that is, "standard setting" procedures to determine minimum proficiency levels needed to perform real-world communication tasks. Subject Matter Experts are assembled who are familiar with the communication tasks needed to fulfill, for example, the exit requirements for world language majors or world language teachers or the requirements to credential teachers. Such standard setting took place during the development of the ACTFL World Language Teaching Standards, to set recommended speaking and writing proficiency standards for World Language teachers, as well as similar projects conducted by individual State Teacher Licensing Boards. Similar "task analyses" and LNAs have been performed with 100 Fortune companies and different federal agencies to set minimum levels of proficiency needed to perform the language tasks associated with over 80 different job titles. Commercial and governmental SMEs found that the *ACTFL Proficiency Guidelines* effectively articulated a hierarchy of language ability. The SMEs agreed that that each level described increasing competence in terms of the

communicative tasks performed across an increasingly broader range of content/context areas, with a higher degree of organization and elaboration, and with an increasing degree of articulation and precision as one moves up the scale. What researchers found when examining the responses of SMEs in a standard setting was that they could consistently identify the same level of proficiency when they used the ACTFL framework to describe language competence needed to perform real-world tasks.

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®.

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 Bärenfänger purport that since the study used very experienced tester trainers and testers for The European Language Certificates (TELC), which norms testers and raters across European languages, 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 determined based on the ability and interests of the test taker, no two interviews are the same and a test of item bias would not be appropriate.

The OPI® is used by many corporations and government agencies who do track the demographic information of their applicants and employees, as well as proficiency ratings issued by ACTFL. These clients periodically conduct "Adverse Impact" analyses to see if the ACTFL assessment is biased against

a gender, sex, race and age. These corporations have not shared the results of their analyses with LTI and ACTFL, but have reported that they did not find any adverse impact as a result of their analyses.

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

The ACTFL Oral Proficiency Interview, or OPI®, is a live 20-30 minute conversation, depending on the level of proficiency of the candidate. It is not a timed test; instead it ends when the tester has gathered a robust speech sample of the not only the floor, but a clear ceiling of the candidates performance.

Provisions for standardizing administration of the examination

OPIs must be proctored by a trusted, responsible individual. In academic settings, this is often a school employee. In business settings, it is ideally a member of the HR department of the organization or a trained Test Control Officer (TCO). When the interviews are administered face-to-face, the purpose of the proctor is to verify the examinee's identity. When interviews are administered telephonically, the proctor both verifies the examinee's identity and ensures no unauthorized material or help is available to the examinee. This individual, nominated by the organizing agency, will read detailed proctor instructions and sign a Proctor Agreement Form, which is collected by LTI in advance of the assessment. This procedure guarantees the identity of the candidate and the conditions under which the test is taken.

Provisions for exam security

The ACTFL OPI® is a live interview with a certified tester. The tester has been trained to adapt and create appropriate test questions based on the background and interests of the candidate. All official OPIs are proctored to ensure that candidates do not record the prompts they are given.

When the OPI® is administered within 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 is permitted to act as proctors.

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 is permitted to act as proctors.

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

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 OPI® 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

There is no pool of items for this exam. When eliciting a ratable sample, successful testers ask open-ended questions in order to encourage the speaker to show his or her language ability at its best. Testers listen to and process a speaker's responses before formulating subsequent questions. Testers ask purposeful questions; every question targets a specific level and function of that level. Testers explore a variety of topics within and across major levels of the ACTFL scale.

Validity of Computer Administration

Size of the operational test item pool for test

This is not applicable as this is not a computer-delivered test.

Cut-score information

Rationale for the particular cut-score recommended

Once a ratable sample of speech has been elicited, that sample is compared 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) 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.*

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