ACTFL
Language Connects

ACTFL ORAL PROFICIENCY INTERVIEW

Part A: General Test Information

Part B: Statistical Analysis & Evidence of Validity

Publication No. AAR/OPI/ACE/I—2020-001

ACTFL
Language Connects

ACTFL ORAL PROFICIENCY INTERVIEW

Part A: General Test Information

Adrian Massei, Ph.D.
Department of Modern Languages,
Furman University

Publication No. AAR/OPI/ACE/I—2020-001

Table of Contents

Table of Figures

Rationale and purpose of the test

The ACTFL Oral Proficiency Interview (OPI) is a standardized, global assessment of functional speaking ability that assesses performance in a range of language tasks according to a set of fixed criteria. This American Council on the Teaching of Foreign Languages (ACTFL) test is delivered by Language Testing International (LTI), the exclusive licensee of ACTFL assessments. It is currently being used by hundreds of institutions of higher learning, government agencies, corporations, and state licensing boards for various purposes, including the credentialing of individuals for work and study opportunities as well as the evaluation and articulation of world language curricula.

The ACTFL OPI® takes the form of a live conversation between an ACTFL-certified OPI® Tester and the individual whose language proficiency is being assessed. The test is a highly structured interview that lasts between twenty and thirty minutes and can be conducted face-to-face or over the telephone. It represents a real-life exchange and is designed to gather a sample of the examinee's functional speaking abilities in the target language. The examinee's performance is then assessed vis-à-vis the ACTFL Proficiency Guidelines and the ACTFL Rating Scale, the rating criteria for the ACTFL OPI®.

The ACTFL Proficiency Guidelines are descriptions of what speakers of a language can and cannot do with regard to the particular linguistic functions they are able to perform, the types of content and contexts in which they are able to function, the way they deliver the message in terms of accuracy and comprehensibility, and the predominant text type they use when organizing oral discourse. The guidelines identify five major levels of proficiency: Distinguished, Superior, Advanced, Intermediate, and Novice; the last three are divided into three sublevels (High, Mid, and Low).

Central to the ACTFL OPI® is the notion of language proficiency understood as the ability to use a language to communicate meaningful information in spontaneous interactions, and in a manner that is deemed acceptable and appropriate to other speakers of the language with whom they are communicating in a particular context. In this sense, the test is not focused on what the examinee knows about the language (a particular set of predetermined content, discrete vocabulary items or grammar rules, for example) or dependent on how the language was acquired or learned, but rather on what the speaker can and cannot consistently do with it in unrehearsed exchanges. Therefore, a proficiency rating is an indicator of current ability to function in a language in real-life situations. The purpose of the ACTFL OPI® is to give the examinee an opportunity to demonstrate a functional floor (the examinee's highest level of consistent performance) and a functional ceiling (the level at which the examinee can no longer function consistently) across a variety of topics. At the floor, the speaker shows consistent ability to sustain all the criteria associated with the required functions of a particular proficiency level. At the ceiling, the speaker shows signs of linguistic breakdown manifested as a failure to sustain all the criteria required to perform the functions of that next proficiency level. To have a ratable sample, an OPI tester must identify and elicit clear evidence of an examinee’s floor and ceiling across a variety of topics, in turn, ensuring the validity of the test.

In obtaining the evidence of sustained performance and linguistic breakdown, the ACTFL-certified OPI® tester adheres to a standardized structure and protocol, alternating questions that target the floor and the ceiling. As the ACTFL OPI® is adaptive in the sense that the test is personalized to both the examinee’s interests and experiences as well as to the evidence resulting from performance in the different functions as the test unfolds, there is no set of pre-established questions that the tester asks in each interview. Instead, by listening attentively to performance, the tester crafts a unique experience for every examinee and presents certain question types targeting the particular functions that need to be explored in order to demonstrate what the candidate can and cannot consistently do with the language from a proficiency perspective. The sample thus obtained is compared with the ACTFL Proficiency Guidelines in order to determine what major level and sublevel best describe the examinee’s performance in terms of patterns of strengths and weaknesses. The sample is evaluated against a set of fixed criteria (the *ACTFL Proficiency Guidelines*) and not against a specific set of learned skills and/or content (achievement tests) or relative to other people’s performance (norm-referenced tests). In this sense, the ACTFL OPI® is a holistic assessment based on standardized criteria (i.e. it is a criterion-referenced test), since it measures a person’s ability to function in a given language according to a set of criteria.

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

Authorship of the ACTFL OPI® is attributed to the American Council on the Teaching of Foreign Languages (ACTFL). The design of a reliable, standardized procedure for the global assessment of functional speaking ability in a language has been a progressive effort of development, revision, and refinement that can be traced back to the 1950s. According to the Interagency Language Roundtable (no date), Language proficiency assessment was born out of necessity in the context of the Korean War (1950-1953). In 1952, the Civil Service Commission was assigned the responsibility of creating a register of government employees’ language skills, background, and experience. The lack of a language proficiency assessment in place and of a set of criteria for test construction led the commission to recommend the creation of a system that was objective, applicable to all languages, and independent of any particular curriculum.

The Foreign Service Institute (FSI) took on the responsibility of developing the first language proficiency scale under the leadership of Dr. Henry Lee Smith. The scale identified six levels of proficiency, ranging from “no ability” to “native speaking ability” and was first used in 1956. In 1958, the FSI created an independent testing office headed by Frank Rice and Claudia Wilds and, since then, all Foreign Service officers have been required to take language proficiency tests. For many years, the assessment criteria and language proficiency test developed by the FSI were used by several other government agencies and bodies, including the Central Intelligence Agency, the Defense Language Institute, and the

Peace Corps. In the late 1960s, a joint effort from several government agencies led to a document that offered descriptions of the base levels in four skills – speaking, listening, reading, and writing. This document would become the basis for the Interagency Language Roundtable Language (ILR) Skill Level Descriptions-Speaking, which provided the foundation for the ILR OPI, as well as assessments in other skills. The ILR scale and instruments based on it are still widely used today by the US government.

In the early 1980s, the American Council on the Teaching of Foreign Languages (ACTFL) would be the national organization that undertook adapting the ILR rating scale to make it more suitable to the academic domain, proposing finer gradations of proficiency at the lower levels. The result of this adaptation was the publication of the *Provisional Oral Proficiency Guidelines* in 1982, followed by the first official *ACTFL Proficiency Guidelines* three years later. In 1989, ACTFL OPI® Tester Certification was made available and the first *Oral Proficiency Interview Tester Training Manual* was published (co-authored by Dr. Pardee Lowe, Jr. and Dr. Judith Liskin-Gasparro).

Today, ACTFL is supported by a group of active and involved 66 master testers who teach OPI certification workshops and mentor candidates through the certification process. Their names and affiliations can be found in Appendix A. This group of experienced mentors work with ACTFL staff to contribute to the monitoring and refinement of ACTFL OPI functions, tasks, and protocols along with participating in the development of new OPI functions and tasks.

Types of scores reported to examinees

The scores reported to examinees follow the *ACTFL Proficiency Guidelines 2012 – Speaking*, which describe language proficiency along a continuum from the very top (highly articulate users of the language) to the very bottom (little or no functional ability) of the scale. In this continuum, the ACTFL guidelines identify and describe five major levels of proficiency: Distinguished, Superior, Advanced, Intermediate, and Novice; the last three are divided into three sublevels (High, Mid, and Low). The current ACTFL OPI® only tests through Superior (general professional proficiency), and this is the highest rating a test taker may receive, even if his or her performance exceeds the criteria for Superior. Therefore, the full range of possible scores reported to examinees includes: Superior, Advanced High, Advanced Mid, Advanced Low, Intermediate High, Intermediate Mid, Intermediate Low, Novice High, Novice Mid, and Novice Low.

The ACTFL OPI® rating scale assumes that proficiency in the language increases exponentially within the various global functions and throughout a hierarchy of those functions, rather than growing linearly in an additive fashion. Each of the levels encompasses a range of performance that grows in various ways as the level increases; partly for this reason, the scale assumes that language use is best assessed holistically by identifying patterns of strengths and weaknesses from the standpoint of overall performance.

This ACTFL rating scale can best be represented by an inverted pyramid, where the solid lines separate the major levels and the dotted lines indicate the sublevels within the Novice, Intermediate, and Advanced levels, as shown below:

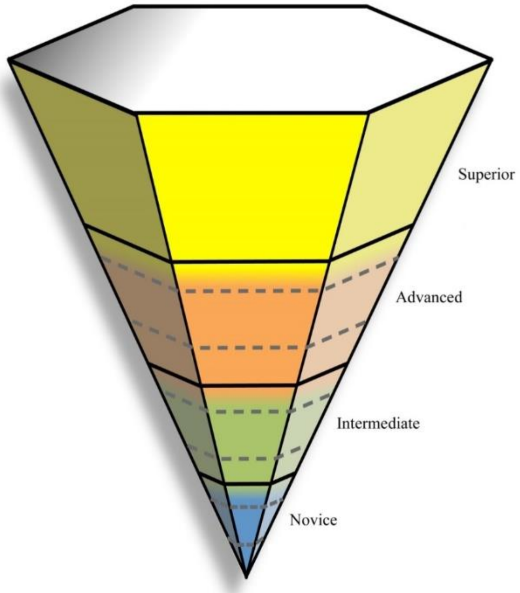

Figure 1: Inverted pyramid

The ACTFL rating scale is based on a hierarchy of global functions, where each major level (Novice, Intermediate, Advanced, and Superior) is defined by a set of linguistic functions that a speaker in that proficiency range is able to perform. These functions increase in complexity as the scale ascends from one level to the next. A rating at any major level subsumes the criteria of the levels below it and requires sustained performance of all the global functions associated with it. However, because each of these major levels is conceived as a range, two speakers with markedly different language performance may still be rated within the same major level. The sublevels (Low, Mid, and High) reflect the quantity and quality with which speakers perform the various non-compensatory global tasks or functions associated with the given major level. At the *Low* sublevel, the speaker functions barely above the major border or threshold; performance is skeletal, fluency and accuracy are typically limited, and the delivery often includes repeated instances of lexical confusion and self-correction. Performance at the *Mid* sublevel is robust; the speaker displays quantity, quality, flow, and solid control of the functions within the level; although not required, a *Mid* speaker may even start to show some features of the next higher level. At the *High* sublevel, the speaker is very solid at the major level in terms of quantity and quality of the speech produced and, in addition to that, there is substantial evidence of performance at the next higher level; in fact, speakers at the *High* sublevel are capable of functioning much of the time at the next higher level but are unable to sustain language consistently at that level; because of this, the dynamic of the “*High*” is best understood in a top-down representation of proficiency: it describes a fall from the next higher level (ceiling) rather than just a strong ability demonstrated at the base level (floor).

Directions for scoring and procedures and keys

Before starting the assessment, the ACTFL-certified tester reads an introduction to the examinee in English. This statement briefly describes the interview and encourages the test taker to participate actively in the conversation in order to show his/her language ability at its best. After clarifying any questions the examinee may have, the tester begins the oral assessment in the target language with a brief warm-up to get to know the examinee a bit so as to pursue topics of interest and to determine where to begin a “working level.” Next the tester presents a series of level checks, i.e. questions targeting the examinee’s level of sustained performance (floor), alternating them with probes or questions aimed at the next major level to demonstrate the level at which the candidate can no longer sustain performance consistently (ceiling). This is done across a variety of topics. The tester adjusts the “working floor” as appropriate, based on the examinee’s performance in the level checks and probes. The interview is digitally recorded and stored in a secure Internet-based archive so it can be accessed for rating purposes. The tester assigns a first rating, which is followed by an independent, blind second rating provided by another certified tester who accesses the data base. Both ratings must agree fully in the major level and sublevel in order for an official ACTFL rating to be issued. In the case of a discrepancy between the first and second ratings, the sample is arbitrated by a third certified tester and a final certified rating is issued when two ratings agree exactly.

The rating process of the ACTFL OPI® is based on a holistic approach and addresses a number of abilities simultaneously and analyzes them from a global perspective rather than from the point of view of the presence or absence of discrete linguistic features. In evaluating the speaker’s performance, the following assessment criteria are considered:

Proficiency Level*	Global Tasks and Functions	Context / *Content*	Accuracy	Text Type
Superior	Discuss topics extensively, support opinions, hypothesize. Deal with a linguistically unfamiliar situation.	Most formal and informal settings from concreate to abstract perspectives. *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 from the message.	Extended discourse
Advanced	Narrate and describe in major time frames and deal effectively with an unanticipated complication.	Some informal settings and a limited number of transactional situations. *Predictable, familiar topics related to daily activities.*	Understood without difficulty by speakers unaccustomed to interacting with language learners	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 interacting with language learners.	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 interacting with language learners.	Individual words and phrases

Figure 2: OPI rating criteria chart

*A rating at any major level is arrived at by the **sustained performance** of the functions of the level, within the contexts and content areas for that level, with the degree of accuracy described for the level, and in the text type for the level. The performance must be sustained across **ALL** of the criteria for the level in order to be rated at that level.

Global tasks or functions

These refer to what speakers can do with the language. At each proficiency level, there are specific functions that the speakers must be able to perform consistently in order to be considered proficient at that level (non-compensatory core requirements). At the Superior level, for example, a speaker needs to show consistent ability to elaborate abstractly on issues, provide structured arguments to support opinions, and hypothesize to explore possible outcomes or consequences, all beyond the realm of personal experience. Speakers lacking these functional abilities will not be at the Superior level, even if they demonstrate unusually strong fluency, lexical or structural control of the language.

Contexts/Contents

Context refers to the socio-cultural settings in which speakers can be expected to carry out communicative tasks. At the lower levels (Novice and Intermediate), these are limited to informal and familiar settings related to self and daily life that usually have a lot of concrete support from personal experience. At the higher levels, the number and type of possible situations in which a speaker can function expands and begins to encompass both informal and some formal settings at Advanced, such as workplaces, and most formal settings, such as public events at Superior. These settings clearly require increasing control and flexibility in the use of linguistic and rhetorical resources.

Being the most variable element of the OPI, content depends, to a large extent, on the speaker's interests and experiences. At the lower levels, content revolves primarily around self, personal experience, and daily life. At the higher levels, content transcends personal experience and expands to the community (Advanced) and the realm of ideas (Superior).

Accuracy/Comprehensibility

Accuracy and comprehensibility refer to the acceptability, quality, and precision of the message conveyed, as well as to the type of interlocutor needed to understand the message and engage in a meaningful exchange. It includes features such as fluency, grammar, pragmatic competence, pronunciation, sociolinguistic competence, and vocabulary. The degree of a speaker's control of these features when communicating may require different types of interlocutors. At the lower levels, for example, communication is made possible by a sympathetic listener, i.e. a listener who can mentally compensate for the gaps that occur in communication due to the speaker's limitations in some or all of the accuracy features. The type of interlocutor expected at the Advanced level is no longer sympathetic but rather neutral, one who is unaccustomed to dealing with learners or low-level speakers. At the Superior level, the speaker's control of the language should be such that the listener is not distracted from the message due to linguistic imperfections in accuracy and delivery.

Text type

Text type refers to the speaker's predominant way of organizing oral discourse. While Novice-level speakers tend to communicate minimally using memorized language such as isolated words, lists, phrases and sentence fragments, Intermediate-level speakers' predominant text type is discrete sentences and strings of sentences. At the Advanced level, speakers can organize speech in oral paragraphs when required by the function (i.e. a group of sentences that are sequenced strategically to convey one organized and cohesive message), while Superior-level speakers can communicate in extended discourse when performing Superior-level functions (i.e. a series of paragraphs tied together through cohesive devices to convey ideas, arguments, opinions, positions, etc.).

Cut scores

The OPI® does not have numeric cut scores. The OPI 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 OPI is a proficiency-oriented assessment with no recommended cut scores. That is, the OPI 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/OPIc
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.

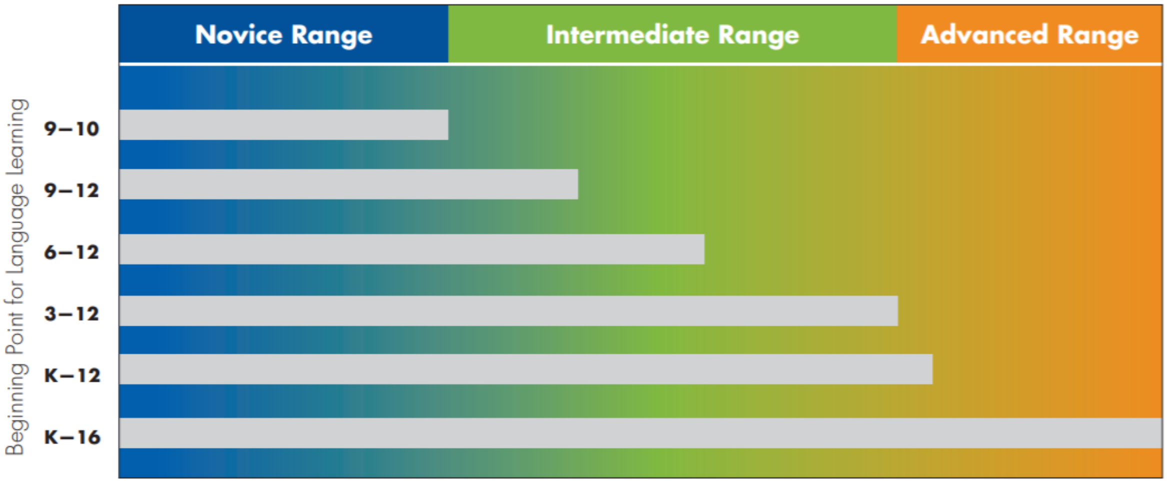

Figure 3: Time as a critical component for developing language performance

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 3).

Equivalence of forms

Each examinee receives a unique set of questions from the tester based on responses during the Warm-Up. As such, each OPI is meant to be a unique experience for each test taker.

OPI testers follow a function-based protocol, using topics identified during the warm-up. This allows for a standardized approach to the assessment such that the content of the Interview along with tasks used to convey the functions differ from examinee to examinee; however, the functions for which they must demonstrate a sustained ability to communicate remain the same. Adherence to the function-based protocol along with adherence to rating according to the ACTFL proficiency descriptors allow for equivalence between interviews.

Information on norms and normative groups (if appropriate)

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

Information about item/test content development

Given the adaptive nature of the ACTFL OPI® in terms of the topics explored during the interview, the test does not focus on any particular set of content items that need to be covered; rather, topics stem from the actual interaction between the examinee and the ACTFL-certified tester. The focus of the ACTFL OPI® is on functional ability and, because of that, it unfolds as a task-based test that follows a standardized structure and protocol. Each level in the ACTFL rating scale has a series of non-compensatory communicative tasks or functions that the speaker needs to show consistent ability to perform across a variety of topics. As explained above, the higher the level, the more complex these communicative functions become in terms of the linguistic resources needed to sustain performance.

Novice speakers do not have much functional ability in the language; they communicate minimally with formulaic and rote utterances, lists and phrases. At the **Intermediate** level, speakers become more interactive, and they now show the ability to create with the language spontaneously by combining and recombining learned material to convey new meaning about self and daily life; they also demonstrate the consistent ability to ask and answer simple questions on familiar topics, and handle a simple transactional situation. **Advanced**-level speakers demonstrate the consistent ability to narrate and describe in the major time frames of the language, controlling all the linguistic features associated with these functions; they also show the ability to move beyond personal experience and engage in conversations about current events or topics of interest to the community; finally, Advanced-level speakers demonstrate the ability to handle a situation with a complication. **Superior**-level speakers show consistent ability to discuss topics and issues both concretely and abstractly, give and support opinions on issues through structured argumentation, and hypothesize by elaborating on possible outcomes or consequences given certain circumstances. Finally, they show the ability to handle situations or topics that are not necessarily familiar to them from the linguistic standpoint by using strategies such as circumlocution.

ACTFL-certified testers are thoroughly prepared to ask questions purposefully in order to elicit the particular functions associated with each level. This is done following a standardized structure consisting of the four phases shown below:

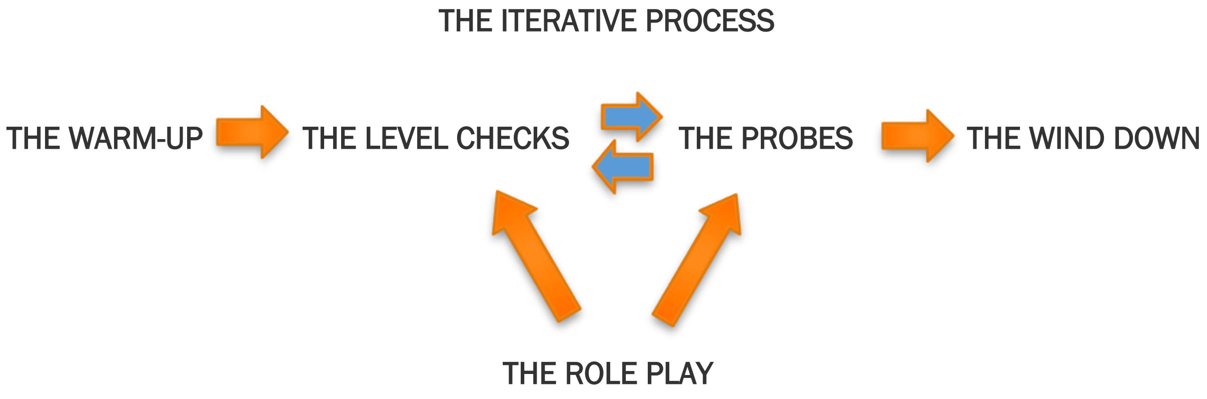

Figure 4: Phases of the OPI

Phase I: The warm-up

During the first four to five minutes, the tester uses conversation openers and open-ended questions that invite the examinee to share general information about self, such as occupation, work and/or educational background, place of residence, travel, leisure activities, etc. This helps the tester gather topics to develop later during the interview. It also helps the tester gather preliminary linguistic evidence leading to the initial working level, that is, the types of questions that the tester will ask to start eliciting the examinee's floor in the next phase of the interview.

Phase II: The level checks

These are questions targeting the functions and content areas that the speaker can handle most comfortably, demonstrating the ability to sustain the assessment criteria while doing so. They establish the speaker's floor or level of consistent performance.

Phase III: The probes

These are questions targeting the functions and content areas of the next higher major level that result in linguistic breakdown. They establish the ceiling or level where performance is no longer consistent, and the assessment features associated with that level are no longer sustained.

Rather than offering level checks and probes consecutively in a linear fashion, the tester selects a topic, develops it within the level by exploring functions at the floor and then spirals it up by exploring functions at the next higher level. Once the ceiling is established, the tester returns to the floor and repeats the same process on different topics throughout the interview until patterns of strengths and weaknesses are clearly demonstrated. This whole process of moving back and forth between level checks and probes is known as the iterative process and constitutes the main body of the test. As part of this dynamic, in the last third of the interview, the tester introduces a role-play, a required component of the ACTFL OPI® between Novice High and Advanced Mid. A role-play may be used as a level check or a probe as needed in order to confirm if the examinee can carry out functions that cannot be elicited by means of a conversational exchange (i.e. asking questions or handling a simple transaction or a situation with a complication).

Phase IV: The wind down

This is the last phase of the ACTFL OPI®. It signals the end of the interview and allows the examinee to regain a comfortable level to leave the interview on a positive note.

Statement of test's emphasis on each of the content, skill, and/or ability areas

The ACTFL OPI® is a global assessment of functional speaking ability and, as such, it does not assess any particular content or curricula. As explained above, speaking skills are evaluated holistically according to the ACTFL Proficiency Guidelines and based on the four major assessment criteria therein: global tasks or functions, context/content, accuracy/comprehensibility, and text type. As one moves up in the rating scale, the range of performance grows exponentially in terms of the global functions that the speaker is able to carry out, the types of settings and the variety of topic areas in which the language can be used effectively, the level of precision and comprehensibility of the message, and the text types in which discourse is organized.

Rationale for the kinds of tasks (items) included in the test/Information about why each task (item) is included

From the ACTFL OPI® perspective, the tester's ability to elicit a ratable sample is necessarily connected to effective elicitation; therefore, asking questions purposefully (with a clear functional target) is key to the assessment. As the tester adapts to the examinee's experiences and interests, each OPI is a unique experience. No two interviews are the same. Based on the information provided by the examinee, the tester develops motivating topics during the iterative process in order to confirm and reconfirm the floor and the ceiling. Rather than a list of questions, the interview is an interactive negotiation of meaning, whereby the tester poses question types depending on the functional role these questions target in the overall interview. The tester chooses questions based upon the evidence of sustained performance or linguistic breakdown provided by the examinee as the exchange unfolds. Once the floor and ceiling are established, the questions span two contiguous major levels across a variety of topics, confirming that the examinee has been given the opportunity to demonstrate their highest level of proficiency. The tester never goes below the floor in the iterative process and never probes two levels above it.

In order to elicit the appropriate functions, the questions asked need to clearly invite responses that naturally demonstrate those functions. At the Novice level, for example, questions target the production of memorized and formulaic languages such as lists and set expressions. At Intermediate, open-ended requests invite the examinee to create with the language spontaneously to have simple conversations. At Advanced, emphasis on detail and follow up questions elicit the narrations and descriptions expected at this level. At Superior, the tester uses prelude questions that model the type of language expected and that invite the examinee to demonstrate the functions associated with the level, such as abstract elaborations, argumentation, and speculation about issues.

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

Since the ACTFL OPI® is not a quantitative test, there is no pre-established number of tasks that might be considered appropriate for assessing functional spoken ability in a language. A ratable sample is obtained when the tester does the following:

1. Proves the floor by confirming and reconfirming the level of sustained functional performance via multiple level checks across a variety of topics.
2. Proves the ceiling by confirming and reconfirming the level at which the examinee can no longer function consistently via multiple probes across a variety of topics.
3. Completes all the phases of the ACTFL OPI®.
4. Adheres to the structure and elicitation protocols of the ACTFL OPI®.

The types of global functions and tasks elicited by the tester depend on the examinee's performance. If the interviewee establishes a floor at Novice level, for instance, the tester proves the level of sustained performance by eliciting memorized and formulaic language in different content areas related to self and daily life (i.e. food, clothing items, rooms in the house, colors, days of the week, etc.). In order to confirm and reconfirm the ceiling, the tester asks open-ended questions inviting creation with the language in sentence-level discourse. The tester does not explore tasks at the higher levels (descriptions, narrations, abstract topics, argumentation, etc.) since the tester only probes one level above the floor.

Information on the currency and representativeness of the test's tasks (items)

The ACTFL OPI® measures real life spoken language ability. Topics stem from the examinee's interests and experiences and are developed by the tester, who presents questions targeting different functions in order to prove the examinee's floor and ceiling. Virtually any topic may be developed at any level. Once it has been successfully addressed within a level, it can be spiraled up to the next higher level by changing the functions. The topic "food," for example, can be developed at Novice level (*List your favorite foods*), Intermediate level (*Tell me about meal time around your house?*), Advanced level (*Describe in detail the inside of your favorite Italian restaurant*) or Superior level (*Discuss the interconnection between eating habits and public health in the United States and the new trends observed in this sphere*).

Description of the item sensitivity panel review

ACTFL-certified testers complete a rigorous certification and norming process during which they are instructed on topics to avoid. Age, sex and sexual orientation, race, color, religion, national origin, marital status, health, and political viewpoints are to be avoided. For assessments delivered to commercial clients, many of these topics are prohibited by law. Testers are also discouraged from pursuing other highly controversial topics (even if volunteered by the candidate), such as abortion, gun control, immigration laws, corporal or capital punishment, war, etc. Testers are also given examples of topics they can explore during the interview.

Testers are instructed to adhere strictly to these guidelines when interviewing so that the purpose of the OPI is not compromised by the introduction of one or more topics that may make the examinee feel uneasy and, consequently, affect their performance in the language negatively. The tester's role is to develop topics that are of interest to the interviewee so that they engage in the discussion spontaneously and naturally, thus showing their language ability at its best.

Finally, the introduction that the tester reads to the examinee before the interview states that the examinee may decline a topic introduced by the tester if it makes them feel uncomfortable or if they are not authorized to talk about it.

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

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

References

Alpine Testing Solutions. (2020a). *Examination of the ACTFL Oral Proficiency Interview® (OPI) in Korean, French, and Mandarin for the ACE Review - Part B: Statistical analysis & evidence of validity*. Orem, UT: Alpine Testing Solutions.

Cubbellotti, S. & Cox, T. (2015). *Examination evaluation of the ACTFL OPI® in French, Korean, Mandarin for the ACE Review* (ACTFL Publication No. AAR/OPI/ACE/R—2015-001) Alexandria, VA: ACTFL.

ILR (no date). *Interagency Language Roundtable: How did the language proficiency scale get started?* https://www.govtilr.org/Skills/IRL%20Scale%20History.htm

Lowe, P. & Liskin-Gasparro, J. (1989). *ACTFL Oral Proficiency Interview tester training manual. Yonkers, NY: ACTFL.*

Appendix A: Active OPI master testers/mentors

Name	Affiliation
Abboudy, Bahia	Canadian Foreign Service Institute
Abuhakema, Ghazi	College of Charleston
Adamowicz-Hariasz, Maria	The University of Akron
Akiyama, Kathy	Mount Angel Seminary
Alosh, Mahdi	King Saud University
Ao, Qun	United States Military Academy
Baumann, Catherine	University of Chicago
Bedi, Susham	Columbia University
Breiner-Sanders, Karen	Georgetown University
Cassidy, Jim	Mt. Angel Seminary
Castro, Percio	University of Dayton
Chi, Richard	University of Utah
Cowles, Maria Antonia	University of Pennsylvania
Cox, Troy	Brigham Young University
Darhower, Mark	North Carolina State University
Demko, Anthony	International Education Center
Dhonau, Stephanie	University of Arkansas
DiBase-Lubrano, MaryJo	Yale University
Dolgova, Irina	Yale University
Hiple, David	University of Hawaii
Jacobe, Monica	The College of New Jersey
Jen, Theresa	University of Pennsylvania
Kamada, Osamu	Nanzan University
Kang, Sahie	Asian School
Kartchner, Eric	Georgia Southern University
Kim, Hee-Sun	Stanford University
Kong, Mei	University of Maryland
Lacorte, Manel	University of Maryland
Laughlin, Lizette	University of South Carolina
Laviosa, Flavia	Wellesley College
Lavie, Rena	Brandeis University
Lindsay, Deborah	South Albany School
Lindseth, Martina	University of Wisconsin-Eau Claire
Lipton, Shlomit	Hebrew at the Center
Liskin-Gasparro, Judith	University of Iowa
Maurer, Virginia	Harvard University
Martin, Cynthia	University of Maryland
Makino, Seiichi	Princeton University
Massei, Adrian	Furman University

McLaughlin, Suzanne	Chemeketa Community
Mir, Montserrat	Illinois State University
Morris, Daniel	Southern Oregon University
Moussa, Nawal	Department of National Defence and the Canadian
Milano, Ali	Stanford University
Miura, Ken-Ichi	Franklin and Marshall College
Otto Jr., Karl	Northwestern University
Overesch-Meister, Lynne	Johnson County Community College
Peyman, Nojoumian	University of Southern California
Prince, Bill	Furman University
Rivera-Martinez, Mildred	Emeritus, Peace Corp
Rockaitis, Ryan	Deerfield School
Rubio, Fernando	University of Utah
Ringvald, Vardit	Brandeis University
Saito, Mariko	Bunka Women’s University
Shankar, Jishnu	University of Texas Austin
Stever, Mari	Yale University
Swender, Jennifer	Independent Consultant
Shimada, Kazuko	East West Japanese Language Institute
Thompson, Chantal	Brigham Young University
Tschirner, Erwin	Herder Institute, Leipzig University
Viana da Silva, Eduardo	University of Washington
Watanabe, Suwako	Portland State University
Weissenrieder, Maureen	Ohio University
Winkler, Helga	Moorpark College
Wilkins, Jim	Lee University
Zhang. Yongfang	Wofford College

Examination Evaluation of the ACTFL Oral Proficiency Interview® (OPI) in Korean, French, and Mandarin 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

This document is structured to parallel the ACE Examination Checklist, addressing the following topics: scoring, statistical performance, and validity evidence.
This report documents the American Council on the Teaching of Foreign Languages (ACTFL) Oral Proficiency Interview (OPI®) from 2016 to 2020 to satisfy a review requirement of the American Council of Education (ACE) College Credit Recommendation Service (CREDIT) program. The ACTFL OPI® is an assessment of functional speaking proficiency in a foreign language that is evaluated by trained and certified experts in an interview format across numerous languages.

Inter-rater reliability and rater agreement were analyzed for three languages of the ACTFL OPI: French, Korean, and Mandarin. 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 OPI surpassed the minimum inter-rater reliability and agreement requirements. All language exam scores were in agreement within one sublevel over 97% 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 0.91- 0.97 across languages. Areas for improvement include a focus to the absolute agreement between raters within the Intermediate High (IH) and Advanced Mid (AM) borders. 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 "sustained functional ability, that is, the level at which speakers show full control over the functions," which means a single holistic score is assigned for the whole exam (see *ACTFL OPI Examinee Handbook*, page 16). 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 (French, Korean, and Mandarin) 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
French	553	455	511	464	82	2065
Korean	141	205	255	277	37	915
Mandarin	899	934	1013	1086	105	4037

*French and Korean data collected through March 27, 2020; Mandarin data collected through March 26, 2020.

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

	N	2 Ratings	3 Ratings	4 Ratings
French	2065	68%	31%	1%
Korean	915	59%	41%	1%
Mandarin	4037	72%	28%	1%

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

Table 3. French OPI: Rating 1 and Rating 2 Agreement from 2016-2020 (N = 2065)

		Rating 1*									
		NL	**NM**	**NH**	**IL**	**IM**	**IH**	**AL**	**AM**	**AH**	**S**
Rating 2*	**NL**	0	0	0	0	0	0	0	0	0	0
	NM	0	5	1	1	0	0	0	0	0	0
	NH	0	4	12	6	1	0	0	0	0	0
	IL	0	0	14	38	32	0	0	0	0	0
	IM	0	0	3	20	179	31	2	0	0	0
	IH	0	0	0	0	8	234	80	2	1	0
	AL	0	0	0	0	1	57	319	86	2	1
	AM	0	0	0	0	0	4	124	262	24	3
	AH	0	0	0	0	0	0	5	29	140	55
	S	0	0	0	0	0	0	1	4	59	215

*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. Korean OPI: Rating 1 and Rating 2 Agreement from 2016-2020 (N = 915)

		Rating 1*									
		NL	**NM**	**NH**	**IL**	**IM**	**IH**	**AL**	**AM**	**AH**	**S**
Rating 2*	**NL**	11	8	0	0	0	0	0	0	0	0
	NM	6	31	22	2	0	0	0	0	0	0
	NH	0	13	75	31	3	1	0	0	0	0
	IL	0	1	17	87	38	0	0	0	0	0
	IM	0	0	1	25	80	33	1	0	0	0
	IH	0	0	0	2	22	63	9	1	0	0
	AL	0	0	0	0	2	13	41	21	1	0
	AM	0	0	0	0	0	3	26	38	16	4
	AH	0	0	0	0	0	0	2	13	27	28
	S	0	0	0	0	0	0	0	1	15	81

*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. Mandarin OPI: Rating 1 and Rating 2 Agreement from 2016-2020 (N = 4037)

		Rating 1*									
		NL	**NM**	**NH**	**IL**	**IM**	**IH**	**AL**	**AM**	**AH**	**S**
Rating 2*	**NL**	0	1	0	0	0	0	0	0	0	0
	NM	0	2	8	0	0	0	0	0	0	0
	NH	1	12	49	24	2	1	0	0	0	0
	IL	0	4	50	120	43	4	0	0	0	0
	IM	0	0	6	68	381	55	11	0	0	0
	IH	0	0	0	3	65	339	160	17	1	1
	AL	0	0	1	0	8	110	274	179	2	0
	AM	0	0	0	0	0	15	94	371	35	7
	AH	0	0	0	0	1	0	2	31	140	44
	S	0	0	0	0	0	0	1	2	89	1203

*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 68% of the time for the French exam, 58% for the Korean exam, and 71% for the Mandarin exam. All three were within one category of each other over 97% 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**
French	2065	68.0%	98.5%	99.9%
Korean	915	58.4%	97.3%	99.9%
Mandarin	4037	71.3%%	97.8%	99.9%

Table 7. Percent (N) of Exact Agreement

Language	**Rating**	**Rating**		
		2	**3**	**4**
French	**1**	68.0% (1404)	52.4% (346)	56.3% (9)
	2	–	39.1% (258)	12.5% (2)
	3	–	–	37.5% (6)
Korean	**1**	58.4% (534)	45.0% (170)	42.9% (3)
	2	–	40.5% (153)	28.6% (2)
	3	–	–	16.7% (1)
Mandarin	**1**	71.3% (2879)	43.4% (498)	40.6% (13)
	2	–	40.6% (466)	15.6% (5)
	3	–	–	37.5% (12)

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

		Rating		
Language	**Rating**	**2**	**3**	**4**
French	**1**	98.5% (2034)	96.8% (639)	81.3% (13)
	2	–	93.6% (618)	62.5% (10)
	3	–	–	93.8% (15)
Korean	**1**	97.3% (890)	92.6% (350)	85.7% (6)
	2		92.3% (349)	85.7% (6)
	3			100.0% (6)
Mandarin	**1**	97.8% (3947)	92.1% (1056)	81.3% (26)
	2		90% (1032)	78.1% (25)
	3			78.1% (25)

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

		Rating		
Language	**Rating**	**2**	**3**	**4**
French	**1**	99.9% (2062)	99.7% (658)	100% (16)
	2	–	99.8% (659)	100% (16)
	3	–	–	100% (16)
Korean	**1**	99.9% (914)	99.5% (376)	100% (7)
	2		100% (378)	85.7% (6)
	3			100% (6)
Mandarin	**1**	99.9% (4031)	99.8% (1145)	96.9% (31)
	2		99.6% (1142)	93.8% (30)
	3			96.9% (31)

The Spearman rank-order correlation (ρ) 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

Ratings Compared	Language	N	ρ	*p*-value
1 and 2	French	2065	0.931	< 0.001
1 and 2	Korean	915	0.958	< 0.001
1 and 2	Mandarin	4037	0.961	< 0.001
1 and 3	French	660	0.873	< 0.001
1 and 3	Korean	378	0.931	< 0.001
1 and 3	Mandarin	1147	0.867	< 0.001
2 and 3	French	660	0.845	< 0.001
2 and 3	Korean	378	0.929	< 0.001
2 and 3	Mandarin	1147	0.844	< 0.001

Table 11. Spearman's Correlations by Year

Language	Ratings	Year	N	ρ	*p*-value
French	1 and 2	2016	553	0.931	< 0.001
	1 and 2	2017	455	0.916	< 0.001
	1 and 2	2018	511	0.936	< 0.001
	1 and 2	2019	464	0.945	< 0.001
	1 and 2	2020	82	0.909	< 0.001
Korean	1 and 2	2016	141	0.951	< 0.001
	1 and 2	2017	205	0.957	< 0.001
	1 and 2	2018	255	0.955	< 0.001
	1 and 2	2019	277	0.967	< 0.001
	1 and 2	2020	37	0.917	< 0.001
Mandarin	1 and 2	2016	899	0.959	< 0.001
	1 and 2	2017	934	0.958	< 0.001
	1 and 2	2018	1013	0.954	< 0.001
	1 and 2	2019	1086	0.966	< 0.001
	1 and 2	2020	105	0.954	< 0.001

Table 11. Spearman's Correlations by Year

Language	Ratings	Year	N	ρ	*p*-value
French	1 and 3	2016	197	0.870	< 0.001
	1 and 3	2017	167	0.876	< 0.001
	1 and 3	2018	157	0.910	< 0.001
	1 and 3	2019	116	0.818	< 0.001
	1 and 3	2020	23	0.769	< 0.001
Korean	1 and 3	2016	59	0.908	< 0.001
	1 and 3	2017	95	0.928	< 0.001
	1 and 3	2018	106	0.933	< 0.001
	1 and 3	2019	98	0.938	< 0.001
	1 and 3	2020	20	0.945	< 0.001
Mandarin	1 and 3	2016	260	0.899	< 0.001
	1 and 3	2017	265	0.835	< 0.001
	1 and 3	2018	329	0.862	< 0.001
	1 and 3	2019	278	0.869	< 0.001
	1 and 3	2020	15	0.898	< 0.001
French	2 and 3	2016	197	0.869	< 0.001
	2 and 3	2017	167	0.828	< 0.001
	2 and 3	2018	157	0.877	< 0.001
	2 and 3	2019	116	0.793	< 0.001
	2 and 3	2020	23	0.728	< 0.001
Korean	2 and 3	2016	59	0.932	< 0.001
	2 and 3	2017	95	0.944	< 0.001
	2 and 3	2018	106	0.902	< 0.001
	2 and 3	2019	98	0.941	< 0.001
	2 and 3	2020	20	0.874	< 0.001
Mandarin	2 and 3	2016	260	0.837	< 0.001
	2 and 3	2017	265	0.793	< 0.001
	2 and 3	2018	329	0.870	< 0.001
	2 and 3	2019	278	0.854	< 0.001
	2 and 3	2020	15	0.838	< 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 above 0.909. The Ratings for the French exam were slightly lower than that of the Korean and Mandarin, but by a small amount. The correlations also decreased from 2019 to 2020 by a small amount, but they still maintained high agreement. This drop 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.

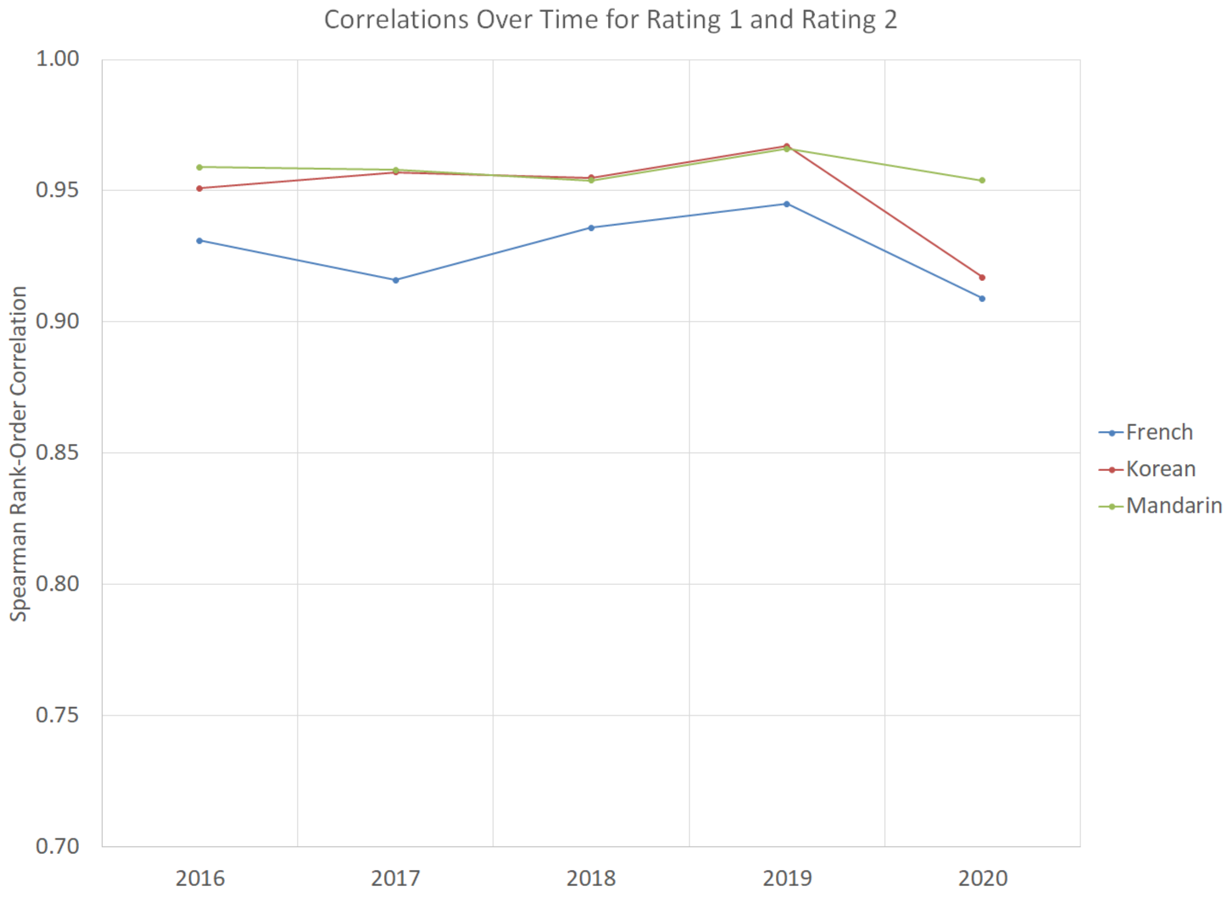

Figure 1. Spearman-rank correlations of Rating 1 and Rating 2 from 2016 to 2020 Score

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 French exam, the distribution of the final exam ratings were similar over time; however, the 2020 ratings were higher for the categories of “AL”, “AM”, and “S” then in any previous year and lower in the categories of “IH” and “IM”. The Korean exam had noticeably more “IL” ratings in 2020 than in the past. There were also more “AL” and “AM” ratings in 2020 compared to previous years and fewer “IH” ratings. Of the 37 examinees that completed the 2020 exam, none of them earned an “S” rating. For Mandarin, there was an unusually high percentage of “S” ratings among the 105 examinees completing the exam in 2020 and a low percentage of “IM” ratings. It is recommended that ACTFL review the high ratings for the Korean “IL” category and the Mandarin “S” category. As with the Spearman correlations shown in Figure 1, it is possible the shortened data collection for 2020 had some impact on these results.

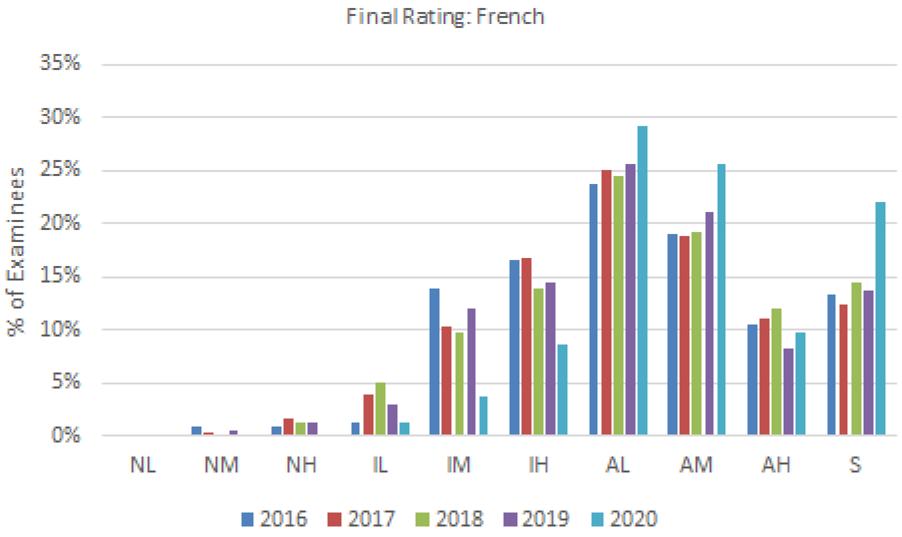

Figure 2. Final ratings from 2016 to 2020 for the French OPI

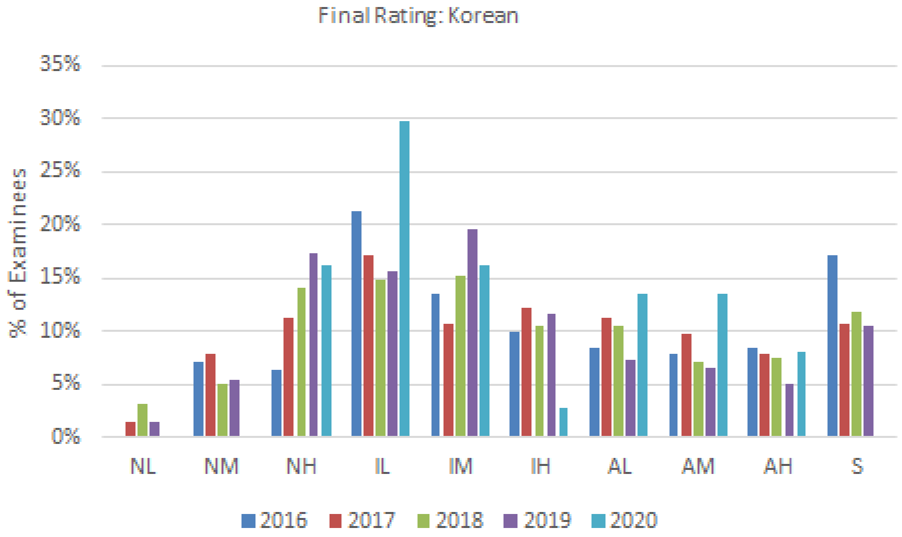

Figure 3. Final ratings from 2016 to 2020 for the Korean OPI

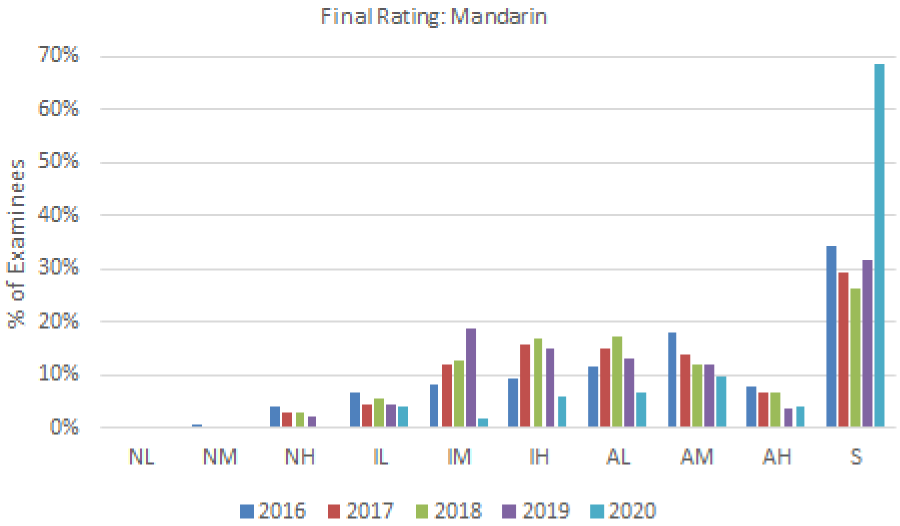

Figure 4. Final ratings from 2016 to 2020 for the Mandarin OPI

Evidence of Validity

Content Related

OPI Elicitation tasks used by the Tester during the Oral Proficiency Interview are standardized and are representative of the domain for which it is designed to measure-- language proficiency (speaking) as per the *ACTFL Proficiency Guidelines*. The Oral Proficiency Interview is a standardized procedure for the global assessment of functional speaking ability. As per the *ACTFL OPI Tester Training Manual* , to assess a test taker’s performance via a ratable sample of the language, “the OPI establishes 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*.

Criterion Related

Scores from the current OPI® have not been compared with any related measures of language performance that would allow for criterion-related validity evidence. The exam scores are used for a variety of purposes including language fluency certification, employment selection, placement, and college credit; therefore, standardized measures of later performance would be difficult to obtain. In addition, the OPI® 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 OPI® have not been compared with any related tests of language ability largely because the OPI® 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 Warm-Up allows the tester to avoid selecting 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 the OPI®, OPI testers are instructed to avoid sensitive topics (e.g., immigration, national origin, sexual preference, religion, marital status, racism) when developing OPI® prompts. The *ACTFL OPI Examinee Handbook* notes that the tester informs the test taker: "If you are uncomfortable with, or not authorized to speak about a topic that I may introduce, please let me know and we will discuss another topic." (p. 7). Further, as per ACTFL's policy on *Record Retention and Test Taker Confidentiality*, testers may not ask a candidate about:

- Age
- Sex / Gender Identity
- Race
- Color
- Religion
- National Origin
- Sexual Preference
- Marital Status
- Health
- Political Viewpoint

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 ACTFL OPI is not a timed assessment. The typical amount of time for an assessment depends on the examinee's language proficiency. The higher a test taker's proficiency, the more language they can produce. As such, those with higher proficiency levels will take more time with the assessment than those with lower proficiency levels. On average, an OPI takes between 15 and 30 minutes; however, a tester will not terminate the test should it take longer than expected. One caveat that should be noted is that testers are trained to attend to test-taker affect and fatigue. Since fatigue can prevent a test taker from demonstrating their actual proficiency, testers make every effort to approach the assessment with efficiency so that examinee affect or fatigue do not negatively affect the assessment outcome.

Provisions for Standardizing Administration of the Examination

The OPI test structure is governed by a detailed protocol and by the global language functions delineated by the criteria referenced by the Assessment i.e. the *ACTFL Proficiency Guidelines*. As per the *OPI Tester Training Manual* and the *OPI Examinee Handbook*, test administration begins with an introduction. Once the interview commences, the tester begins with a warm-up that allows the test taker to become accustomed to participating in an Oral Proficiency Interview. The Interview carries on by identifying a working level and eliciting language that demonstrates a test taker’s ability to consistently complete linguistic tasks which provide evidence of their proficiency. These tasks are designated by the protocol and are in-line with the functions that are identified within the *ACTFL Proficiency Guidelines.*

Irrelevant Sources of Difficulty Affecting Test Scores

A formal study of construct irrelevant variance for the OPI® 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 *OPI Tester Training Manual*, pages 15-19). The use of the Warm-Up to select prompts most likely to be familiar to the examinee may help to minimize context effects (see the *OPI Tester Training Manual*, pages 22, 26). As described above, administration procedures are standardized to make sure 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.

The ACTFL OPI® is a live interview with a certified tester. The tester has been trained to adapt and create appropriate test items based on the background and interests of the candidate. The interview is digitally recorded by the tester within the Language Testing International (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.

All official OPIs are proctored to make sure that candidates do not record the prompts they are given. As per ACTFL’s standard operating procedure document, proctors must apply and be accepted by the test administration office. They must sign an agreement verifying that they understand and can apply ACTFL proctoring protocols.

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 members are 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 members are 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. 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.

Finally, OPI tasks per language are retired based on their ability to elicit the targeted linguistic features (i.e. performance) and/or due to overexposure.

Interpretations and Conclusions

To conclude, the ACTFL OPI exceeded the minimum inter-rater reliability and agreement requirements. All language exam scores were in agreement within one sublevel over 97% of the time. Overall, the highest absolute agreement across languages was found at the highest level (i.e. Superior), and the lowest absolute agreement was found at the Intermediate High and Advanced Low border.

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 raising absolute rater agreement within the Intermediate High and Advanced Mid borders across languages. The results of this analysis confirm the reliability of the ACTFL OPI 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, 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 and Language Testing International (2019). *ACTFL OPI Examinee Handbook. https://www.languagetesting.com/pub/media/wysiwyg/manuals/opi-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.