Two Studies Investigating the Reliability and Validity of the English ACTFL OPIc® with Korean Test Takers

The ACTFL OPIc® Validation Project Technical Report

Updated March 23, 2008[1]

Authored by:
Eric A. Surface, PhD
Reanna M. Poncheri, MS
Kartik S. Bhavsar, MS
SWA Consulting
Raleigh, NC

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

Language Testing International
White Plains, NY

Abstract

Two studies were conducted with Korean test takers to assess initial psychometric evidence for the English ACTFL OPIc® as a measure of speaking proficiency in English. The initial study (Study 1) established evidence of reliability and validity for the assessment, specifically interrater reliability, test-retest reliability, and construct validity evidence. Study 1 led to several recommendations to improve the assessment, which were implemented by ACTFL. A smaller, second study (Study 2) was conducted after the modifications were implemented and yielded further evidence of the reliability and validity of the ACTFL OPIc®. Although they use different interview modalities, the results suggest both assessments measure the same construct, have similar reliabilities, and provide similar inferences. The findings from the two studies provide sufficient evidence to justify the initial use of the ACTFL OPIc® for commercial testing. However, ACTFL should maintain its commitment to using research to inform the test development and validation process as it extends the computerized interview format to other languages and test takers. This technical report describes both English studies and presents and discusses the results.

[1] The original version of this technical report covering Study 1 only was completed on March 17, 2006. Study 2 was added in 2007 to create this version. This version is being officially released on March 23, 2008.

Two Studies Investigating the Reliability and Validity of the English ACTFL OPIc® with Korean Test Takers

The ACTFL OPIc® Validation Project Technical Report

Overview

Factors, such as globalization and world political and military events, have increased the need for foreign language skills in business, governmental, and military organizations. Speaking is often the most commonly required language skill in these organizations. These unfulfilled foreign language skill requirements have led to an increased demand for hiring language qualified individuals and to a corresponding increase in language training and testing.

One of the most frequently used speaking proficiency assessment techniques involves an interviewer-based protocol that requires scheduling an interview between two individuals that may be located anywhere on the planet. Thus, increased demand for measuring speaking proficiency could create a testing capacity issue because of logistical constraints. As demand for speaking proficiency testing explodes, scheduling and conducting oral proficiency interviews (OPI) with human interviewers will need to be supplemented with new methods to meet the future testing needs of education, business, government, and military organizations.

In anticipation of the need to supplement existing testing capacity created by increased demand, the American Council on the Teaching of Foreign Languages (ACTFL) has developed an Internet-delivered, semi-direct version of its hallmark ACTFL OPI® assessment of speaking proficiency. This assessment elicits and collects a ratable sample of speech, eliminating the need for the interviewer and allowing the sample to be rated by certified raters located anywhere in the world. This Internet-based assessment of speaking proficiency is called the ACTFL OPIc® with the "c" representing the computerized nature of the assessment.

In accordance with the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999), test publishers must document the psychometric properties of their instruments by providing empirical evidence of reliability and validity. As a new assessment, evidence of the validity and reliability of the ACTFL OPIc® must be provided. Therefore, two studies were conducted as initial investigations of the psychometric properties of the English version of the assessment.

This technical report presents the results of two studies which examined the reliability and validity of the ACTFL OPIc® for the purpose of assessing the English speaking proficiency for a sample of Korean employees from the sponsoring organization. The findings of the two studies taken together provide evidence supporting the use of the ACTFL OPIc® as an assessment of speaking proficiency in English. The studies also demonstrate that ACTFL has made a serious commitment to a data-driven approach to test development and improvement.

The initial study—referred to as Study 1 throughout this document—was conducted with a sample of Korean employees and led to a number of recommendations for improving the assessment. These recommendations were instituted by ACTFL as part of their iterative improvement philosophy. A second smaller study—Study 2—was commissioned as follow up to implementing the recommendations. Results from both studies provide an initial assessment of the psychometric properties of the ACTFL OPIc® and provide support for its use.

These studies are in fulfillment of the requirements of the ACTFL OPIc® development

contract. The studies conform to current technical and professional standards and have been designed to provide the best quality data available given the constraints of the situation.

This technical report provides some basic background information, describes the study methods and results, and discusses the implications of our findings for the ACTFL OPIc® and future research. It should be noted the original version of this technical report (Surface, Poncheri, & Bhavsar, 2006) focused solely on Study 1, and this version is a revision to integrate Study 2. After a general discussion of the background issues, studies 1 and 2 are presented, followed by a general conclusions section. Any technical questions about the study should be addressed to Dr. Eric A. Surface (esurface@swa-consulting.com) at SWA Consulting Inc.

Background

Need for the ACTFL OPIc®

Although language skills have always been important, recent factors, such as globalization and world political and military events, have increased the need for these skills by education, business, government, and military organizations. As Swender (2003) states, "In today's workplace, many companies, agencies, corporations, and other institutions are experiencing ever-increasing demands to hire personnel with language skills" (p. 524). This increased demand to hire personnel with foreign language skills has led to an increase in language training.

Noe (2005) indicates that approximately 20 to 30 percent of US organizations have a budget allocated for foreign language training. Additionally, many US and foreign organizations devote resources to training English—typically called *English as a Second Language* (ESL) programs in the US. For example, multinational corporations and organizations that employ many non-English speaking workers have found language training to be a necessity (Weber, 2004). Wyndham Hotels recently implemented a self-guiding, voluntary English language program for its Spanish-speaking employees with the goal of boosting morale, employee retention, customer service, and promotion potential (Hammers, 2005).

Large companies have been moving jobs abroad, and the offshoring trend is expected to continue (Smith & Frangos, 2004). This movement has implications for language usage and assessment. Worker mobility and its language implications have also increased due to expatriate assignments by multinational corporations. Both within the U.S. and elsewhere, language skills can affect expatriate adjustment to the host country (Chao & Sun, 1997; Takeuchi, Yun, & Russell, 2002). Other salient examples of work situations where language skills are critical, such as call centers, abound.

Call centers are growing rapidly, both in terms of the number of people employed and the increasing size of the sector (Callaghan & Thompson, 2002). Many call centers employ personnel whose native language differs from the language spoken by the customers they serve. Many companies in the United States and the United Kingdom are outsourcing their call centers to countries such as India and the Phillipines (Pristin, 2003; Vina & Mudd, 2003).

A recent article reported that there are approximately 171,000 individuals working in call centers in India (Vina & Mudd, 2003). The same article reports that the United States and other countries have already lost roughly 400,000 back-office bank and low-level computer-coding jobs to India; this is likely to climb to 3.3 million by 2015. Offshoring seems to be a trend that will only increase (Chittum, 2004; Pristin, 2003). An article by Flynn (2003) addresses one effect of this trend – training workers abroad to be sensitive to the cultures of customers located in the United States and elsewhere. At the most basic level, this sensitivity requires language skills.

All these examples reinforce the importance of and need for work-related foreign language

proficiency. As the need for language skills has become more prevalent, the need to be able to measure language proficiency has become more important. Many of the jobs requiring language in education (e.g., teacher certification), business (e.g., call center personnel), government (e.g., diplomatic corps), and military (e.g., Special Forces Soldiers) organizations require speaking proficiency. The spoken communication requirements of these jobs will lead to increased demand for the assessment of speaking proficiency in foreign languages.

One of the most frequently used speaking proficiency assessment techniques involves an interviewer-based protocol that requires scheduling an interview between two individuals that may be located anywhere on the planet. Thus, increased demand for measuring speaking proficiency could create a testing capacity issue because of logistical constraints. As demand for speaking proficiency testing explodes, scheduling and conducting oral proficiency interviews (OPI) with human interviewers will need to be supplemented with new methods to meet the future testing needs of business, government, and military organizations.

A solution utilized by providers of paper-and-pencil surveys and assessments is to move them to the Internet. This allows for assessment at any place, at any time, and allows for the reduction of the resources required for testing. Examples of written measures that have shifted from paper to online formats include personality tests, situational judgment tests, cognitive ability tests, and surveys (Potosky & Bobko, 2004; Salgado & Moscoso, 2003; Thompson, Surface, Martin, & Sanders, 2003).

In anticipation of the need to supplement existing testing capacity created by increased demand, the American Council on the Teaching of Foreign Languages (ACTFL) has developed an Internet-delivered, semi-direct version of its hallmark ACTFL OPI® assessment of spoken proficiency. This assessment elicits and collects a ratable sample of speech, eliminating the need for the interviewer and allowing the sample to be rated by certified raters located anywhere in the world. This Internet-based assessment of speaking proficiency is called the ACTFL OPIc® with the "c" representing the computerized nature of the assessment.

Although there will always be a need for face-to-face and telephonic assessments of speaking proficiency, the ACTFL OPIc® is a large step toward increasing testing capacity and flexibility through use of computerized testing. Our studies were designed to assess the psychometric properties of this new assessment for speaking proficiency in English.

The Standards

The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) is the primary document that provides evaluative guidelines for the users, developers, and publishers of tests. According to the *Standards*, test publishers must document the psychometric properties of their instruments by providing empirical evidence of reliability and validity. The *Standard*s provide guidelines for presenting reliability and validity information about a test or other type of assessment.

A test refers to any "evaluative device or procedure in which a sample of an examinee's behavior in a specified domain [test content area] is obtained and subsequently evaluated and scored using a standardized process" (AERA, APA, & NCME, 1999; p. 3) and is not simply restricted to paper-and-pencil assessments. Therefore, these guidelines apply to the ACTFL OPIc®. As a new assessment, evidence of the validity and reliability of the ACTFL OPIc® must be provided for its intended use, the assessment of speaking proficiency.

Our studies were designed to ascertain empirical reliability and validity evidence for the ACTFL OPIc® within the constraints of the testing environment and to be consistent with accepted professional practices and technical standards specified in the *Standards* (AERA, APA, & NCME, 1999), the *Principles for the Validation and Use of Personnel Selection Procedures* (Society for Industrial and Organizational Psychology, 2003) and generally accepted,

reputable publications related to relevant psychological research (e.g., *International Journal of Selection and Assessment*) and methodology (e.g., *Psychological Methods*). Other sources on test development and validation were consulted as well (e.g., Downing & Haladyna, 2006).

Although the *Standards* document provides the primary source of guidance for evaluating the acceptability of testing and assessments, it is not by design a prescriptive document. As such, the *Standards* document acknowledges and advocates the use of and need for professional judgment that is based on knowledge of behavioral science and psychometrics to guide the evaluation of assessments. This professional judgment is specifically meant to encompass the choice of evaluation methodology deemed most appropriate for the specific testing context.

The *ACTFL OPIc® Validation Project* has been designed by qualified experts in the field of industrial/organizational psychology to ensure that the highest quality psychometric evidence is provided in accordance with the *Standards* given the constraints of the testing context. Since any assessment needs to demonstrate sufficient evidence of reliability and validity to be used, our studies were designed to focus on assessing these psychometric properties.

Reliability Evidence

Reliability as Consistency

Consistency, defined as the extent that separate measurements retain relative position, is the essential notion of classical reliability (Anastasi, 1988; Cattell, 1988; Feldt & Brennan, 1989; Flanagan, 1951; Stanley, 1971; Thorndike, 1951). Simply put, reliability is the extent to which an item, scale, procedure, or instrument will yield the same value when administered across different times, locations, or populations.

In the specific case of rating data, the focus of reliability estimation turns to the homogeneity of judgments given by the sample of raters. One of the most commonly used forms of rater reliability estimation is interrater reliability, which reflects the overall level of consistency among the sample of raters involved in a particular judgment process. When interrater reliability estimates are high, the interpretation suggests a large degree of consistency across sample raters. This is similar to the concept of internal consistency of items in a scale or test.

Another common approach to examining interrater consistency is to use measures of absolute agreement. Whereas interrater reliability estimates are parametric and correlational in nature, measures of agreement are non-parametric and assess the extent to which raters give concordant or discordant ratings to the same objects (e.g., interviewees). Technically speaking, measures of agreement are not indices of reliability *per se*, but are nevertheless quite useful in depicting levels of rater agreement and consistency of specific judgments, particularly when data can be considered ordinal or nominal. The ACTFL proficiency scale is ordinal.

Standards for absolute agreement vary depending on the number of raters involved in the rating process. When two raters are utilized, an absolute agreement of 80% or greater is generally considered to be excellent. Although absolute agreement closer to 100% is desired, a minimum of 70% is acceptable. However, in assessment development, rater training, or initial fielding contexts, lower agreement levels might be acceptable, depending on the circumstance and as long as agreement levels increased before high stakes use. Each additional rater (e.g., adding a third rater) employed in the process decreases the minimum acceptable agreement percentage. This recognizes that the agreement between more than two raters is increasingly difficult.

There can be a disconnection between rater agreement and interrater reliability. Interrater reliability can be high when the concordance of raters is lower than desired, especially if the disagreements are few, consistent in direction, and slight in terms of magnitude. Both interrater reliability and agreement provide useful information about the judgment of raters.

Reliability as Repeatability

In addition to consistency, reliability can also be defined in terms of repeatability of measurement or test-retest reliability. Repeatability is just as important for rater-based assessments as it is for multiple choice tests. An assessment should provide a stable measurement of a construct across multiple administrations, especially when the time interval in between the administrations limits the potential for the amount of the underlying construct to change. To function properly, assessments must yield highly consistent scores or ratings across measurements, within a time period when no acquisition or decay or change in the construct (i.e., English speaking proficiency in this case) can be expected. This means that the assessment process must function equivalently across multiple administrations as well as across raters. It also implies that raters at any point are functioning equivalently. Therefore, interrater reliability and test-retest reliability are both important. Correlation coefficients are usually used as test-retest reliability coefficients.

Importance of Reliability

Items, tests, raters, or procedures generating judgments must yield reliable measurements to be useful and have psychometric merit. Data that are unreliable are—by definition, unduly affected by error—and decisions based upon such data are likely to be quite tenuous at best and completely erroneous at worst. Although validity is considered the most important psychometric measurement property (AERA, APA, & NCME, 1999), the validity of an assessment is negated if the construct or content domain cannot be measured consistently. In this sense, reliability can be seen as creating a ceiling for validity.

The *Standards* provide a number of guidelines designed to help test users evaluate the reliability data provided by test publishers. According to the *Standards*, a test developer or distributor has the primary responsibility for obtaining and disseminating information about an assessment procedure's reliability. However, under some circumstances, the user must accept responsibility for documenting the reliability and validity in its local population. The level of reliability evidence that is necessary to assess and to be reported depends on the purpose of the test or assessment procedure. For example, if the assessment is used to make decisions that are "not easily reversed" or "high stakes" (e.g., employee selection or professional school admission), then "the need for a high degree of precision [in the reliability data reported] is much greater" (p. 30).

Study 1 was designed to assess the test-retest reliability (i.e., repeatability) of the ACTFL OPIc® across two administrations with the same individuals as well as the interrater reliability and consistency for the ACTFL OPIc® raters at both time points (two administrations of the ACTFL OPIc®). Study 2, because of testing constraints, focused on interrater reliability and agreement at a single administration only.

Validity Evidence

Validity is a unitary concept referring to "the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed purpose" (AERA, APA, & NCME., 1999, p.11). Validity is the most important psychometric property of any test and must be demonstrated through the accumulation of empirical, scientific evidence that scores can be appropriately interpreted and used for a specified purpose.

The *Standards* provide guidelines for assessing and reporting evidence of validity. Although there are five categories of validity evidence outlined in the *Standards*, the two categories—"evidence based on internal structure" and "evidence based on relations to other variables"—are the focus of this section of our technical report because they provide the basis for our validity examinations of the ACTFL OPIc®.

An additional type of evidence, "evidence based on test content," was addressed by language experts and test developers at ACTFL during the ACTFL OPIc® development and initial testing. This is what was formerly referred to as "content validity" evidence and is typically established by expert judgment and ensures test content

overlaps with the content domain of the construct in question. In other words, it provides evidence of the degree to which the content of the ACTFL OPIc® relates to the construct of speaking proficiency as defined by the *ACTFL Proficiency Guidelines – Speaking: Revised 1999* (Breiner-Sanders, Lowe, Miles, & Swender, 2000).

Evidence based on internal structure

This refers to a type of validity evidence that provides an evaluation of the degree to which items or ratings are related and are a representation of the construct in question. If the construct definition implies a single dimension of behavior, such as a specific type of proficiency, then the items or ratings measuring that construct should be related and fairly homogenous. According to the *Standards*, the analyses used to evaluate this type of evidence and their interpretation depends on the test and testing context.

For example, Confirmatory Factor Analysis (CFA) techniques can be used to confirm or disconfirm an a priori construct or set of relationships between constructs. CFA allows for the evaluation of alternative models to determine if the a priori model provides the best fit of the data. One limitation is that the data requirements for CFA are fairly strict in terms of the number of items per construct and number of cases needed for the analysis. Typically, to operationalize a construct in CFA, a minimum of four items or ratings are needed. However, it is possible to estimate "fit statistics" and parameter estimates for a CFA with only three ratings per construct, if multiple latent constructs are modeled.

In Study 1, an additional (third) rater was added to allow for the use of CFA. Each participant took both assessments and both assessments were rated by three raters. This study deviates from the two rater protocol solely to allow for the use of CFA for construct validity evidence. Therefore, the final ACTFL OPI® and ACTFL OPIc® proficiency ratings assigned to each test-taker were based on the agreement of at least two out of three testers, consistent with the typical process. The use of CFA allows for the estimation of model fit (allowing for comparisons between models), for the calculation of validity coefficients, and for the estimation of a latent correction between the two test constructs (i.e., ACTFL OPI® and ACTFL OPIc® in this case). CFA results provide a source of strong evidence to confirm or disconfirm the construct validity of a measure. In study 2, because of constraints, it was not practical to have three raters, preventing us from using CFA.

Evidence based on relations to other variables

The statistical relationship of a test to established measures of the same construct, related constructs, or different constructs can provide validity-related evidence (AERA, APA, & NCME, 1999). The relationship between scores on a test and scores on measures that assess the same or similar constructs provides convergent evidence of validity.

In this study, a strong correlation between the ACTFL OPI® and the ACTFL OPIc® would provide validity evidence because one would expect two language assessments of the same skill modality in the same language utilizing the same definition of the construct to be highly related. Therefore, whereas a strong relationship between the ACTFL OPI® and ACTFL OPIc® would provide strong validity evidence, the lack of a robust, statistically significant relationship with an established assessment of speaking proficiency (ACTFL OPI®) would raise questions about the validity of the ACTFL OPIc®.

The *Standards* suggests integrating various "strands" of empirical evidence and expert judgment into a coherent, sound validity argument to support the intended use of the assessment. Therefore, the purpose of our studies was to conduct initial investigations of the psychometric properties of the English version of the assessment to start the process of accumulating "strands" of empirical evidence.

Concordance[2] of Final Ratings

Since the ACTFL OPIc® is considered to be a different modality of the ACTFL OPI® and not a different assessment (i.e., same assessment different delivery method), both assessments should produce concordant final ratings in theory. In other words, since both versions use the same definition of speaking proficiency and rating protocol, the final ratings produced by the assessments should be the same or very similar. This is the same issue as face-to-face and telephonic interviews producing the same final ratings.

The issue according to the *Standards* is interchangeability of the test scores. "Support should be provided for any assertion that scores obtained using different items or testing materials, or different testing procedures, are interchangeable for some purpose" (AERA, APA, & NCME, 1999; p. 57). This includes alternative delivery modes of the same assessment (paper-and-pencil vs. web-based tests). For example, if a person is classified as an ENTJ on the paper-and-pencil MBTI®, then that person should be classified as an ENTJ on the web version as well as on the short form of the MBTI®. The *Standards* discusses the need to "equate" alternate forms of assessments.

Absolute agreement between final scores or final ratings is a high standard to achieve. An assessment can be valid without producing the same score or rating as another assessment of the same construct. This is possible because validity coefficients are correlational in nature and based on the degree of relationship between scores or ratings. Therefore, scores or ratings can be different in the absolute sense and still have a high degree of relationship, providing strong validity evidence, as long as the differences are consistent in direction and magnitude.

For example, when equating two forms of the same multiple choice assessment (parallel forms), it is possible that the two forms could yield highly correlated scores but never assign the same exact score to the same test taker. Although it is unlikely that the concordance of scores would be zero between two parallel forms, it highlights the point that exact agreement of scores needs to be considered when appropriate. It also suggests how difficult having 100% agreement would be in practice.

In other words, when two assessments have a high validity coefficient (i.e., high correlation between the two assessments), it means that the score or rating on one assessment can be consistently predicted from the score or rating on the other assessment. Therefore, 100% or high concordance between final ratings or scores is not strictly necessary to establish the validity of the new measure based on its relationship to the established measure of the construct. However, identical conceptual definitions and statistical evidence is required to demonstrate the relationship in order to "equate" the scores from the two assessments.

In the case of the ACTFL OPIc®, in addition to the validity evidence, the exact agreement of its final rating with the ACTFL OPI® final rating is also important because the assessments are designed to measure the same construct (same definition of speaking proficiency) using the same protocol and scale. Therefore, the two modalities should produce "identical" final ratings in theory.

In reality, achieving and maintaining 100% absolute agreement between two assessments is not a feasible expectation (as aforementioned) because of the measurement error that impacts all tests to some degree. Actually, the reliability (or consistency) of measurement associated with each assessment impacts the agreement between the two assessments. If the two assessments do not individually measure the construct with 100% reliability (consistency), regardless of using the same scale and protocol, then the final

[2] Concordance refers to exact agreement between final ratings as used in ratings research. In the test linking literature, concordance refers specifically to the linking of assessments that measure similar but not identical constructs (e.g., Kolen, 2004). Equating is often used when the tests are meant to be parallel forms of the same assessment.

scores or rating yielded by the two assessments can never achieve 100% concordance. Since perfect reliability is not possible in testing, there will not be perfect concordance. Two assessments designed to yield the same score can each have very good reliability and still not be concordant at 100% as aforementioned.

This does not mean concordance of final ratings should be ignored in the case of rater-based assessments. However, there is no clear-cut, agreed-upon minimum standard for the concordance of final ratings between different modalities of a rater-based assessment. Most guidance refers only to the magnitude of the validity coefficients (correlational evidence) or articles make an obscure reference to above 80% concordance being excellent. Jackson (1998) does suggest a standard of 70% based on research with Cohen's Kappa. Further interpretational insights can be gleaned from two empirical studies investigating that impact of modality on the assessment of speaking proficiency.

Jackson (1999) at the Defense Language Institute conducted a study comparing the results of the Speaking Proficiency Test (SPT) administered across several different modalities, including telephonic and face-to-face. Swender (2003) conducted a direct comparison between a face-to-face and telephonic version of the ACTFL OPI®. These were the only two studies with direct comparison (same test takers completing both assessment modalities) that could be found for rater-based speaking proficiency assessments. However, both provide a solid reference point for our interpretation of final rating concordance.

Jackson (1999) conducted a series of studies on the SPT with volunteers in Russian and Arabic. The SPT yields speaking proficiency ratings on the ILR scale. Although the interrater reliabilities were high, the absolute agreements between the final ratings were not. For the Russian study comparing face-to-face and telephonic modalities, the absolute agreement between final ratings was 54.7% (n = 64). When collapsing categories within the major levels (e.g., levels 1 and 1+ are within the level 1 of the ILR), the absolute agreement increased to 75%. When the ratings were allowed to be the exact same or off by +/- one step (regardless of whether or not it crossed a major level), the agreement jumped to 87.5%.

Jackson (1999) conducted two studies with Arabic learners comparing face-to-face interviews to desktop video-teleconferencing and tape-mediated interviews, respectively. In short, the exact agreement between the face-to-face and desktop video-teleconferencing final ratings was 68.2%, and the agreement between the face-to-face and tape-mediated was 50%. When the ratings were allowed to be the exact same or off by +/- one step (regardless of whether or not it crossed a major level), the agreement in both studies jumped to 90% or higher.

Swender (2003) conducted a study comparing face-to-face and telephonic ACTFL OPI® final ratings in Spanish for a group of learners at the Middlebury Language Schools. These students were volunteers, but they had the incentive of being awarded an official OPI rating and certificate. The two interviews were counterbalanced and conducted within 48 hours to prevent order and history effects from impacting the findings. The telephonic and face-to-face testing agreed exactly in 32 of 34 cases (94%).

Jackson (1999) and Swender (2003) report very different results in terms of concordance of final ratings between modalities. Differences in the study designs can help explain the results. Jackson (1999) suggests that the results might be related to low examinee motivation (no stakes volunteers) and lack of tester or examine familiarity with the technologies. Swender (2003) had participants who were motivated by the award of an official proficiency rating and the technology (telephone) was not unfamiliar to tester or examinees. Differences in interview protocol between the assessments used in the two studies may have contributed as well. Regardless, both these studies can provide insights into setting a concordance standard between final ratings.

For the initial development (piloting) of a new delivery modality of a rater-based assessment, a concordance (exact agreement) of 60-70% with the established modality of the assessment would be a sufficient start. For initial use, the 70% minimum standard for exact agreement between final ratings suggested by Jackson (1998, 1999) makes is feasible and appropriate. We recommend the within major level agreement (i.e., +/- one rating step within a major level, such as, novice) should be calculated as well and should be above 90%. This will diminish the impact of disagreements on test takers.

In the future, we recommend 84% concordance between the final ratings of two delivery modalities as the eventual goal—the assessments produce identical results five out of six times. This level of concordance coupled with good reliability and validity coefficients would allow the test user to have great confidence in using the ACTFL OPIc® to measure speaking proficiency.

Study 1 and Study 2

The next two sections present the descriptions and findings of Study 1 and Study 2. For each study, the following are reported: (1) research questions, (2) study methodology, (3) results by research question; and (4) discussion for that study. The methods sections provide enough detail about the two studies for a thorough review of the research (evidence quality) and for a study replication if desired. After both studies are presented, a general conclusions section is presented.

ACTFL OPIc® Validation Study 1

An initial validation study—referred to as Study 1—was conducted with a sample of 151 employees from a Korean company to investigate the reliability and validity of the ACTFL OPIc®, a computerized-version of the ACTFL Oral Proficiency Interview (ACTFL OPI®), as an assessment of speaking proficiency in English. This section of the report presents the research questions, methods, results and discussion for Study 1. The main goal of Study 1 is to gather initial psychometric evidence on the ACTFL OPIc to determine whether or not its use as a measure of English speaking proficiency is initially justified.

Study 1: Research Questions

The *ACTFL OPIc® Validation Study* 1 was designed to address the research questions presented in *Table 1*. The next section describes the methodology employed to address these research questions.

Study 1: Method

Participants—Overall

An initial group of 151 individuals were selected from the workforce at a company in Korea to participate in the *ACTFL OPIc® Validation Study 1*. This group consisted of 67 males (44.4%) and 84 females (55.6%) whose age ranged from 21 to 42 with an average age of 29. The majority of these individuals (83.4%) indicated that they were university graduates, while 3.3% had either graduated high school or not completed high school and 13.2% indicated attending graduate school. These individuals were randomly assigned to four groups: Pilot Study ($N = 20$), Validation Study Condition 1 ($N = 50$), Validation Study Condition 2 ($N = 50$), and a Holdout Sample ($N = 31$).

Table 1. Study 1 Research Questions

RQ1.	What is the overall interrater reliability and consistency of the ACTFL OPIc®?
RQ2.	How does the interrater reliability and consistency of the ACTFL OPIc® (first administration) compare to that of the ACTFL OPI® for the same sample of test takers?
RQ3.	What is the relationship between ACTFL OPIc® and ACTFL OPI® final ratings?
RQ4.	How do the underlying constructs of the ACTFL OPI® and ACTFL OPIc® compare?
RQ5.	What is the absolute agreement between ACTFL OPIc® and ACTFL OPI® final ratings?
RQ6.	Does the order of test administration impact the relationship between the ACTFL OPI® and ACTFL OPIc® (first administration)?
RQ7.	What impact does the self assessment have on the agreement of the ACTFL OPI® and ACTFL OPIc® (first administration) final ratings?
RQ8.	What is the test-retest reliability of the ACTFL OPIc® across two administrations with the same sample of test takers?
RQ9.	How do the underlying constructs of the first and second administrations of the ACTFL OPIc® compare?
RQ10.	What is the absolute agreement between the final ACTFL OPIc® rating at time one and the final ACTFL OPIc® rating at time two?
RQ11.	How do study participants view the ACTFL OPIc®? How do they view the ACTFL OPIc® in relationship to the ACTFL OPI®?

Of the 151 participants assigned to conditions, only 142 individuals participated in the study. Of these 142 participants, 20 individuals participated in Pilot Study. The purpose of Pilot Study was to pilot test a multiple user implementation of the ACTFL OPIc® system and to capture samples of speech to be used for rater training. Previous alpha and beta tests had been conducted in the United States to refine the ACTFL OPIc® process and system.

Individuals assigned to the Holdout Sample were asked to participate in the Validation Study when others originally assigned to the Validation Study dropped out of the study. A total of 99 individuals participated in the Validation Study. These individuals completed both an ACTFL OPI® and ACTFL OPIc® in addition to the pre- and post-assessment surveys. More details for the Validation Study are included in the *Study Design* section.

Participants—Validation Study

Thirty-seven males (37.4%) and 61 females (61.6%) participated in the Validation Study (one participant did not indicate gender). Most participants (68.7%) indicated that the highest level of education they had completed was a B.A. or B.S. degree. In terms of work experience, 64.6% of the participants had worked in their current job for 1-5 years, although 29.3% of participants had worked in their current job for less than one year. No participants reported working for more than 10 years in their current job. Only 36.4% of participants indicated serving in a supervisory role in their current job. The majority of participants (73.7%) reported that they do not use English as part of their job. In addition, the majority of participants (69.7%) indicated that they had to speak with people via the telephone with whom they had not had previous contact.

Participants were asked several questions on the pre-assessment survey about their experiences using the telephone and computers. The majority of participants indicated that they had never taken part in a telephonic job interview (82.8%) or taken a test via the telephone (90.9%). Approximately 58% of the participants indicated that they had been using computers for 1-5 years, while 31.3% indicated that they had been using computers for 6-10 years. A majority of respondents had applied for a job on the internet (89.9%) and had taken an online course (91.9%). Approximately 65% of participants had taken a language course online and 59.6% had taken a test on the internet.

Almost all participants indicated that they were required to use the internet as part of their job (96%) and that they use online messaging (99%). A majority of participants (73.7%) indicated that they use the internet for more than five hours at work in a typical day. Participants indicated using the internet at home less frequently than at work with 48.5% indicating internet usage for less than one hour and 37.4% indicating internet usage between one and two hours in a typical day.

Participants were also asked some questions about their previous English training/education and their previous experience with English testing. Most participants indicated that they first started to study English in primary school (39.4%) or middle school (56.6%). There was some variability in terms of the number of English courses that individuals had taken either at school or through private institutes. Most participants (66.7%) indicated taking between one and three courses, although 13.1% indicated taking between four and six courses and 10% indicated that they had taken 10 or more courses.

In terms of experience with English testing, the majority of participants had never taken an ACTFL OPI® (96%) or any of the Cambridge ESOL exams (98%). A majority of participants (66.7%) had taken the Test of English as a Foreign Language (TOEFL) or the Test of English for International Communication (TOEIC). Approximately 18% of participants indicated that they had taken another standardized test of English proficiency not mentioned in the other questions.

Study Design

Pre- and Post-Assessment Questionnaire Development. Before beginning data collection, it was necessary to develop and test both the pre- and post-assessment user surveys. The pre-

assessment survey included questions about the participant's background as well as measures of individual differences (e.g., test-taking self-efficacy). It was necessary to measure individual differences variables related to the test taking prior to exposure to the ACTFL OPI® and ACTFL OPIc® in order to establish temporal precedence for the data. Data from the pre-assessment survey can be used to statistically control for individual differences or demographic group differences when appropriate. The post-assessment included questions about reactions to the assessments and provided the opportunity for test takers to give feedback on the assessments. The post-assessment data will be useful for improving the assessment. Alpha and beta testing was conducted for the pre- and post-assessment surveys as well as the ACTFL OPIc® to ensure the instruments were functioning properly prior to the deployment of the system in Korea. The data from the alpha tests were analyzed iteratively and recommendations for improvement were made.

Participant Sampling. Language Testing International (LTI) planned the logistics of the data collection for the Validation Study. The Korean company provided a sample of 151 individuals who resembled the target testing population in terms of key demographics (e.g., gender, age, education, job, etc.). The purpose of stratified randomly sampling from the target population was to help to eliminate any systemic impacts of non-measured individual differences and ensure the results will generalize to the target population.

Participant Random Assignment. The 151 individuals selected by the Korean company were then randomly assigned to four conditions: Pilot Study (PS; N = 20), Validation Study Condition 1 (VS C1; N = 50), Validation Study Condition 2 (VS C2; N = 50), and a Holdout Sample (N = 31). The Holdout Sample was used as replacement for individuals who dropped out of either validation study condition.

Pilot Study. The PS was conducted in order to pilot test the ACTFL OPIc® and associated pre- and post-assessment user surveys with a small sample from the target population in Korea. This group received the ACTFL OPIc® and user surveys, but not an ACTFL OPI®. The sample of ACTFL OPIc® data was also used for rater training purposes. The data from PS was reviewed to determine if any changes needed to be made to the process or measures prior to beginning the Validation Study.

Validation Study. Individuals assigned to VS C1, took the pre-assessment survey, the ACTFL OPIc®, the ACTFL OPI®, and the post-assessment survey (in that order). Individuals assigned to VS C2, took the pre-assessment survey, the ACTFL OPI®, the ACTFL OPIc®, and the post-assessment survey (in that order). Additionally, individuals in VS C2 participated in the test-retest reliability study, taking a second ACTFL OPIc® approximately seven days after their first ACTFL OPIc®. The administration order of the ACTFL OPI® and ACTFL OPIc® assessments was counterbalanced in order to assess and control for the impact of test-taking order if necessary. The specific study procedures for VS participants are outlined in *Table 2* (see next page). *Figure 1* (see next page) provides a diagram of the specific design of the VS.

Rating Speech Samples

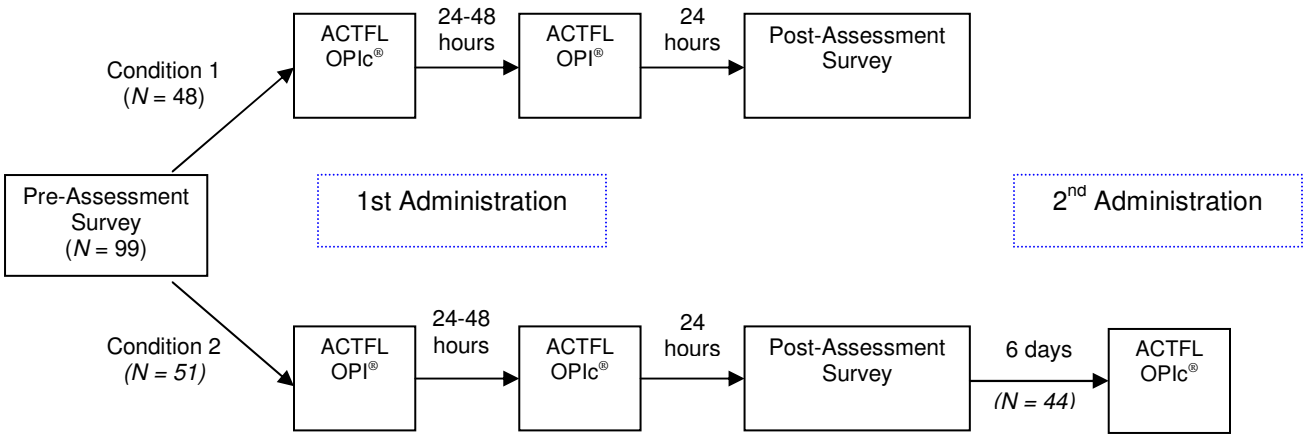
After participants completed all assessments, the ACTFL OPI® and ACTFL OPIc® samples were rated. Since the validity and reliability of any rater-based assessment is a function of the raters, the selection of raters and the design of the rating protocols were very important considerations.

Raters. Nine experienced, high-performing ACTFL OPI® testers were trained to rate the ACTFL OPIc® format. This eliminated the possibility of inexperience with the ACTFL guidelines for speaking proficiency (Breiner-Sanders, Lowe, Miles, & Swender, 2000) or the rating task being a source of systemic error variance in the study. Additionally, restricting the rater pool allowed us to look at the impact of rater characteristics and functioning with a small sample.

Table 2. Validation Study 1 Procedures

1	All participants in the Validation Study took the pre-assessment survey within the same specified time frame prior to the administration of the first test (either the ACTFL OPI® or ACTFL OPIc® depending upon the participant's assigned condition).
2	During the first round of testing, each person in VS C1 received an invitation to take the ACTFL OPIc® and took the ACTFL OPIc® (ACTFL OPIc® first administration). Each person in VS C2 was scheduled to take the ACTFL OPI® and took the ACTFL OPI®.
3	The second round of testing was scheduled to begin 24-48 hours after the completion of the first round of testing. Individuals assigned to VS C1 took the ACTFL OPI® and individuals assigned to VS C2 took to the ACTFL OPIc® (ACTFL OPIc® first administration). The approximate standardization of test-taking times was important for interpretation of the results and controlling for the impact of differential maturation and history between the two testing events.
4	All participants were asked to complete the post-assessment survey approximately 24 hours after the second round of testing.
5	Individuals assigned to VS C2 participated in the test-retest reliability study. Approximately, one week (7 days) after completing the second round of testing, each participant assigned to VS C2 completed a second ACTFL OPIc® (ACTFL OPIc® second administration).

Figure 1. Validation Study 1 Design



Rating Protocol. For the validation study, all raters followed the same content rating protocols (i.e., ACTFL guidelines) they would normally follow for the ACTFL OPI® and ACTFL OPIc® assessments. However, LTI utilized more raters than is typical for ACTFL OPI® testing. Each ACTFL OPI® and ACTFL OPIc® sample was rated by a *minimum of three raters*. This allowed for the highest quality assessment of reliability and validity evidence and for the use of more sophisticated rater agreement statistics (e.g., intraclass correlations; ICC; Shrout & Fleiss, 1979) and confirmatory factor analysis (CFA) – providing a more in-depth and powerful assessment of the ACTFL OPIc®.

Table 3 presents the protocols that were followed for all rater assignments regardless of assessment mode (ACTFL OPI® or ACTFL OPIc®) or individual case (participant). Once the ratings were completed, LTI transferred all data to SWA for analysis and reporting.

Measures

ACTFL Oral Proficiency Interview (ACTFL OPI®). The ACTFL Oral Proficiency Interview (ACTFL OPI®) is a standardized assessment of speaking proficiency in a foreign language. The assessment is administered in the form of a face-to-face or telephonic interview in which a certified ACTFL tester—serving as the interviewer—assesses the speaking proficiency of a test taker by asking the test taker a series of questions in the context of a structured conversation. The question content is based on the test taker's interests as determined by a preliminary set of questions in the interview and is adapted during the course of the interview based on the test taker's speaking proficiency level. In our studies, the ACTFL OPI® was administered in English to group of native Korean speakers over the telephone.

Table 3. Rater Assignment Protocols

Ratings were completed as part of a "masked" process. Raters *did not* know who they were rating (i.e., no names or other identifying information associated with the samples were presented to the raters). This prevented raters from recognizing the name across multiple assessments.

For VS C1, each participant had five unique raters, one for each rating position, who provided six ratings across 2 assessments (ACTFL OPI®, ACTFL OPIc® first administration). The five rating positions were as follows: (1) ACTFL OPI® interviewer and rater one; (2) ACTFL OPI® rater two; (3) ACTFL OPIc® first administration rater one; (4) ACTFL OPIc® first administration rater two; and (5) Dual rater for ACTFL OPI®/ ACTFL OPIc® first administration (ACTFL OPI® rater three; ACTFL OPIc® first administration rater three).

For VS C2, each participant had seven unique raters, one for each rating position, who provided nine ratings across 3 assessments (ACTFL OPI®, ACTFL OPIc® first administration, ACTFL OPIc® second administration). The seven rating positions were as follows: (1) ACTFL OPI® interviewer and rater one; (2) ACTFL OPI® rater two; (3) ACTFL OPIc® first administration rater one; (4) ACTFL OPIc® second administration rater one; (5) ACTFL OPIc® second administration rater three; (6) Dual rater for ACTFL OPI®/ ACTFL OPIc® first administration (ACTFL OPI® rater three; ACTFL OPIc® first administration rater three); and (7) Dual rater for ACTFL OPIc® first administration / ACTFL OPIc® second administration (ACTFL OPIc® first administration rater two, ACTFL OPIc® second administration rater two).

LTI made rater assignments and recorded the specific raters that were used in each position for each participant. All raters rotated through all seven rating positions.

Each rater did not know the position they were assigned on any particular case except when they were the interviewer.

The rating assignments were presented to the raters in random order within each modality.

All raters were asked to complete their ACTFL OPI® and ACTFL OPIc® ratings without communicating.

ACTFL Oral Proficiency Interviews (ACTFL OPI®) are conducted and rated by certified ACTFL testers. The interviews are recorded and typically rated by two certified testers—one who interviews the individual and rates the sample after the interview and one who serves as a rater only. The ACTFL testers compare the test taker's responses with criteria for ten proficiency levels (e.g., Intermediate Mid) specified in the *ACTFL Proficiency Guidelines – Speaking: Revised 1999* (Breiner-Sanders, Lowe, Miles, & Swender, 2000). The range of proficiency assessed with this test is Novice to Superior; individuals can achieve scores of Novice Low, Novice Mid, Novice High, Intermediate Low, Intermediate Mid, Intermediate High, Advanced Low, Advanced Mid, Advanced High, and Superior.

Previous research has produced support for the ACTFL OPI® construct (Dandonoli & Henning, 1990), and the ACTFL OPI® has been found to be reliable (Magnan, 1986; Surface & Dierdorff, 2003; Thompson, 1995). Surface & Dierdorff (2003) provided reliability evidence for the ACTFL OPI® for 19 languages.

ACTFL OPIc®. The ACTFL OPIc® is intended to be an internationally used semi-direct test of spoken English proficiency designed to elicit a sample of speech via computer-delivered prompts. An individual student, wishing to have his/her English language proficiency evaluated, will be able to access an ACTFL OPI®-like test without the presence of a live tester to conduct the interview. The range of proficiency assessed by this test is Novice to Advanced; individual scores of Novice Low, Novice Mid, Novice High, Intermediate Low, Intermediate Mid, Intermediate High, and base-line Advanced are the reporting options. The ACTFL OPIc® uses the same guidelines and scale as the ACTFL OPI®.

Each test is individualized. An algorithm selects prompts at random from a database of thousands of prompts. The task and topic areas of these prompts correspond to the test taker's linguistic, interest, and experience profiles. The approximate test time is 10 - 30 minutes, depending on the level of proficiency of the test taker. The speech sample is digitally saved and rated by certified ACTFL OPIc® raters. The *ACTFL Proficiency Guidelines – Speaking: Revised 1999* (Breiner-Sanders, Lowe, Miles, & Swender, 2000) are the basis for assigning a rating.

The ACTFL OPIc® is intended for language learners and language users. This test is potential appropriate for a variety of purpose: placement into instructional programs, screening for placement for hiring purposes, demonstration of an individual's linguistic progress, evidence of program effectiveness, and indication of readiness for a full ACTFL OPI® at the Advanced and Superior levels.

Pre-Assessment Survey. The pre-assessment web-based survey contained seven sections (Sections A-G; See *Appendix A* for items). The survey was developed in English and translated into Korean for administration to the participants.

Section A contained a few questions to gather identifying information from participants (e.g., last name). This information was necessary for linking responses between the pre-assessment survey, post-assessment survey, ACTFL OPI®, and ACTFL OPIc®.

Section B included eight demographic items (e.g., gender, date of birth) in order to gather background information about participants. Section C contained questions related to participant's experiences using the telephone and computers.

The items in Section D are related to attitudes toward computerized/telephonic tests. Items in Section E assessed student's test-taking self-efficacy (based on items used in Bauer, Maertz, Dolen, & Campion, 1998). Test-taking self-efficacy is a measure of an individual's confidence in their ability to take tests.

Section F contained questions from the Sensory Modality Preference Inventory to assess visual and auditory learning style of the examinee (Sensory Modality Preference Inventory, 2002). This inventory assesses the extent to which test

takers prefer to learn through visual means or through auditory means.

Finally, Section G contains a goal orientation scale developed by Vandewalle (1997). Goal orientations are defined as dispositions toward developing or demonstrating ability in achievement situations (Dweck, 1986). There are three types of goal orientation measured with this instrument: learning, prove performance, and avoid performance. Individuals who have a learning goal orientation pursue goals related to learning, are motivated to learn new skills, and increase their knowledge in various areas. Individuals who have a prove performance goal orientation like to show how knowledgeable they are around others and enjoy it when individuals are impressed by their knowledge. Individuals who have an avoid performance goal orientation try to avoid showing when they do not have much knowledge in a particular area.

***Post-Assessment Survey*.** The post-assessment web-based survey contained four sections (Sections A-D; See *Appendix B* for items). The survey was developed in English and translated into Korean for administration to the participants.

Section A contained the same questions that were included on the pre-assessment to gather identifying information from participants (e.g., last name). This information was necessary for linking responses between the pre-assessment survey, post-assessment survey, ACTFL OPI®, and ACTFL OPIc®.

Section B contained items related to the ACTFL OPIc®, including an item asking participants to indicate if they read the ACTFL OPIc® instructions in English or Korean. Other items in this section assessed reactions to the initial instructions, background survey, self-assessment, test description/instructions, test format/tutorial sample, and general reactions to the ACTFL OPIc®. Section C contained items related to the ACTFL OPI®. This section focused on general reactions to the ACTFL OPI®. Both Section B and C contained open-ended questions related to the ACTFL OPI® and ACTFL OPIc®, respectively.

Section D contained items which are meant to compare test takers' reactions to the ACTFL OPI® and ACTFL OPIc®. This section also contained closed-ended and open-ended questions which ask test takers to indicate which assessment offered a better opportunity to demonstrate their best speaking proficiency and which assessment they would prefer to take in the future.

It is important to note that due to some issues with the data collection procedures, not all open-ended responses were recorded for all participants. Therefore, it is important to be cautious when interpreting the responses to these comments as they may not be representative of all of the participants.

Analytic Procedures

After receipt, the data from LTI and the survey contractor were cleaned, formatted, and aggregated for analysis. Basic descriptive statistics for the sample were calculated from the pre- and post-assessment survey data as well as the ACTFL assessments.

The method used to address the research objectives are presented by research question.

***RQ1 and RQ2*.** Interrater reliability was calculated using intraclass correlations (ICC; Shrout & Fleiss, 1979).

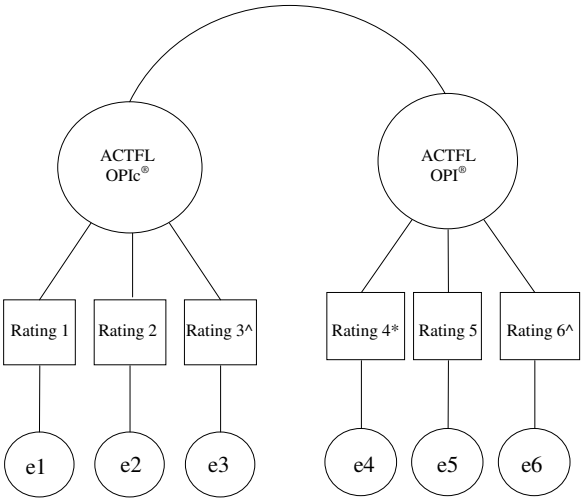
Intraclass correlations are often used as reliability coefficients among evaluations of items that are deemed to be in the same category or class. The ICC assesses rating reliability by comparing the variability of different ratings of the same subject to the total variation across all ratings and all subjects. ICCs can also account for rater agreement as well as reliability.

Shrout and Fleiss (1979) show that one can use the ICC in two ways: To estimate the reliability of a single rating, or to estimate the reliability of a mean of several ratings. The one-way random ICC is appropriate when raters are different for different cases and when the unit of analysis is a single observation as opposed to an averaged observation (McGraw & Wong, 1996; von Eye & Mun, 2005).

Additionally, an estimate of maximum scale reliability, $R_{Max,}$ was calculated using data from the confirmatory factor analysis (CFA) model represented in *Figure 2*. R_{Max} is defined as the square of the canonical correlation of the scale items (or ratings) with the latent scale factor (Drewes, 2000). R_{Max} is the maximum squared correlation of a weighted composite of the items with the underlying factor that can be attained.

R_{Max} can be used as a single index of composite scale reliability much like coefficient alpha, but with better properties, For example, while coefficient alpha provides an underestimation of reliability, R_{Max} provides a maximum estimate of scale reliability. Providing that the final set of manifest variables is scalable and some items are judged to have acceptable reliability, R_{Max} can be computed.

Figure 2. CFA Measurement Model of ACTFL OPI® and ACTFL OPIc® Relationship



The scale can be said to exhibit high reliability if R_{Max} is .80 or above, moderate reliability if R_{Max} is between .60 and .80, and low reliability if R_{Max} is less than .60. Additionally, R_{Max} can be used to calculate a single index of composite scale validity or construct validity, $R_{cv.}$ This construct validity coefficient is calculated by taking the square-root of $R_{Max.}$

Finally, the absolute agreement between the three raters (i.e., percentage of the time all three raters agreed) was calculated as a measure of interrater consistency for both administrations. As a cautionary note, a high percentage of absolute agreement is a very difficult standard to achieve across more than two raters. When two raters are utilized, absolute agreement between raters of 70% or higher is generally considered acceptable. Each additional rater employed in the process decreases the minimum acceptable agreement percentage.

***RQ3*.** To explore the relationship between final ratings from the ACTFL OPI® and the ACTFL OPIc®, we computed two correlation coefficients (Pearson's *r* & Spearman's *R*) between the final ratings obtained from the ACTFL OPI® and the ACTFL OPIc® (first administration) and between the final ratings from the ACTFL OPI® and the ACTFL OPIc® (second administration). In addition, the correlations were computed for the ratings from the common rater position [i.e., each participant has a rater who rated both the ACTFL OPI® and ACTFL OPIc® (first administration) sample for that individual].

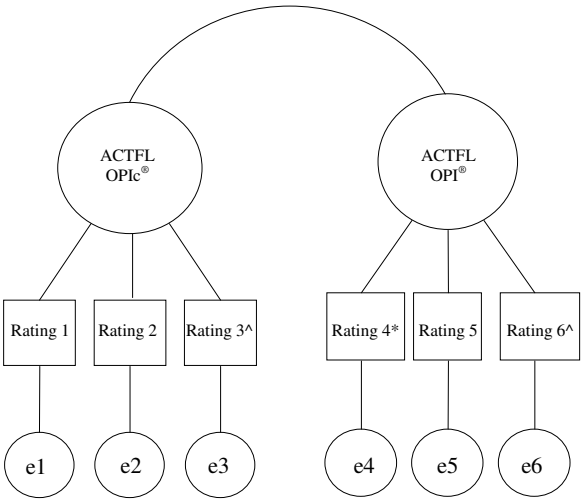
Sometimes called a *product–moment* correlation, Pearson's correlation (*r*) is one the most widely used methods of assessing interrater reliability. This correlation assesses the degree to which ratings covary. In this sense, reliability can be depicted in the classical framework as the ratio of true score variance to total variance (i.e., variance in ratings attributable to *true* speaking proficiency divided by total variance of ratings).

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

RQ4. Confirmatory factor analysis (CFA; Hatcher, 1994) was employed to assess the relationship between the two speaking proficiency constructs— ACTFL OPI® and ACTFL OPIc® (first administration). In order to address *RQ4*, we tested the following CFA models.

Model 1 (see *Figure 3*) consists of two correlated constructs, one for the ACTFL OPI® and one for the ACTFL OPIc® (first administration).

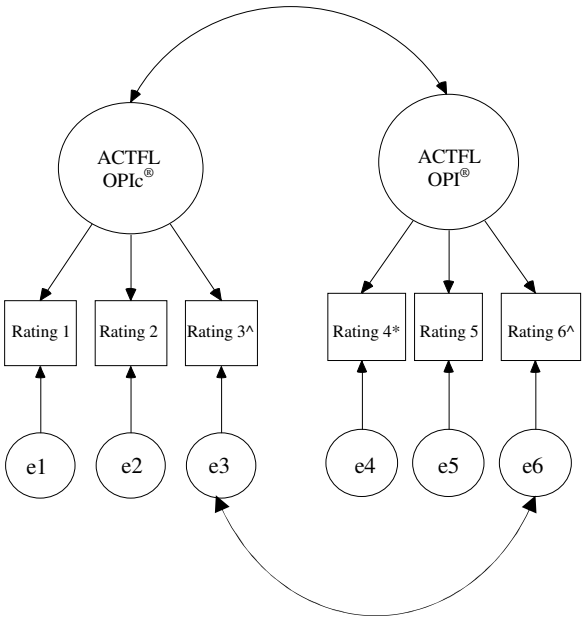
Figure 3. Basic CFA Model of Relationship between the Two Assessments



The results from the CFA analysis testing this Model 1 provide a correlation between ACTFL OPI® and ACTFL OPIc® (first administration) at the latent level. Additionally, the model provides factor loadings for each rating (i.e., the relationship of each rating to the common factor) that can be used to calculate maximum reliability (R_{Max}) and construct validity (R_{cv}) for the ACTFL OPI® and ACTFL OPIc® (first administration) to assess reliability and construct validity more thoroughly. R_{cv} is the square root of the maximum reliability. As with the R_{Max} coefficient, R_{cv} values close 1.00 indicate excellent validity for the items or ratings used to operationalize the construct.

Model 2 (*see Figure 4*) is also a correlated two-factor model. However, this model takes into account the fact that there was a rater position occupied by a common rater across the ACTFL OPI® and ACTFL OPIc® (first administration) constructs for each case. By correlating the "error" terms for the common rater positions across the ACTFL OPI® and ACTFL OPIc® (first administration), any variance that might be idiosyncratic to the common raters was modeled, therefore, accounting for its impact on model fit. Typically, the influence of an unspecified factor is relegated to the error terms of the items or ratings.

Figure 4. Two-factor CFA Model Accounting for the Common Rater Across Assessments



If multiple items are influenced by an unspecified (unmodeled) the factor, then a portion of the error variance between these items is correlated. At this point, the unmodeled correlated error variance between the items may degrade model fit. This "systematic" error variance must be modeled to accurately assess fit. In this case, we knew there was a potential for characteristics of the common raters to influence the ratings across the assessments, so we correlated the error terms for those items to ensure this was not a problem.

RQ5. To determine the concordance between the ACTFL OPI® and ACTFL OPIc® final ratings, the absolute agreement between the final ratings from both assessments was computed. In other words, the percentage of cases in which the final ACTFL OPIc® (first administration) rating agreed with the final ACTFL OPI® rating and the percentage of cases in which the final ACTFL OPIc® (second administration) rating agreed with the final ACTFL OPI® rating were calculated. Please note the ACTFL OPI® was measured once.

RQ6. To determine if the administration order of the ACTFL OPI® and ACTFL OPIc® (first administration) impacted their ratings or the agreement between their ratings, we conducted several analyses. Two one-way ANOVAs were used to determine if differences between the groups [i.e., ACTFL OPI® administered first verse the ACTFL OPIc® (first administration) administered first] impacted the ACTFL OPI® and ACTFL OPIc® final ratings. A Pearson's chi square (χ^2) was used to determine if the order had an impact on the agreement between the ACTFL OPI® and ACTFL OPIc® (first administration).

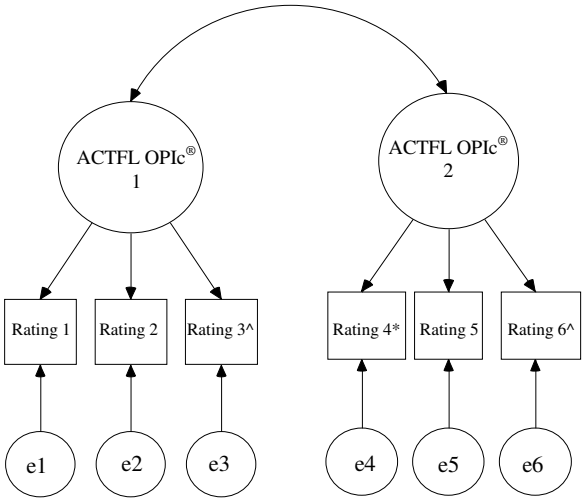
RQ7. Participants were asked to provide a self-assessment of their language proficiency at the beginning of the ACTFL OPIc®. This assessment was used to modify the difficulty level of the prompts received by the individual. To determine if participant self-assessments impacted the match between the ACTFL OPI® and ACTFL OPIc® (first administration) final ratings, a cross-tabulation between ACTFL OPIc® (first administration) self-assessments and the match of the ACTFL OPI® and ACTFL OPIc® (first administration) final ratings was computed. A chi square (χ^2) was conducted to determine if the impact was statistically significant.

RQ8. Using data from participants who took the ACTFL OPIc® twice, we computed correlations between the final ratings obtained from the first administration (ACTFL OPIc® first administration) and the second administration (ACTFL OPIc® second administration) in order to determine the test-retest reliability of the assessment. Pearson's *r* and Spearman's *R* were calculated. Additionally, the coefficients were calculated between the common rater position ratings for the ACTFL OPIc® (first administration) and ACTFL OPIc® (second administration) to determine the test-retest reliability for the consistent rater position across these two tests.

RQ9. CFA was employed to assess the relationship between the ACTFL OPIc® constructs across administrations (test-retest).

Model 3 (see *Figure 5*) is a correlated, two-factor model for assessing the relationship between the ACTFL OPIc® (first administration) and ACTFL OPIc® (second administration).

Figure 5. CFA Model of Relationship between ACTFL OPIc® Test-Retest Administrations



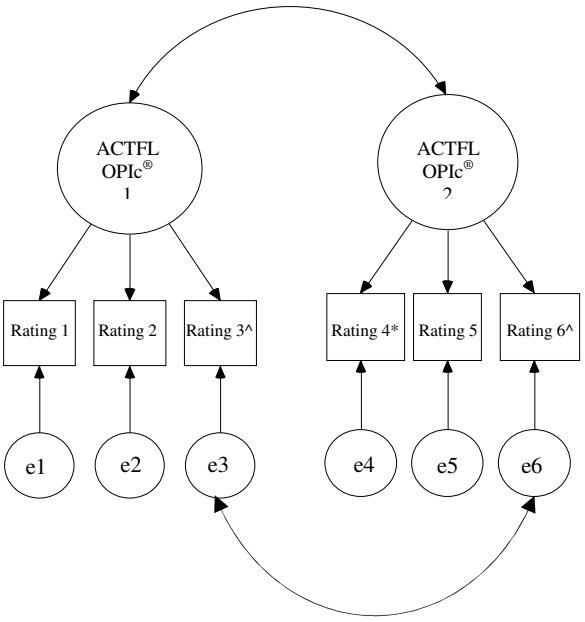
CFA results from a test of this model indicate how the underlying constructs from the first and second administration of the ACTFL OPIc® compare, providing a latent correlation between them. Additionally, R_{Max} and R_{cv} can be calculated for ACTFL OPIc® (first administration) and ACTFL OPIc® (second administration).

Model 4 (*see Figure 6*) is the same correlated, two-factor model between the ACTFL OPIc®

(first administration) and ACTFL OPIc® (second administration) with one exception. This model takes into account the fact that there was a rater position occupied by a common rater across the ACTFL OPIc® (first administration) and ACTFL OPIc® (second administration) constructs for each case.

By correlating the "error" terms for the common rater positions across the ACTFL OPIc® (first administration) and ACTFL OPIc® (second administration), any variance that might be idiosyncratic to the common raters was modeled, therefore, accounting for its impact on model fit.

Figure 6. CFA Model of Relationship between ACTFL OPIc® Administrations with Common Rater Modeled



RQ10. To determine the concordance between ACTFL OPIc® (first administration) and ACTFL OPIc® (second administration) final ratings, the absolute agreement—the percentage of cases the final ACTFL OPIc® (first administration) rating agreed with the final ACTFL OPIc® (second administration) rating—was computed.

RQ11. We analyzed the post-assessment survey data to determine user reactions to the ACTFL OPI® and ACTFL OPIc®. Although a number of process-oriented questions were asked, reviewing the results of every question is beyond the scope of this document. However, *Appendix C* presents the complete item-level data from the post-assessment survey. To address *RQ11*, we examined the results of the agreement items (1 = *Strongly Disagree* to 5 = *Strongly Agree*) listed in *Table 4 (next page)*.

Additionally, the participants were asked, "In which format (ACTFL OPI®/ ACTFL OPIc®) did you feel you were able to demonstrate your best speaking proficiency?" Participant responses to these and other questions in *Appendix C* can be used to determine user reactions and to refine the ACTFL OPIc®.

Study 1: Results

This section presents the findings for the *ACTFL OPIc® Validation Study 1* by research question (*RQ1-RQ11*).

RQ1: What is the overall interrater reliability and consistency of the ACTFL OPIc®?

To assess the interrater reliability for the ACTFL OPIc®, intraclass correlations (ICCs) were calculated across the three ACTFL OPIc® rater positions (first administration) and across the three ACTFL OPIc® rater positions (second administration).

The ICC for the ACTFL OPIc® (first administration), .94, was significant ($F = 46.63$, $p < .001$, $n = 96$, *95% C.I.* = .92 - .96).

The ICC for the ACTFL OPIc® (second administration), .79, was significant as well ($F = 12.40$, $p < .001$, $n = 42$, *95% C.I.* = .68 - .87). For an ICC, a coefficient of .79 is at the bottom of the acceptable range. However, it should be noted that the ICC for the second administration ACTFL OPIc® (.79) was most likely lower than the ICC for first administration of the ACTFL OPIc® because of the smaller sample size.

Table 4. Participant Feedback Items

1. *I believe my performance on the ACTFL OPIc® accurately reflects my current speaking proficiency level.*
2. *I believe the ACTFL OPIc® is an effective way to measure English speaking proficiency.*
3. *I would recommend taking an ACTFL OPIc® to a friend who needs their speaking proficiency assessed.*
4. *I believe my performance on the ACTFL OPI® accurately reflects my current speaking proficiency level.*
5. *I believe the ACTFL OPI® is an effective way to measure English speaking proficiency.*
6. *I would recommend taking an ACTFL OPI® to a friend who needs their speaking proficiency assessed.*
7. *I thought it was more difficult to demonstrate my speaking proficiency via the computer than with a live interviewer over the telephone.*
8. *Both the computer and telephonic interviews provided an adequate opportunity for me to demonstrate my speaking proficiency.*
9. *The ACTFL OPIc® was more user friendly than the ACTFL OPI®.*
10. *The ACTFL OPIc® provided a better opportunity for me to demonstrate my speaking proficiency.*
11. *I preferred the testing format with a live interviewer than with the Avatar.*
12. *It was easier to understand questions from a live interviewer than from the Avatar.*
13. *I felt more comfortable recording my answers on the computer than providing answers to an interviewer.*

As an additional means of assessing reliability, maximum reliability (R_{Max}) was calculated for both administrations of the ACTFL OPIc® using the CFA results and was found to be .98 in both cases, suggesting high consistency among the raters at the construct level for the ACTFL OPIc®.

Taken together, the ICC and R_{Max} results provide sufficient evidence of reliability for the ACTFL OPIc®. Since R_{Max} is more "accurate" (other reliability estimations frequently underestimate reliability; Drewes, 2000), we are less concerned about the .79 ICC.

Finally, the absolute agreement between the three raters was calculated as a measure of interrater agreement for both administrations. As a cautionary note, a high percentage of absolute agreement is a very difficult standard to achieve across more than two raters. There was 59% absolute agreement across all three raters for the ACTFL OPIc® (first administration) and 41% for the ACTFL OPIc® (second administration).

Keep in mind that the final rating was assigned based on the agreement of two of the three raters, so the absolute agreement between the three raters had very little impact on the final ratings. The typical ACTFL rating process uses two raters with a third being used to arbitrate disagreements.

To address any potential concerns about low absolute agreement between three raters, we calculated absolute agreement for each of the three rater pairs (i.e., rater 1 and rater 2, rater 1 and rater 3, rater 2 and rater 3) on the first administration of the ACTFL OPIc and found the agreement to range between 71% and 76%, which is acceptable for a new assessment format.

RQ2: How does the interrater reliability and consistency of the ACTFL OPIc® compare to that of the ACTFL OPI® for the same sample of test takers?

The ICC for the ACTFL OPI®, .93, was significant ($F = 40.35$, $p < .001$, $n = 96$, *95% C.I.* = .90 - .95). The R_{Max} coefficient for the ACTFL OPI® was .98, and the absolute

agreement across the three ACTFL OPI® raters was 53%. Again, a high percentage of absolute agreement is a difficult standard to achieve with three or more raters. The reliability results for the ACTFL OPIc® (first administration) appear to be very similar to the ACTFL OPI® for the same group of test takers. Additionally, the same pattern of increased absolute agreement was seen for the rater pairs over the rater triplet. Both assessments appear to have sufficient reliability for testing purposes.

RQ3. What is the relationship between ACTFL OPIc® and ACTFL OPI® final ratings?

To determine the relationship between the ACTFL OPIc® and ACTFL OPI®, correlations between the final ratings of the ACTFL OPI® and the ACTFL OPIc® (first administration) and the final ratings of the ACTFL OPI® and ACTFL OPIc® (second administration) were calculated.

The correlations between the ACTFL OPI® and ACTFL OPIc® (first administration) were significant ($r = .92$, $p < .001$; $R = .91$, $p < .001$) and indicate a strong positive relationship between the assessments.

The correlations between the ACTFL OPI® and ACTFL OPIc® (second administration) were significant ($r = .94$, $p < .001$; $R = .94$, $p < .001$) and also indicate a strong positive relationship.

Additionally, we computed a correlation between the common rater position for the ACTFL OPI® and ACTFL OPIc® (first administration). The result of this analysis indicated a strong positive relationship ($r = .90$, $p < .001$; $R = .90$, $p < .001$).

RQ4. How do the underlying constructs of the ACTFL OPI® and ACTFL OPIc® compare?

Two confirmatory factor analysis (CFA) models were tested to address this research question. If the models provide an acceptable fit for the data, then we can use and interpret the latent correlation coefficient between the constructs and the standardized loadings for each rating. To evaluate model fit, a number of indices are used. These can be found in *Table 5 (see next page)*.

Criteria specified by Hu and Bentler (1999), Millsap (2002), and Vandenburg and Lance (2000) were examined to assess the overall fit of the measurement models. The ratio of chi-square to degrees of freedom (χ^2/df) was computed, with ratios of less than 2.0 indicating a good fit.

Vandenberg and Lance (2000) suggest using two absolute indices – the root mean square error of approximation (RMSEA) and the standardized root-mean-square residual (SRMR). For RMSEA, good fit is indicated by values less than 0.05; values from 0.05 to 0.10 are indicative of moderate fit and values greater than 0.10 are taken to be evidence of a poorly fitting model (Browne & Cudeck, 1993). For SRMR, values less than .10 are indicative of acceptable model fit (Kline, 1998).

However, since absolute indices can be adversely affected by sample size (Loehlin, 1992), two other relative indices, the comparative fit index (CFI) and the Tucker and Lewis index (TLI) were computed to provide a more robust evaluation of model fit (Tanaka, 1987; Tucker & Lewis, 1973). For CFI and TLI, coefficients closer to unity indicate a good fit, with acceptable levels of fit being above 0.90 (Marsh, Balla, & McDonald, 1988).

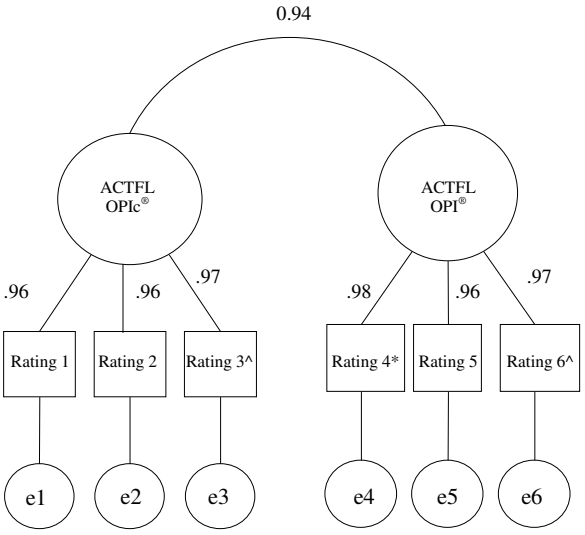
Model 1, as shown in *Figure 7 (see next page)*, represents a correlated two-factor model and assesses the relationship between the ACTFL OPI® and ACTFL OPIc® (first administration) constructs. As can be seen in *Table 5*, which presents the fit statistics for each CFA model tested, the correlated two-factor model provided an excellent fit for the data. All of the fit indices were within the margins of good to excellent fit. The latent correlation between the ACTFL OPI® and ACTFL OPIc® (first administration), .94, provided evidence of a very strong relationship between the constructs. The R_{Max} (reliability) and R_{CV} (construct validity) coefficients for the ACTFL OPI® were .98 and .99, respectively. The R_{Max} (reliability) and R_{cv} (construct validity) coefficients for the ACTFL OPIc® (first administration) were .98 and .99, respectively.

Table 5. Comparison of Model Fit Statistics for CFAs

Model	X^2	Df	X^2/Df	CFI	TLI	RMSEA	RMSEA 90% CI	SRMR
Model 1: ACTFL OPI® and ACTFL OPIc® as two correlated factors	9.36	8	1.17	1.00	1.00	.04	[.00 - .13]	.01
Model 2: ACTFL OPI® and ACTFL OPIc® (Model 1) with common rater	4.95	7	.71	1.00	1.00	.00	[.00 - .10]	.01
Model 3: ACTFL OPIc® (first administration) – ACTFL OPIc® (second administration) as two correlated factors	5.36	8	.67	1.00	1.00	.00	[.00 - .14]	.01
Model 4: ACTFL OPIc® (first administration) – ACTFL OPIc® (second administration; Model 3) with common rater	4.14	7	.59	1.00	1.00	.00	[.00 - .13]	.01

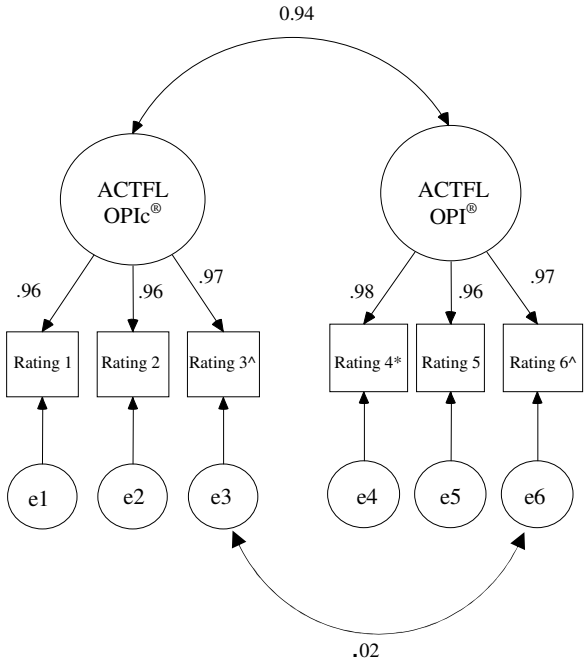
Note. CFI = comparative fit index; TLI = Tucker-Lewis index (also known as the non-normed fit index); RMSEA = root mean square error of approximation; SRMR = standardized root-mean-square residual. Summary consists of each group validated with itself, as well as its cross-validation with the other two groups. A one-factor model was tested and demonstrated degraded fit compared to the two-factor models.

Figure 7. Model 1 - Correlated Two-Factor Model: ACTFL OPI® and ACTFL OPIc® (first administration)



As shown in *Figure 8*, Model 2 accounts for the common rater position across the ACTFL OPI® and ACTFL OPIc® (first administration) by correlating "error" terms for the rating positions that share the common rater.

Figure 8. Model 2 - Correlated Two-Factor Model: ACTFL OPI® and ACTFL OPIc® (first administration) with Correlated Error Terms



As can be seen from *Table 5*, the fit for Model 2 is exceptional and slightly better than the fit for

Model 1. However, the improvement is not great enough in magnitude to believe an issue exists, and correlating the error terms had no impact on the latent correlation, the standardized loadings, or the R_{Max} and R_{cv} coefficients. Additionally, the path between the two error terms was virtually zero. The fit of Model 2 is not significantly better than Model 1.

Overall, the ACTFL OPI® and ACTFL OPIc® were found to be highly related at the construct level. Therefore, the assessments are measuring the same construct.

RQ5. What is the absolute agreement between ACTFL OPI® and OPIc® final ratings?

In terms of the absolute agreement or concordance between the ACTFL OPI® (first administration) and ACTFL OPIc® (first administration), the final ratings of the ACTFL OPI® and ACTFL OPIc® (first administration) agreed for 63% of the participant cases. The ACTFL OPI® (first administration) and ACTFL OPIc® (second administration) final ratings agreed for 67% of the cases.

Overall, the relationship between the final ACTFL OPI® and ACTFL OPIc® ratings was robust, although the absolute agreement should be slightly higher for use (e.g., 70% for initial use and 84% for prolonged use). The agreement was sufficient for initial development of the assessment (60-70%). Of note, for the first administration of the ACTFL OPIc®, the agreement between final ratings jumped to 85% within the major categories (novice, intermediate, advanced) and to 98% when the major category boundaries were ignored and agreement was defined as an exact match or being off by +/- one step.

The strong correlation coefficients between final ratings, coupled with the absolute agreement results, suggest that the differences in proficiency ratings are small in magnitude and consistent in direction. Given that the CFA results provide strong construct validity evidence for the ACTFL OPI® and ACTFL OPIc®, there may be a systematic source of measurement error impacting the absolute agreement of the final ratings. We assessed two potential sources—administration order and self-assessment.

RQ6. Did the order of test administration impact the relationship between the ACTFL OPI® and ACTFL OPIc® (first administration)?

The administration order of the initial assessment did not have a statistically significant impact on the final rating of the ACTFL OPI® (F = .22, p = .64) or the ACTFL OPIc® (first administration; F = .02, p = .89). Additionally, the administration order did not have a statistically significant impact on the agreement between the ACTFL OPI® and ACTFL OPIc® (first administration) final ratings (χ^2 = 2.23, df = 3, p = .53).

RQ7. Since the ACTFL OPIc® relies on the user's language proficiency self assessment, what impact (if any) did the self assessment have on the agreement of the ACTFL OPI® and ACTFL OPIc® final ratings?

Table 6 (see next page) provides the results of the cross-tabulation between the individual's self-assessment and the match (agreement or concordance) between the ACTFL OPI® and ACTFL OPIc® (first administration) ratings.

37.5% of the cases did not agree (36 of 96). Of those 36 disagreements, 26 (72%) were at self-assessment level one. Of the 26 disagreements at level one, 85% (22) resulted from cases where the ACTFL OPIc® final rating provided an underestimation of the ACTFL OPI® final rating. Across all the levels, 75% of the disagreements were underestimates of the ACTFL OPI® final ratings and only 25% were overestimates.

Although the descriptive results suggest the ACTFL OPIc® may produce an underestimate of the ACTFL OPI® at proficiency self-assessment level one, this interpretation should be tempered with caution because the relationship between self-assessment level and ACTFL OPI® -ACTFL OPIc® (first administration) agreement was not statistically significant (χ^2 = 4.66, df = 3, p = .19). However, the small sample size and the skew of the sample to the low proficiency end of the spectrum could be impacting the statistical significance.

Table 6. Cross of Participant Proficiency Self-assessment with the Agreement of their ACTFL OPI and ACTFL OPIc Final Ratings

Self-Assessed Speaking Proficiency	*ACTFL OPI® and ACTFL OPIc® Final Ratings*		*Break Down of Disagreements by Direction*	
	Agree	*Disagree*	*ACTFL OPIc® Underestimated ACTFL OPI®*	*ACTFL OPIc® Overestimated ACTFL OPI®*
Level 1: In English, I can understand and respond to basic, predictable greetings and expressions. I can name basic objects, colors, days of the week, foods, clothing items, numbers, etc. I cannot always ask simple questions or speak in sentences.	40	26	22	4
Level 2: In English, I can participate in a simple conversation about myself, familiar people and places, and daily routines. I can satisfy some basic, daily survival needs. I can say a few simple sentences and ask simple questions.	8	8	4	4
Level 3: In English, I can participate in short conversations about myself, daily routines, work/school and hobbies. I can easily produce a series of simple sentences on these familiar topics and routines. I can also ask questions when needed.	9	1	-	1
Level 4: In English, I can participate in conversations about topics and activities related to home, work/school, personal interests, and current events. I can talk at length about activities or events in the past, present and future. I can give explanations when asked and can handle routine situations, even when there may be an unexpected complication.	3	1	1	-
Total Across All Levels	60	36	27	9

Therefore, the issue of self-assessment at level one and underestimation of the ACTFL OPI® final rating should be investigated further. It appears that some test takers may be underestimating their proficiency, and this may be impacting the concordance between assessments.

RQ8. What is the test-retest reliability of the ACTFL OPIc® across two administrations with the same sample of test takers?

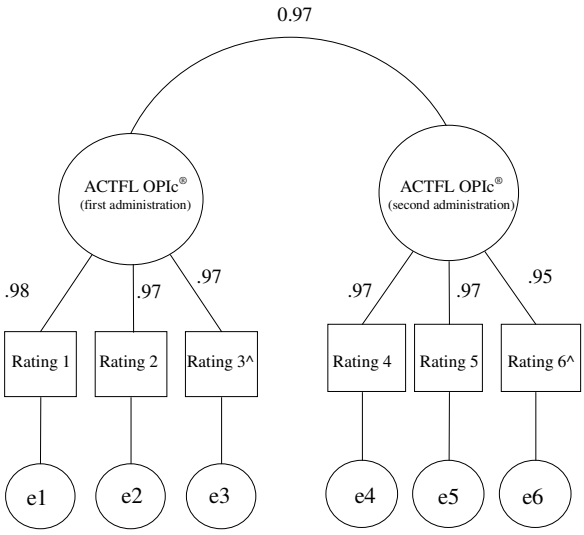
The correlation between the final ratings obtained from the first administration of the ACTFL OPIc® and the second administration of the ACTFL OPIc® was calculated using two different correlation coefficients. Regardless of coefficient, a strong degree of relationship was found between ACTFL OPIc® (first administration) and ACTFL OPIc® (second administration) final ratings ($r = .94$, $p < .001$; $R = .91$, $p < .001$), providing evidence for test-retest reliability of the ACTFL OPIc®.

In addition, we computed a correlation between the common rater for the ACTFL OPIc® (first administration) and ACTFL OPIc® (second administration). The result of this analysis indicated a fairly strong relationship between the ratings from the common rater ($r = .89$, $p < .001$; $R = .89$, $p < .001$).

RQ9. How do the underlying constructs of the first and second administration of the ACTFL OPIc® compare?

Model 3, as shown in *Figure 9*, represents a correlated two-factor model and assesses the relationship between the ACTFL OPIc® (first administration) and ACTFL OPIc® (second administration) at the latent level of analysis.

Figure 9. Model 3: ACTFL OPIc® (first administration) – ACTFL OPIc® (second administration) as two correlated factors



As can be seen in *Table 5*, the correlated two-factor model provided an excellent fit for the data. The latent correlation (see *Figure 9*) between the ACTFL OPIc® (first administration) and ACTFL OPIc® (second administration), .97, provides evidence of a very strong relation between the constructs. The R_{Max} and R_{CV} coefficients for the ACTFL OPIc® (first administration) were .98 and .99, respectively. The R_{Max} and R_{cv} coefficients for the ACTFL OPIc® (second administration) were .98 and .99, respectively.

Model 4 (not pictured) accounts for the common rater position across the ACTFL OPIc® (first administration) and ACTFL OPIc® (second administration) by correlating "error" terms for the rating positions that share the common rater. As can be seen from *Table 5*, the fit for Model 4 is exceptional and slightly better than the fit for Model 3. However, the improvement is not great enough in magnitude to believe an issue exists, and correlating the error terms had no impact on the latent correlation, the standardized loadings, or the R_{Max} and R_{cv} coefficients. The fit of Model 4 is not significantly better than that of Model 3.

The results of these analyses provide strong evidence for the ACTFL OPIc® construct and its repeatability across administration.

RQ10. What is the absolute agreement between the final ACTFL OPIc® rating at time one and the final ACTFL OPIc® rating at time two?

The final ratings from the ACTFL OPIc® (first administration) and the ACTFL OPIc® (second administration) agreed for 76% of the cases, demonstrating acceptable absolute agreement when interpreted in light of the test-retest reliability and CFA results.

RQ11. How did participants view the ACTFL® OPIc? How did they view the ACTFL OPIc® in relationship to the ACTFL OPI®?

The post-assessment survey was used to assess what users thought about their experiences taking the ACTFL OPI® and the ACTFL OPIc®. *Appendix C* contains tables which present complete item-level information for the questions on the post-assessment survey. *Table 7* (see next page) provides the results for the specific user feedback items highlighted for review.

Overall, a review of the item-level results in *Table 7* indicates that participants favored the ACTFL OPI® experience over the ACTFL OPIc® experience. In addition to the agreement items, each participant was asked, "In which format (ACTFL OPI®/ ACTFL® OPIc) did you feel you were able to demonstrate your best speaking proficiency?" Forty-four of the post-assessment survey respondents indicated the ACTFL OPI®, whereas, 15 indicated both assessments, and 10 reported the ACTFL OPIc®.

Although user preferences and reactions to the assessment are not psychometric qualities from a test validation perspective, these data can be useful in improving and refining the assessment. The data provided in *Appendix C* should be used to adjust the ACTFL OPIc® as appropriate.

Study 1: Discussion

This initial *ACTFL OPIc® Validation Project* study provided an investigation of the psychometrics of the ACTFL OPIc® as an assessment of speaking proficiency in English. The goal of this study was to start accumulating evidence on the validity and reliability of the ACTFL OPIc®. To this end, a sample of Korean employees completed both an ACTFL OPIc® and an ACTFL OPI® within a rigorous field experiment design that included random assignment and counterbalancing of assessments. This first study yielded some impressive preliminary evidence of validity and reliability for the ACTFL OPIc®. Although there are a few issues to be addressed to improve the assessment, this study should be viewed as a robust initial step in establishing the psychometric properties of the English version of the ACTFL OPIc® and the ACTFL OPIc® protocol in general. The next sections summarize the key findings, discuss several issues, and provide recommendations to the test developers.

Summary of Key Findings

This section highlights the key findings for the initial validation study. The study 1 results section provides a complete reporting of the findings.

- Taken together, the ICC and R_{max} results provide sufficient evidence of interrater reliability for the ACTFL OPIc®. The ICC for the first and second administrations were .94 ($F = 46.63$, $p < .001$, $n = 96$, *95% C.I.* = .92 - .96) and .79 ($F = 12.40$, $p < .001$, $n = 42$, *95% C.I.* = .68 - .87), respectively. The R_{max} was .98 in both cases.

- The interrater reliability and agreement for the ACTFL OPIc® was consistent with the ACTFL OPI® for the test takers.

Table 7. Test taker feedback on the ACTFL OPI® and ACTFL OPIc®

				Percentage (%) of Responses				
	N	M	SD	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
I believe my performance on the ACTFL OPIc® accurately reflects my current speaking proficiency.	75	3.20	.87	4.0	12.0	49.3	29.3	5.3
I believe the ACTFL OPIc® is an effective way to measure English speaking proficiency.	75	3.35	.89	2.7	12.0	40.0	38.7	6.7
I would recommend taking an ACTFL OPIc® to a friend who needs their speaking proficiency assessed.	75	3.31	.87	2.7	13.3	40.0	38.7	5.3
I believe my performance on the ACTFL OPI® accurately reflects my current speaking proficiency level.	70	3.46	.94	4.3	8.6	34.3	42.9	10.0
I believe the ACTFL OPI® is an effective way to measure English speaking proficiency.	70	3.66	1.00	5.7	5.7	21.4	51.4	15.7
I would recommend taking an OPI to a friend who needs their speaking proficiency assessed.	70	3.64	1.00	5.7	5.7	22.9	50.0	15.7
I thought it was more difficult to demonstrate my speaking proficiency via the computer (ACTFL OPIc®) than with a live interviewer over the telephone (ACTFL OPI®).	69	3.41	1.05	5.8	14.5	23.2	46.4	10.1
Both the computer and telephonic interviews provided an adequate opportunity for me to demonstrate my speaking proficiency.	69	3.25	.76	1.4	13.0	46.4	37.7	1.4
The ACTFL OPIc® was more user friendly than the ACTFL OPI®.	69	2.70	.96	7.2	39.1	34.8	14.5	4.3
The ACTFL OPIc® provided a better opportunity for me to demonstrate my speaking proficiency.	69	2.90	.91	7.2	21.7	47.8	20.3	2.9
I preferred the testing format with a live interviewer than with the Avatar.	69	3.70	.91	2.9	7.2	21.7	53.6	14.5
It was easier to understand questions from a live interviewer than from the Avatar.	69	3.30	.91	2.9	14.5	39.1	36.2	7.2
I felt more comfortable recording my answers on the computer than providing answers to a live interviewer.	69	2.71	.94	8.7	33.3	39.1	15.9	2.9

- The correlations between the ACTFL OPI® (only administered once) and ACTFL OPIc® (first administration) were significant (r = .92, p < .001; R = .91, p < .001) and indicate a strong positive relationship between the assessments.

- Confirmatory Factor Analysis (CFA) results provide further construct validity evidence for the ACTFL OPI® and ACTFL OPIc. The fit statistics were excellent, the latent correlation between the two assessments was .94, and the construct validity coefficient was .99.

- Although slightly lower than desired, the absolute agreement results were sufficient for the development of the assessment. For example, the final ratings of the ACTFL OPI® and ACTFL OPIc® (first administration) agreed for 63% of the participant cases. Of note, for the first administration of the ACTFL OPIc®, the agreement between final ratings jumped to 85% within the major categories (novice, intermediate, advanced) and to 98% when the major category boundaries were ignored and agreement was defined as an exact match or being off by +/- one step.

- The order of assessment administration had no impact on the results.

- The results suggest that test takers who are indicating the most basic level of proficiency are being underestimated by the ACTFL OPIc®. Of the 26 disagreements at level one, 85% (22) resulted from cases where the ACTFL OPIc® final rating provided an underestimation of the ACTFL OPI® final rating. This negatively impacted the agreement between the two assessments. Since the ACTFL OPIc® is drive by test taker self-assessment and the ACTFL OPI® is driven by tester assessment, it suggests that some test takers are under assessing.

- Regardless of coefficient, a strong degree of relationship was found between ACTFL OPIc® (first administration) and ACTFL OPIc® (second administration) final ratings (r = .94, p < .001; R = .91, p < .001), providing evidence for test-retest reliability of the ACTFL OPIc®.

- CFA results provide further reliability and validity evidence across the two administrations of the ACTFL OPIc®. The latent correlation between the first and second administrations of the ACTFL OPIc® was .97, providing evidence of a very strong relation between the constructs. The R_{Max} and R_{CV} coefficients for the ACTFL OPIc® (first administration) were .98 and .99, respectively. The R_{Max} and R_{cv} coefficients for the ACTFL OPIc® (second administration) were .98 and .99, respectively.

- Although not the focus of the study, participants indicated a preference for the ACTFL OPI® over the ACTFL OPIc®.

Study 1 Issues

Although the evidence was positive overall for the ACTFL OPIc®, there are four issues that must be addressed to improve the assessment.

First, the absolute agreement was slightly lower than desired between ratings across the three rating positions for each assessment. As mentioned, obtaining high levels of absolute agreement between more than two raters is difficult to achieve. However, when we calculated absolute agreement for each of the three rater pairs (i.e., rater 1 and rater 2, rater 1 and rater 3, rater 2 and rater 3) on the first administration of the ACTFL OPIc®, we found the agreement to range between 71% and 76%. Although this is acceptable on a new assessment, the assessment should strive for absolute agreement of 80% or higher when the two-rater protocol is used.

Fortunately, the slightly-lower-than-desired absolute agreement did not impact the interrater reliability or maximum reliability coefficients. This suggests that the disagreements between raters were small in magnitude and consistent in direction. Additionally, since the final rating was assigned based on two-out-of-three raters, validity evidence was not impacted. Historically, the interrater reliability of ACTFL assessments

has been well above .80 (Surface & Dierdorff, 2003). This issue should be addressed and monitored periodically.

Second, the absolute agreement between the final ratings produced by assessments was sufficient for test development and piloting but should be slightly higher for use. For initial use, we would like to see a standard of 70% concordance achieved in a second study. This would be in line with the recommendations of Jackson (1998, 1999). Ideally, we would like to see the concordance of final ratings between the ACTFL OPI® and ACTFL OPIc® reach 84% (5 out of 6).

Fortunately, the lower-than-desired absolute agreement between the ACTFL OPI® and ACTFL OPIc® final ratings did not impact the reliability and validity coefficients. However, this issue should be addressed since the ACTFL OPI® and ACTFL OPIc® should produce the same rating a high percentage of the time in order to be considered alternative forms of the same assessment.

One potential cause of low concordance mentioned in Jackson (1999) that might be operating here is test taker motivation—that is, these test takers knew this was for research not administrative purpose. Addressing the next issue—underestimation or underreporting proficiency—may eliminate another potential cause of the concordance issue.

Third, when a test taker self assesses at level one during the ACTFL OPIc® background questionnaire, the ACTFL OPIc® appears to be susceptible to underestimating the speaking proficiency of the test taker. For most of the disagreements between the final ratings of the ACTFL OPI® and ACTFL OPIc® (first administration), the ACTFL OPIc® underestimated the test taker's proficiency, and the majority of these disagreements occurred at self-assessment level one.

This suggests that some individuals with higher proficiency are choosing level one and the ACTFL OPIc® does not provide these individuals with enough opportunities to demonstrate higher levels of speech, which relegates them to receiving an underestimate of their speaking proficiency. Although this is primarily an issue of individuals underestimating (poor self-assessment) or underreporting (intentionally trying to game the system) their proficiency, the process should have mechanisms to mitigate this issue.

In support of the underestimation hypothesis, several of the ACTFL OPIc® cases were marked as most likely underestimating the speaking proficiency of the test taker by the raters. In other words, the performance on the level one prompts was so strong that raters believed the test taker to be capable of producing a higher level of speech. However, because the speech sample was constrained at the lower level, there was no data or empirical evidence to justify or support a higher rating.

As an experiment, we adjusted the final ratings of the ACTFL OPIc® by increasing each of these marked cases by one proficiency rating on the ACTFL scale. This increased the levels of concordance (absolute agreement) between the final ratings to levels that are more than acceptable for our current purposes: 76% between the ACTFL OPI® and ACTFL OPIc® (first administration), 81% between the ACTFL OPI® and ACTFL OPIc® (second administration), and 83% between the ACTFL OPIc® (first administration) and ACTFL OPIc® (second administration). This suggests that the concordance issue is in part a function of underestimation or underreporting.

Another hypothesis might involve the interaction of an individual difference with the self-assessment choice. This investigation is beyond the current scope of this report, but we plan to examine this possibility in the future. We did assess basic demographic variables and process characteristics to determine if they impacted the self-assessment. For example, in terms of process characteristics, the order of administration (ACTFL OPI® first verse ACTFL OPIc® first) did not impact the self-assessment or the ratings.

The only demographic variable that had a

statistically significant relationship with the self-assessment was gender. Significantly more women tended to self assess at level one than men. However, this finding must be taken with extreme caution because the sample consisted of a majority of women and the organizational context is unknown to us. For example, women may occupy jobs that require less language proficiency. Therefore, the gender finding could be the result of gender's relationship to other structural, contextual factors in the organization, not to the ACTFL OPIc® self-assessment. However, that said, this should be investigated in future research to ensure it is not an issue.

Fourth, the user reactions to the ACTFL OPIc® were not as favorable as would be desired. Luckily, the study participants provided feedback and many suggestions for improving the interface. Of note, the respondents did not like the Avatar. Animation and Avatars are very prevalent and sophisticated in the Korean culture. Results suggest that the quality of the avatar was a major source of dissatisfaction for the Korean participants.

Study 1 Recommendations

We have five main recommendations to improve the ACTFL OPIc® process based on the findings of the validation study.

First, the self-assessment process needs to be adjusted. It is likely one of the major causes of the lower-than-desired absolute agreement between ACTFL OPI® and ACTFL OPIc® final ratings. One suggestion might be to rescale the self-assessment to include more proficiency levels. Another might be to give the people who self-assess in the lowest level the assessment protocol for the next level up—this has some potential negatives as well (i.e., test taker frustration for people who are truly at the lowest level). Providing audio samples of the proficiency levels to which test taker can listen will help with proficiency estimation. Finally, clarifying the instructions to make sure test takers accurately self-assess may also help. Some participants may be underreporting their proficiency because they believe this course of action will lead to an easier test and a higher score, whereas it actually leads to an easier test which caps their opportunity to receive a higher score.

Second, the core ACTFL OPIc® assessment should provide sufficient opportunities or prompts to demonstrate speaking proficiency regardless of the individual's self-assessment. So, extending or expanding the core assessment may help to address the agreement issue and reduce the number of ACTFL OPIc® samples that are rated "Not Ratable" (i.e., not a sufficient sample of speech to be rated). Additionally, providing the opportunity for test takers to request additional prompts after the core assessment is completed might be useful as well, although this would likely lead to non-standardization in the length of the assessment across test takers. Therefore, this decision should be made carefully.

Third, we recommend additional training on rating the ACTFL OPIc® for all raters. This should help improve the agreement and continue to maintain the high interrater reliability. We recommend periodic monitoring of rater agreement and re-norming of raters for who drop below the 80% agreement threshold in a two rater system. ACTFL has well-established processes in place for training, certification, and managing ACTFL OPI® raters. We recommend that these be modified for ACTFL OPIc® raters. Additionally, we recommend conducting rater diagnostic research and training effectiveness research to improve the rater training.

Fourth, we recommend that ACTFL and LTI continue to engage in an iterative development process using empirical data (such as this validation study) to improve the ACTFL OPIc®. We recommend conducting a second validation study if possible to assess the impact of any changes made based on our recommendations.

Finally, the user feedback suggests that the user interface needs to be improved, especially the quality of the avatar (technically an embodied agent according to literature definitions). We suggest making every effort to improve the look and feel of the testing experience. The will likely translate into more favorable impressions of the assessment, although likely does not

impact psychometric qualities of the assessment. It would be interesting to study the impact of having an avatar verse a video of a human interviewer on construct measurement.

Study 1 Limitations

There were several limitations of this study that need to be mentioned.

First, there was a technical issue with the post-assessment data collection and a number of open-ended responses were lost. Therefore, the open-ended responses in *Appendix C* might not be representative. Additionally, we do not know if technical issues prevented other participants from responding to the post-assessment survey. Second, the logistics of the study were carried out by another organization in Korea and were out of our direct control. Third, the sample size was small-to-moderate for this type of field research, which may have impacted our ability to detect significant findings and may limit the generalizability of our findings.

Fourth, the sample was skewed in terms of the proficiency of the study participants. Most of the individuals were at the Novice level. There were very few individuals at the higher proficiency levels. Future research should investigate the ACTFL OPIc® with a broader range of proficiency.

Fifth, we do not know to what extent the "artificial" nature of the study (i.e., test-takers typically do not take both assessments nor do they take pre- and post-assessments and the test are usually for some kind of administrative decision as opposed to research) impacted the responses of individuals, especially on the pre- and post-assessment surveys. As with all research, it was possible that the study influenced the behavior being studied.

Sixth, we required that the speech samples be rated by three individuals instead of the usual two raters to facilitate the use of CFA analyses for construct validity. Although this did not affect the speech sample or the quality of the individual ratings, it had an impact on the results of the study because it is more difficult to obtain high levels of absolute agreement with more than two raters.

Seventh, the findings may be idiosyncratic to this particular sample, organization, culture, or language, and may not generalize to other groups, testing purposes, or languages. Therefore, additional validation studies are needed until a sufficient body of evidence supporting the ACTFL OPIc® across multiple contexts has accumulated. We recommend that ACTFL pursue a validation study with each new population or language to which the ACTFL OPIc is extended until the stability of the modality is established.

Study 1 Conclusion

This study was design to investigate the ACTFL OPIc® for the assessment of English speaking proficiency. Although there are a few areas for improvement, the initial evidence of validity and reliability are impressive and support the initial use of the ACTFL OPIc®.

Since developing an assessment is an iterative process, any changes, that are made based on the *ACTFL OPIc® Validation Study 1* findings, should be researched in the future, especially if the change is sufficient enough to influence the construct and its psychometric properties.

ACTFL OPIc® Validation Study 2

The *ACTFL OPIc® Validation Study 1* provided strong initial evidence for the validity and reliability of the ACTFL OPIc® as an assessment of English speaking proficiency in the Korean population. However, four areas for improvement were identified: (1) increasing the agreement between raters (interrater reliability was high but agreement was slightly lower than desired); (2) increasing the concordance between ACTFL OPI and ACTFL OPIc® final ratings for the same individuals (correlations were high between the two assessments but concordance was slightly lower than desired); (3) test takers were able to underestimate or underreport their proficiency and impact the test outcome (this underestimation is likely one of the causes of the

concordance issue); and (4) improving the user interface, especially the avatar.

Following the initial validation study, ACTFL and LTI reviewed the evidence-based recommendations and instituted some of the suggested modifications to the ACTFL OPIc® assessment protocol. Specifically, they improved the self-assessment protocol to mitigate underestimation and underreporting of proficiency, they modified the core test, they provided ACTFL OPIc® raters with additional training, and they switched from an avatar based system to a video recording of a human interviewer asking the question.

ACTFL OPIc® Validation Study 2 is part of the continuing efforts to accumulate psychometric evidence in support of the ACTFL OPIc® as an assessment of English speaking proficiency. Up front, a number of constraints with this study should be noted. The second study was conducted under impromptu conditions and, therefore, many of the design features incorporated into the first study (e.g., adding a third rater for CFA) were not feasible given the client's timetable, access to participants, and available funding.

Despite these constraints, we believe this study has value because it was conducted under more realistic testing and rating conditions and followed modifications to the ACTFL OPIc® recommended by the previous study. Although not as rigorously controlled as the study 1, the results of this field study can be used to infer whether or not the modifications had an impact on the functioning of the assessment. Validation research is an on-going effort to collect evidence on the psychometrics of an assessment. This study is another strand of evidence. Additional studies should be conducted to add to the foundation of evidence supporting the ACTFL OPIc® testing modality.

Study 2: Research Questions

The purpose of the *ACTFL OPIc® Validation Study 2* was to address the research questions in *Table 8*. These questions were addressed according to the methods described in the next section.

Table 8. Study 2 Research Questions

RQ1.	*What is the overall interrater reliability and consistency of the ACTFL OPIc®?*
RQ2.	*How does the interrater reliability and consistency of the ACTFL OPIc® compare to that of the ACTFL OPI® for the same sample of test takers?*
RQ3.	*What is the relationship between ACTFL OPIc® and ACTFL OPI® final ratings?*
RQ4.	*What is the absolute agreement between ACTFL OPIc® and ACTFL OPI® final ratings?*

Study 2: Method

Participants

The participants for the *ACTFL OPIc® Validation Study 2* were selected from the same Korean workforce that was studied in Study 1. A total of 27 individuals participated in Study 2.

Of these participants, 21 completed the pre-assessment survey and, therefore, additional information is available about these participants. Twenty males (95.2%) and one female (4.8%) completed the pre-assessment survey. Most participants (52.4%) indicated that the highest level of education they had completed was a B.A. or B.S. degree while 38.1% indicated that they had completed a M.A. or M.S. degree.

In terms of work experience, 42.9% of the participants had worked in their current job for 11-20 years, 38.1% indicated 6-10 years, and 19% indicated 1-5 years. Approximately 67% of participants indicated serving in a supervisory role in their current job. The majority of participants (76.2%) reported that they do use English as part of their job. In addition, the majority of participants (76.2%) indicated that they had to speak with people via the telephone with whom they had not had previous contact.

Participants were asked several questions on the pre-assessment survey about their experiences using the telephone and computers. The majority of participants indicated that they had never taken part in a telephonic job interview (95.2%). Approximately half of the participants (47.6%) indicated that they have taken a test via the telephone. Approximately 48% of the participants indicated that they had been using computers for 11-20 years, while 38.1% indicated that they had been using computers for 6-10 years. A majority of respondents had never applied for a job on the internet (76.2%), but all participants who responded to the pre-assessment survey had taken an online course. Approximately 76% of participants had taken a language course online, but less than half (42.9%) had taken a test on the internet.

Most participants indicated that they were required to use the internet as part of their job (90.5%) and that they use online messaging (81%). Participants indicated a wide range of internet usage at work. 28.6% indicated using the internet for less than one hour at work, 28.6% indicated 1-2 hours, and 28.6% indicated more than 5 hours. Most participants (66.7%) indicated using the internet at home between one and two hours in a typical day.

Participants were also asked questions about their previous English training/education and their previous experience with English testing. All participants indicated that they first started to study English in primary school (14.3%) or middle school (85.7%). There was some variability in terms of the number of English courses that individuals had taken either at school or through private institutes. Most participants (66.7%) indicated taking between one and three courses, although 19% indicated that they had taken 10 or more courses.

In terms of experience with English testing, the majority of participants had never taken an ACTFL OPI® (95.2%). All participants had taken the Test of English as a Foreign Language (TOEFL) or the Test of English for International Communication (TOEIC).

Study Design

The *ACTFL OPIc® Validation Study 2* was designed to mimic a realistic ACTFL OPIc® testing situation as much as possible. The same pre-assessment survey that was developed for use in Study 1 was administered to participants in Study 2 prior to the administration of the two ACTFL assessments (i.e., ACTFL OPI® and ACTFL OPIc®). All participants took the ACTFL OPIc® first and then took the ACTFL OPI® 24-48 hours later.

Since Study 1 demonstrated that the order of test administration did not impact the relationship between the two assessments, we decided to administer the ACTFL OPIc® prior to the ACTFL OPI® in order to ensure that the ACTFL OPIc® testing experience was as realistic as possible. In other words, the participants would be experiencing the ACTFL OPIc® as a typical participant would with no ACTFL OPI® as a referent prior to ACTFL OPIc® administration. Participants did not complete a post-assessment survey in Study 2 because of logistical issues and participant time constraints.

Rating Speech Samples

After participants completed all assessments, the ACTFL OPI® and ACTFL OPIc® samples were rated. Since the validity and reliability of any rater-based assessment is a function of the raters, the selection of raters and the design of the rating protocols were very important considerations.

Raters. Raters were drawn pool of fifteen ACTFL OPIc® raters and ACTFL OPI® testers who were randomly selected to rate ACTFL OPIc® and ACTFL OPI® assessments. They followed the standardized protocols for each assessment. Some raters/testers differed in experience level.

Rating Protocol. For Validation Study 2, all raters followed the same rating protocols (i.e., ACTFL guidelines) they would normally follow for the ACTFL OPI® and ACTFL OPIc® assessments. Unlike the first validation study—which used three raters per assessment per person—the second study used two raters per assessment, which is the standard ACTFL

process. A third rater was only used to arbitrate disagreements, as the protocol dictates. Interrater absolute agreement is typically calculated on the initial two raters because only disagreements use a third rater.

Measures

ACTFL OPI®. A description of the ACTFL OPI® is provided in the *Study 1: Method* section.

ACTFL OPIc®. A description of the ACTFL OPIc® is provided in the *Study 1: Method* section. However, several changes were made to the user interface as a result of recommendations from Study 1. One change was related to the delivery of prompts. In Study 1, an avatar was used to deliver prompts. In Study 2, a video of a live person was used instead of the avatar. Additionally, modifications of the proficiency self-assessment, such as recorded examples of proficiency for participants to review were added.

Pre-Assessment Survey. The same pre-assessment survey that was developed for use in Study 1 was also used in Study 2. A description of the ACTFL OPI® is provided in the *Study 1: Method* section.

Analytic Procedures

Data were received from LTI and the survey contractor and were cleaned, formatted, and aggregated for analysis. The method used to address the research questions is presented by research question.

RQ1 and RQ2. Interrater reliability was calculated using Pearson's correlation (*r*), Spearman rank-order correlation (*R*), and Goodman-Kruskal's gamma (G). These coefficients were selected because they have been reported in previous research on the interrater reliability of the rater-based speaking proficiency measures (e.g., Surface & Dierdorff, 2003). Because of the use of only two raters and the small sample size, interclass correlations (ICC) could not be calculated, and confirmatory factor analysis (CFA) models could not be estimated. Interrater consistency (absolute agreement) was calculated between the rater one and rater two positions. The ACTFL process arbitrates disagreements by using of a third rater.

RQ3. We computed three correlation coefficients (Pearson's *r*, Spearman's *R*, and Goodman-Kruskal's G) between the final ratings obtained from the ACTFL OPI® and ACTFL OPIc® in order to determine the relationship between these ratings. Again, because of the use of only two raters and small sample size, latent correlations from CFA models could not be estimated.

RQ4. To determine the absolute level of agreement between the ACTFL OPI® and ACTFL OPIc®, the percentage of cases in which the final ACTFL OPI® ratings agreed with the final ACTFL OPIc® ratings was calculated.

Study 2: Results

This section presents the findings for the *ACTFL OPIc® Validation Study 2* by research question.

RQ1: What is the overall interrater reliability of the ACTFL OPIc®?

To assess the interrater reliability for the ACTFL OPIc®, Pearson's correlation (*r*), Spearman rank-order correlation (*R*), and Goodman-Kruskal's gamma (G) were calculated for the two ACTFL OPIc® rater positions. All three estimates of interrater reliability (r = .86, p = .00; R = .85, p = .00; G = .93, p = .00) suggest acceptable interrater reliability for the ACTFL OPIc®. Rater 1 and rater 2 positions on the ACTFL OPIc® agreed exactly 58% of the time. Although this is lower than desired, it is acceptable given the limited rater experience with the assessment (at this point) and that the process allows for a third rater who will assign an independent rating breaking the stalemate between rater one and two (without knowing he or she is the third rater).

RQ2: How does the interrater reliability of the ACTFL OPIc® compare to that of the ACTFL OPI® for the same sample of test takers?

The interrater reliability coefficients for the ACTFL OPI® indicated an acceptable level of reliability (r = .93, p = .00; R = .89, p = .00; G = .98, p = .00). These are consistent with the

levels of interrater reliability previously reported for the ACTFL OPI®. The interrater reliability results for the ACTFL OPIc® appear to be very similar to the ACTFL OPI® for the same group of test takers, although the ACTFL OPI® has slightly more robust coefficients. Additionally, the absolute agreement for raters one and two on the ACTFL OPI® was higher than that of the ACTFL OPIc®, likely because of more rating experience with the ACTFL OPI® format. Both assessments appear to have sufficient reliability for testing purposes.

RQ3. What is the relationship between ACTFL OPIc® and ACTFL OPI® final ratings?

To determine the relationship between the ACTFL OPIc® and ACTFL OPI®, correlations between the final ratings of the ACTFL OPI® and the ACTFL OPIc® were calculated. The correlations between the ACTFL OPI® and ACTFL OPIc® were significant (r = .97, p = .00; R = .95, p = .00) and indicate a strong positive relationship between the assessments. These correlations are consistent and slightly higher in magnitude than the correlations found in study 1 (r = .92, p = .00; R = .91, p = .00), indicating a slightly stronger relationship between the assessments.

RQ4. What is the absolute agreement between ACTFL OPI® and OPI-C final ratings?

In terms of the absolute agreement or concordance between the ACTFL OPI® and ACTFL OPIc®, the final ratings of the ACTFL OPI® and ACTFL OPIc® agreed for 87% of the participant cases. This level of concordance exceeds the 70% concordance level recommended for use (Jackson, 1998, 1999). Additionally, 100% of the cases agreed exactly or were in +/- one rating within the same major boundary (no disagreements crossed major boundaries). This indicates that disagreements between ACTFL OPI® and OPIc® final ratings were minor when they occurred. Overall, the relationship between the final ACTFL OPI® and ACTFL OPIc® final ratings was robust.

Study 2: Discussion

The *ACTFL OPIc® Validation Study 2* provided a second investigation of the psychometric properties of the ACTFL OPIc® as an assessment of speaking proficiency in English. The goal of this study was to continue accumulating evidence on the reliability and validity of the ACTFL OPIc®. Additionally, several modifications were made to the ACTFL OPIc® following Study 1—including improving the self-assessment protocol and replacing the animated avatar with video of a human interviewer—and data needed to be collected on these modifications.

To this end, a small sample of Korean employees completed both the ACTFL OPIc® and OPI®. As with Study 1, the second study yielded evidence of validity and reliability for the ACTFL OPIc®. Although there were some limitations of this study and a few issues remain to be addressed, this study should be viewed as providing additional initial evidence supporting the ACTFL OPIc® for commercial use. The next sections summarize the key findings, discuss several issues, and provide a couple of recommendations to test developers.

Summary of Key Findings

This section highlights the key finding for the second validation study. The results section provides a complete reporting of the findings.

- The interrater reliability coefficients for the ACTFL OPIc® were found to range between .86 and .93, which continues to provide evidence for the reliability of the assessment. The reliability results were consistent with those of the ACTFL OPI®.

- The absolute agreement between ACTFL OPIc® raters continues to be lower than desired. Fortunately, the process includes a third rater whose rating is used to break the stalemate between the initial raters and decide the final rating.

- The relationship between the ACTFL OPI® and OPIc® was found to be robust with validity coefficient of R = .95 and r = .97.

These coefficients along with the validity coefficients and CFA results from Study 1 support the expert judgment (content validity evidence) and rational argument (same construct definition, testing/rating protocols and content domains) that both assessments are measuring the same construct.

- The absolute agreement (concordance) between the final ratings of the ACTFL OPI® and OPIc® was found to be greatly improved over Study 1. The final ratings of the two assessments agreed 87% of the time (and 100% of the time if you count being off by -/+ one step within the same major boundary), which more than exceeds the 70% standard set by Jackson (1998, 1999). It also exceeds the 84% standard that we would like to see for high-stakes use.

- There was no evidence to suggest any underestimation of proficiency, which suggests that the underestimation of proficiency on the self-assessment issue may have been resolved by the changes. However, the small sample size prevents definitive investigation of this issue.

Study 2 Issues

Study 2 provides strong additional evidence for the use of the ACTFL OPIc®. However, one issue continues to remain, the lower-than-desired interrater agreement between rater one and rater two. The interrater reliability remains high because the disagreements are few and small in magnitude, but we would like to see absolute agreement between the raters of 80% or higher. Surface & Dierdorff (2003) reported that the ACTFL OPI® achieved this standard across all languages. Therefore, it is not impossible.

Fortunately, the process is designed to deal with disagreements between the initial two raters by using a third rater to break the stalemate. Since the concordance of final ratings between the ACTFL OPI® and OPIc® is fairly high, the process is working and prevents the lower interrater agreement between ACTFL OPIc® raters one and two from impacting the test taker's final rating.

The real impact is for the test provider and the client in terms of cost. The more third ratings are required the higher the cost of assessments will have to be. It is in the test provider's best interest to keep the number of tests needing arbitration to a minimum. Additionally, it would be impossible to justify using a single rater system for high stakes testing with this level of interrater agreement.

The reliability and validity of ratings-based assessments depends on how well the rating model fits the construct definition, how well the model is internalized by raters, and how consistently it is applied across raters. This requires effective training, a rigorous certification process, supervised practice as a rater, and periodic re-norming of raters. ACTFL has all these processes in place.

ACTFL raters are trained using a frame-of-reference (FOR; Bernardin & Buckley, 1981) training technique (Dierdorff, Surface, & Brown, 2008). FOR is a prevalent rater training protocol in which the primary goal is to "train raters to share and use common conceptualizations of performance [any construct/behavior of interest] when making evaluations" (p. 525; Woehr, 1994). This approach ensures that the raters have a "shared mental model" of the construct, a standardized testing protocol, and apply them consistently. FOR training is very effective across a variety of contexts, including training ACTFL raters (for details see Dierdorff et al., 2008).

The lower-than-desired absolute agreement between the raters on the ACTFL OPIc® may be related to the newness of the assessment, the newness of the associated rater training, and/or the lack of rating opportunities to date because of the relatively few tests administered at the time of the studies. The ACTFL OPIc® had only been administered for the pilot study prior to the validation studies. These studies were literally the first rating opportunities on the ACTFL OPIc®. As mentioned earlier, ACTFL has all the processes in place to ensure high-quality ratings. The absolute agreement should improve quickly

as time and rating opportunities allow for these processes to take effect.

Study 2 Recommendations

Based on the results of study 2, we offer the following recommendations to ACTFL and LTI.

First, we suggest all ACTFL OPIc® raters should be sufficiently trained on ACTFL OPIc® samples with a rigorous certification process that focuses on demonstrating agreement with the expert rating at the 84% or higher level. Regardless of whether an individual is a certified ACTFL OPI® tester, the person should still receive training and certification on the ACTFL OPIc® because the different interview environment.

Second, we suggest that all new ACTFL OPIc raters start out providing "shadow" ratings until they are agreeing above 84% of the time with the expert raters. This will ensure high agreement and less need for third rater arbitration and additional cost.

We are suggesting 84% as a standard here strictly because it limits the number of third rater arbitrations to 1 in every 6 ACTFL OPIc® assessments. However, LTI and ACTFL are free to adjust the number if they want to have less or more third rater arbitrations. Because of the process specifies a third rater to resolve disagreements, the issue is more related to cost than to the assessment's psychometrics.

Third, we suggest that ACTFL improve the ACTFL OPIc® rater training as planned. The initial rater training materials were limited to the samples from the pilot study. Now, there should be a much larger pool of samples to use for training and certification.

Fourth, we recommend revisiting the interrater agreement every three months until it is consistently above 84% or a different standard set by LTI and ACTFL.

Fifth, rater diagnostic studies should be conducted to determine if any raters or rater pairs are consistently lower in terms of agreement. This could be used to pinpoint raters for additional training or re-norming or to make rater assignments. This could also be used to improve rater training processes.

Sixth, as ACTFL and LTI extend the ACTFL OPIc format to other languages and countries/cultures, additional validation studies need to be conducted to assess test function under these new conditions.

Finally, if ACTFL and LTI ever plan to use the ACTFL OPIc® as a single rater assessment and it is used for high-stakes testing, then the raters utilized should be functioning above 84% agreement with other raters consistently on two-rater assessments, with closer to 100% being desirable. Periodic checks of the single ratings for accuracy should be made to ensure rating accuracy.

For a rater-based assessment to be effective, the raters must have a "shared mental model" and apply it uniformly and consistently. The results suggest that the ACTFL OPIc® raters are doing well but there are areas for improvement. These recommendations are offered in the spirit of continuous improvement.

Study 2 Limitations

This study had a number of limitations because it was an impromptu follow up to Study 1. Despite these limitations, Study 2 has value because it provides additional validity and reliability evidence for the ACTFL OPIc® and provides insights into improving the assessment. It also provides evidence of assessment functioning after the recommended modifications were made.

First, the sample size was very small, which limits our use of advanced statistical techniques and potentially limits the accuracy and generalizability of our findings for the larger population of test takers in Korean. Because of the small sample, the range of proficiency on the ACTFL scale is not fully represented in Study 2. Also, the Study 1 and Study 2 samples are somewhat different on several demographic variables, such as gender. For Study 1, the majority of participants were women, whereas

for Study 2 the participants were overwhelming male.

Second, unlike the first study, there were no experimental controls, such as random assignment or counterbalancing. However, we do not believe the lack of experimental controls impacted the findings. The first study demonstrated that order of administration was not a factor, so counterbalancing was not strictly needed. In the second study, we chose to have all participants take the ACTFL OPIc® first to standardize the testing experience and for the ACTFL OPIc® to be as consistent with real testing conditions as possible, even though, this is an artificial study.

Third, the participants knew the assessments were for test development purposes and may not have been motivated to give their maximum performance.

Fourth, there was no post-assessment. This was a logistical issue with the Korean organization. It would have been useful to receive feedback on the use of the video of the human interviewer to contrast with the avatar in Study 1.

Fifth, the rater pool was not as experienced on the ACTFL OPIc® as they will be for "real" testing. Both Study 1 and Study 2 were conducted as part of the development process, and there were few samples for rater practice and norming available prior to these studies (i.e., pilot study mentioned in Study 1). The effectiveness of the raters should improve greatly over time.

Finally, the sample size and the use of only two raters (unless there was a disagreement and a third was used) prevented us from using techniques such as ICCs and CFA. This limited our ability to make direct comparison with Study 1 in some cases.

Study 2 Conclusion

The findings from Study 2 provide additional support for the ACTFL OPIc® as a measure of English speaking proficiency. The validity and reliability coefficients in Study 2 continued to exceed sufficient levels. The Study 2 absolute agreement between the final ratings of the ACTFL OPI® and OPIc® improved to well above the minimum standard recommended by Jackson (1998, 1999). Although the interrater agreement between raters one and two could be higher, the results for Study 2 continue to support the initial use of the ACTFL OPIc®. Despite the small sample size and other limitation of Study 2, we believe it contributes to our understanding of the ACTFL OPIc®.

General Discussion for Studies 1 and 2

Two studies were conducted with Korean employees to investigate the psychometric properties of the ACTFL OPIc® as an assessment of speaking proficiency in English. Although several areas for improvement were identified, the results of both studies support the use of the ACTFL OPIc® for initial commercial testing in Korea. However, the test publisher should continue to take an action research perspective to test improvement and quality assurance. Additional studies should be conducted as the assessment is rolled out in Korea, used with different populations, or used with different languages.

Summary of Findings

Both of the studies provided evidence on the validity and reliability of the ACTFL OPIc®, and the specific results can be found in the respective results and discussion sections of each study. Additionally, ACTFL and LTI should pay close attention to the recommendations provided after each study.

Reliability evidence was provided in two forms—*reliability as consistency* (e.g., interrater reliability) and *reliability as repeatability* (test-retest reliability). Both studies provided evidence of interrater reliability and consistency (rater agreement) for the ACTFL OPIc®. Because of the use of CFA in Study 1, an additional measure of reliability as consistency, maximum reliability (R_{max}; Drewes, 2000), could be and was calculated. Study 1 had a test-retest reliability component as well.

Overall, the results suggest sufficient levels of reliability were achieved for initial use. The ICCs, R_{max}, and other reliability indices were typically above .90. Of note, the absolute agreement between raters on the ACTFL OPIc® was lower than desired. However, as discussed, this will likely correct itself in the future and is less of a concern because the ACTFL process incorporates a third rater to arbitrate disagreements between the initial raters, yielding an accurate final rating. However, ACTFL and LTI need to monitor interrater reliability and agreement periodically because a drop in agreement is one of the first signs of a problem. Plus, the more third rater arbitrations used, the higher the cost of the assessment program.

In terms of validity evidence for the ACTFL OPIc®, there were several types of evidence accumulated during the validation studies. First, it should be noted that validity was built into the ACTFL OPIc® by the rigorous designed process used by language testing experts and their strict adherence to the ACTFL speaking proficiency guidelines (Breiner-Sanders et al., 2000). Expert judgment and strict adherence to an underlying model or framework provide initial validity evidence for the assessment (what used to be referred to as content validity).

Validity evidence was provided in two forms by our studies—*evidence based on internal structure* (CFA results) and *evidence based on relations to other variables* (relationship with an established measure—the ACTFL OPI®). The results of the CFA in Study 1 provide strong support for the internal structure of the construct measured by the ACTFL OPIc®. The excellent fit statistics for the model and the high construct validity coefficient are prime examples. The CFA results also show a very strong latent correlation between the ACTFL OPI® and OPIc®, suggesting the assessments are measuring the same construct. The correlations (*R* and *r*) between the final ACTFL OPI® and OPIc® final ratings in both studies were very robust as well. Taken together, the results suggest that both assessments are measuring the same conceptualization of speaking proficiency regardless of the difference in interview mode.

The absolute agreement or concordance of the final ACTFL OPIc® ratings with the final ACTFL OPI® ratings is an important issue because the ACTFL OPIc® is considered to be a different delivery mode of the ACTFL OPI®, measuring the same construct and producing equivalent results.

For two rater-based assessments to be considered parallel, they must have the same construct definition, have the same test specifications, have the same test protocols, have similar reliabilities, and produce the same rating in direct comparison. This is a high standard to achieve as noted. If we were interested in an assessment that measured a similar construct or a different construct, we would not be discussing as rigorous a standard.

The results suggest that all of these standards have been met (same construct, specifications, and protocols, similar reliability coefficients, and concordance of final ratings). We have discussed all of these results in this section with the exception of absolute agreement between final ratings.

Jackson (1998, 1999) argued for a 70% minimum exact agreement standard for different modes of the same rater-based assessment. Although the exact agreement between the final ratings of the ACTFL OPI® and OPIc® missed the 70% standard slightly in the first study, the agreement percentage in the second study exceeded the 70% as well as our recommended 84% level (5 out of 6 agreements). We believe this is sufficient to justify initial use of the ACTFL OPIc®.

With increased rating training and practice with the new assessment, the agreement of final ratings should continue to improve or remain at sufficient. However, ACTFL and LTI should periodically equate the ACTFL OPIc® and OPI®. LTI also needs to monitor rater functioning closely. If two poor-performing raters are rating samples together and agreeing on an inaccurate rating, then it will likely decrease the concordance with the ACTFL OPI®. Poor-performing raters (relative to the others) should be identified early and corrective action taken.

Finally, although not a psychometric characteristic, the user reactions suggest that the Korean population had unfavorable views towards the avatar because of its low quality in comparison to typical avatars used in Korean virtual culture. This was such a strongly held opinion that the ACTFL OPIc® format changed to a video of a human interviewer in future implementations in Korea. It is unknown whether or not other groups would have the same reaction to the quality of the avatar. However, ACTFL should consider improving the quality of the avatar technology or implementing the video-based approach on all assessments. Regardless, ACTFL needs to monitor the reactions of the test takers to assessment and the avatar.

Future Research

This section contains some research ideas to consider. Future research on the ACTFL OPIc® should focus on the following areas:

- Continuing to assess the psychometric properties of the English version as used in Korean (e.g., periodic reliability studies);
- Assessing the psychometric properties of the English version with other cultures or in other countries (e.g., using the ACTFL OPIc® in China);
- Assessing the psychometric properties of versions of the ACTFL OPIc® in different languages (e.g., Spanish);
- Assessing the measurement equivalence of the ACTFL OPIc® format across languages and cultures/countries;
- Assessing user reactions to ACTFL OPIc® within each language or culture/country and between them (comparisons);
- Assessing the impact of any modifications to the ACTFL OPIc® format that might impact the psychometric properties of the assessment;
- Studying the impact of specific design features on measurement and user reactions (e.g., comparing the avatar verse human interviewer);
- Investigating the individual differences (other than their language proficiency) that might influence an individual's ACTFL OPIc® rating (e.g., computer test-taking anxiety).
- Studying raters to determine which raters are more effective and why (e.g., rater diagnostics).
- Studying the effectiveness rater training and certification processes.
- Studying effectiveness of raters who have been trained solely as ACTFL OPIc® raters verse ACTFL OPI® testers who are subsequently trained to rate ACTFL OPIc®.
- Assessing the relationship between the ACTFL OPIc® and other non-ACTFL assessments.

These are just a few of the research areas and ideas that can be explored.

Once sufficient validity and reliability evidence has been established for the ACTFL OPIc® format and it becomes a mature assessment (i.e., testing in multiple languages and cultures/countries with sufficient evidence of psychometric properties in each), we recommend periodic follow-up evaluations of reliability and validity every three years or whenever substantial changes are made to the assessment or rating protocol that might impact the measurement properties.

Updates

Since the validations studies were conducted, thousands of ACTFL OPIc® English assessments have been administered in Korea. This data will be used in future research to assess interrater reliability and consistency and rater functioning. A study is planned and the results will be published, along with the results of the initial validation studies.

ACTFL instituted a new ACTFL OPIc® rater selection and training program. Data have been collected from the program trainees, and they are being followed through certification to work as actual ACTFL OPIc® raters. This will allow for the connection of individual characteristics with training and certification outcomes and job performance as a rater, allowing us to validate and improve the rater program. The validity and reliability of rater-based assessments have their foundations in the mental models and subsequent rating behaviors of the raters. Therefore, selection and training of rater is critical.

ACTFL has developed a Spanish version of the ACTFL OPIc®, and we have been collecting data on this version of the assessment. When complete, the findings will be published in a technical report.

ACTFL is developing versions of the OPIc® in additional languages. Although no specific plans have been finalized, ACTFL intends to conduct validation studies for each additional language.

Conclusion

ACTFL and LTI should be commended for taking an empirical approach to test development and validation. Developing an assessment is an iterative process that should be guided by empirical evidence from test takers. Repeated measurements that yield high-quality data on the new assessment are required to generate recommendations for thoughtful decisions makers to adopt and implement to modify the assessment. Then, the process of collecting data for improvement starts again.

ACTFL and LTI have demonstrated support for this action research perspective of test development. We hope both organizations continue to support this approach in the future. Although initial evidence supports the use of the ACTFL OPIc® as a measure of English speaking proficiency in the Korea, we encourage ACTFL and LTI to continue their research and improvement efforts, especially as they move the assessment to new populations or to new languages. The initiatives in the update section suggest that both organizations will continue to support evidence-based improvement of the ACTFL assessments.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, DC: Author.

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

Bauer, T. N., Maertz, C. P., Jr., Dolen, M. R., & Campion, M. A. (1998). Longitudinal assessment of applicant reactions to employment testing and test outcome feedback. *Journal of Applied Psychology, 83*, 892-903.

Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. *Academy of Management Review, 6,* 205–212.

Breiner-Sanders, K. E., Lowe, P., Miles, J., & Swender, E. (2000). ACTFL Proficiency Guidelines—Speaking revised 1999. *Foreign Language Annals, 33*, 13–17.

Browne, M. W., & Cudeck, R. (1993). Alternate ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (p. 136-162). Thousand Oaks, CA: Sage.

Callaghan, G., & Thompson, P. (2002). 'We recruit attitude': The selection and shaping of routine call centre labour. *Journal of Management Studies, 39*, 233-254.

Cattell, R. B. (1988). The meaning and strategic use of factor analysis. In R. B. Cattell & J. R. Nesselroade (eds.), *Handbook of multivariate experimental psychology: Perspectives on individual differences*, 2nd ed. (pp. 131–203). New York: Plenum Press.

Chao, G. T., & Sun, Y. J. (1997). Training needs for expatriate adjustment in the People's Republic of China. In Z. Aycan (Ed.), *New approaches to employee management: Vol. 4 expatriate management: Theory and research* (pp. 207-226). Greenwich, CT: JAI Press.

Chittum, R. (2004, May 5). Rise in offshore jobs expected. *Wall Street Journal*, p. B6.

Dandonoli, P., & Henning, G. (1990). An investigation of the construct validity of the ACTFL Oral Proficiency Guidelines and Oral Interview Procedure. *Foreign Language Annals, 23*, 11–22.

Dierdorff, E.C., Surface, E.A., & Brown, K. (2008, August). *For whom is frame-of-reference training effective? The motivational role of goal orientation.* A paper to be presented at the 2008 Academy of Management conference in Anaheim, CA.

Downing, S.M., & Haladyna, T.M. (Eds.). (2006). *Handbook of Test Development*. Mahwah, NJ: Lawrence Erlbaum.

Drewes, D. W. (2000). Beyond the Spearman-Brown: A structural approach to maximal reliability. *Psychological Methods, 5*, 214-227.

Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist, 41*, 1040-1048.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement*, 3rd ed. (pp. 105–46). Washington, DC: American Council on Education.

Flanagan, J. C. (1951). Units, scores, and norms. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 695–763). Washington, DC: American Council on Education.

Flynn, L. J. (2003, December 8). Companies sending work aboard are learning cultural sensitivity—to their American customers. *The New York Times*, p. C4.

Hammers, M. (2005). Wyndham looks to leap language gap. *Workforce Management*, p. 17.

Hatcher, L. (1994). *A step-by-step approach to using the SAS system for factor analysis and structure equation modeling*. Cary, NC: SAS Institute Inc.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equations Modeling, 6*, 1-55.

Jackson, G. L. (1998). A score is a score. Isn't it? *The CALLer*, *5*(4). Center for the Advancement of Language Learning.

Jackson, G. L. (1999, February). *Oral proficiency testing modality study.* [Technical Report 99-01]. Presidio of Monterey, CA: Defense Language Institute Foreign Language Center Research and Analysis Division.

Kline, R. B. (1998). *Principles and practices of structural equation modeling.* New York: Guilford Press.

Kolen, M. J. (2004). Linking assessments: Concepts and history. *Applied Psychological Methods, 28*, 219-226.

Loehlin, J. D. (1992). *Latent variable models: An introduction to factor, path, and structural analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associations, Inc.

Magnan, S. S. (1987). Rater reliability of the ACTFL Oral Proficiency Interview. *The Canadian Modern Language Review, 43*, 267–76.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1998). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103*, 391-410.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30-46.

Millsap, R. E. (2002). Structural equation modeling: A user's guide. In F. Drasgow & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis* (pp. 257-301). San Francisco: Jossey-Bass.

Noe, N. A. (2005). *Employee Training & Development* (3rd ed.), Burr Ridge, IL: Irwin McGraw-Hill.

Pristin, T. (2003, October 8). One victim when jobs go overseas: U.S. office space. *The New York Times,* p. C8.

Potosky, D., & Bobko, P. (2004). Selection testing via the Internet: Practical considerations and exploratory empirical findings. *Personnel Psychology, 57*, 1003-1034.

Salgado, J. F., & Moscoso, S. (2003). Internet-based personality testing: Equivalence of measures and assessees' perceptions and reactions. *International Journal of Selection and Assessment, 11*, 194-205.

Sensory Modality Prefence Inventory. (2002). *Brookhaven College.* Retrieved on August 15, 2005 from http://www.brookhavencollege.edu/learningstyle/modality_test.html.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420-428.

Smith, R. A., & Frangos, A. (2004, June 2). Outsourcing likely to slow office rebound; trend undercuts demand for space, research finds; suburbs lose call centers. *The Wall Street Journal*, p. B4.

Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Author.

Stanley, J. C. (1971). Reliability. In R. L. Thorndike (ed.), *Educational measurement*, 2nd ed. (pp 356–442). Washington, DC: American Council on Education.

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

Swender, E. (2003). Oral proficiency testing in the real world: Answers to frequently asked questions. *Foreign Language Annals, 36*, 520-526.

Takeuchi, R., Yun, S., & Russell, J. (2002). Antecedents and consequences of the perceived adjustment of Japanese expatriates in the USA. *International Journal of Human Resource Management, 13*, 1224-1244.

Tanaka, J. S. (1987). 'How big is big enough? Sample size and goodness of fit in structural equation models with latent variables. *Child Development, 58*(1), 134-146.

Thompson, I. (1995). A study of interrater reliability of the ACTFL Oral Proficiency Interview in five European languages: Data from ESL, French, German, and Spanish. *Foreign Language Annals, 28*, 407-422.

Thompson, L. F., Surface, E. A., Martin, D. L., & Sanders, M. G. (2003). From paper to pixels: Moving personnel surveys to the Web. *Personnel Psychology, 56*, 197-227.

Thorndike, R. L. (1951). Reliability. In E. F. Lindquist (ed.), *Educational measurement* (pp. 560620). Washington, DC: American Council on Education.

Tucker, L. R. & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*(1), 1-10.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4-69.

Vandewalle, D. (1997). Development and validation of a work domain goal orientation instrument. *Educational and Psychological Measurement, 57*(6), 995-10151.

Vina, G., & Mudd, T. (2003, November 5). Call centers migrate to India, and north of England loses jobs. *Wall Street Journal,* p. 1.

von Eye, A., & Mun, E. Y. (2005). *Analyzing rater agreement: Manifest variable methods*. Mahwah, NJ: Lawrence Erlbaum Associates.

Weber, G. (2004). English RULES. *Workforce Management, 8*3, p. 47-51.

Woehr, D. J. (1994). Understanding frame-of-reference training: The impact of training on the recall of performance information. *Journal of Applied Psychology, 79*, 525–534.

Appendix A

Pre-Assessment Survey

Section A: Survey ID

1. Last name/Family name
   Fill in the blank

2. First name
   Fill in the blank

3. Survey ID
   Fill in the blank

Section B: Demographics

Directions: The following items are intended to gather important information about your background. Please read each question carefully and choose the appropriate response.

1. Please indicate your date of birth.
   Fill in the blank

2. Please indicate your gender.
   1 = Male
   2 = Female

3. Please indicate your occupation.
   Fill in the blank

4. Do you use English as part of your job?
   1 = Yes
   2 = No

   a) If yes, how often do you use English as part of your job?
      1 = Never
      2 = Hardly
      3 = Often
      4 = Very Often
      5 = Always

5. How long have you been working in your current job?
   1 = Less than one year
   2 = 1-5 years
   3 = 6-10 years
   4 = 11-20 years
   5 = More than 20 years

6. As part of your current job, do you supervise others?
   1 = Yes
   2 = No

7. As part of your job, do you have to speak with people via the telephone with whom you have not had previous contact?
   1 = Yes
   2 = No

8. What is the highest level of education that you have completed?
   1 = Some high school
   2 = High school
   3 = Some college
   4 = B.A. or B.S. Degree
   5 = M.A. or M.S. degree
   6 = Ph.D or Ed.D

Section C: Biodata

Directions: The following items are related to your experiences using the telephone and computers. Please read each question carefully and choose the appropriate response.

1. Have you ever taken part in a telephonic job interview?
   1 = Yes
   2 = No

2. Have you ever taken a test via the telephone?
   1 = Yes
   2 = No

3. Are you required to use the internet as part of your job?
   1 = Yes
   2 = No

4. Have you ever applied for a job on the internet?
   1 = Yes
   2 = No

5. Have you ever taken an online course?
   1 = Yes
   2 = No

6. Have you ever taken a language course online?
   1 = Yes
   2 = No

7. Have you ever taken a test on the internet?
   1 = Yes
   2 = No

8. Do you use online messaging?
   1 = Yes
   2 = No

9. Previous experience using computers.
    a. How many years have you been using computers?
        1 = Less than one year
        2 = 1-5 years
        3 = 6-10 years
        4 = 11-20 years
        5 = More than 20 years

    b. How often do you use the internet at work in a typical day?
        1 = I never use the internet at work
        2 = Less than 1 hour
        3 = 1-2 hours
        4 = 3-4 hours
        5 = More than 5 hours

    c. How often to you use the internet at home in a typical day?
        1 = I never use the internet at home
        2 = Less than 1 hour
        3 = 1-2 hours
        4 = 3-4 hours
        5 = More than 5 hours

10. Previous English training/education.
    a. When did you first start to study English?
        1 = Primary school
        2 = Middle school
        3 = High school
        4 = College
        5 = Job-related training

    b. How many English courses have you taken both at school and through private institutes?
        1 = I have never taken an English course
        2 = 1-3
        3 = 4-6
        4 = 7-9
        5 = 10 or more

11. Previous experience with English testing.
    a. Have you ever taken an Oral Proficiency Interview (OPI)?
        1 = Yes
        2 = No

    b. Have you ever taken the Test of English as a Foreign Language (TOEFL) or the Test of English for International Communication (TOEIC)?
        1 = Yes
        2 = No

    c. Have you ever taken any of the Cambridge examinations in English for speakers of Other Languages (Cambridge ESOL exams)?
        1 = Yes
        2 = No

d. Have you ever taken any other standardized tests of English proficiency not mentioned above?
1 = Yes
2 = No

Section D: Attitudes toward Computerized/Telephonic Tests

Directions: Please read the following items related to your attitudes toward computerized and telephonic tests and indicate how much you agree or disagree with each item using the 5-point scale provided.

[Scale: 1 = Strongly Disagree, 2 = Disagree, 3 = Neither Disagree nor Agree, 4 = Agree, 5 = Strongly Agree]

1. I feel comfortable taking an exam on the internet.
2. I feel comfortable taking an exam on the telephone.
3. I feel comfortable speaking into a computer.
4. I feel comfortable speaking on the telephone.
5. I feel comfortable speaking when I know I am being recorded.

Section E: Test-taking self-efficacy

Directions: Please read the following items related to your level of confidence in performing the tasks identified. Please use the 5-point scale provided to indicate how much you agree or disagree with each item.

[Scale: 1 = Strongly Disagree, 2 = Disagree, 3 = Neither Disagree nor Agree, 4 = Agree, 5 = Strongly Agree]

1. I am confident in my ability to perform well on written tests.
2. I am confident in my ability to perform well on telephonic tests.
3. I am confident in my ability to perform well on computer-based tests.
4. I am confident in my ability to take a test in English via the telephone.
5. I am confident in my ability to take an internet-based test in English.
6. I am confident in my ability to take a paper and pencil test in English.
7. I am confident in my ability to have a conversation in English with someone I do not know.
8. I am confident in my ability to have a conversation in Korean with someone I do not know.
9. I am confident in my ability to perform well in high-pressure situations.
10. I am confident in my ability to communicate in English when I know I am being recorded.

Section F: Auditory and Visual Learning Preferences

Directions: Please review the statements below related to your learning preferences and indicate how often you engage in the following behaviors using the 5-point scale provided.

[Scale: 1 = Never, 2 = Rarely, 3 = Sometimes, 4 = Often, 5 = Always]

Visual

1. I remember information better if I write it down.
2. Looking at the person helps keep me focused.
3. I need a quiet place to get my work done.

4. When I take a test, I can see the textbook page in my head.
5. I need to write down directions, not just take them verbally.
6. Music or background noise distracts my attention from the task at hand.
7. I don't always get the meaning of a joke.
8. I doodle and draw pictures on the margins of my notebook pages.
9. I have trouble following lectures.
10. I react very strongly to colors.

Auditory

1. My papers and notebooks always seem messy.
2. When I read, I need to use my index finger to track my place on the line.
3. I do not follow written direction well.
4. If I hear something, I will remember it.
5. Writing has always been difficult for me.
6. I often misread words from the text (i.e., "them" for "then").
7. I would rather listen and learn than read and learn.
8. I'm not very good at interpreting an individual's body language.
9. Pages with small print or poor quality copies are difficult for me to read.
10. My eyes tire quickly, even though my vision check-up is always fine.

Section G: Goal orientation

Directions: People have different views about how they approach work. Please read each statement below and select the response that reflects how much you agree or disagree with the statement.

[Scale: 1 = Strongly Disagree, 2 = Disagree, 3 = Sort of Disagree, 4 = Neither Disagree nor Agree, 5 = Sort of Agree, 6 = Agree, 7 = Strongly Agree]

1. I am willing to select a challenging work assignment that I can learn a lot from.
2. I often look for opportunities to develop new skills and knowledge.
3. I enjoy challenging and difficult tasks at work where I'll learn new skills.
4. For me, development of my work ability is important enough to take risks.
5. I prefer to work in situations that require a high level of ability and talent.
6. I like to show that I can perform better than my coworkers.
7. I try to figure out what it takes to prove my ability to others at work.
8. I enjoy it when others at work are aware of how well I am doing.
9. I prefer to work on projects where I can prove my ability to others.
10. I would avoid taking on a new task if there was a chance that I would appear rather incompetent to others.
11. Avoiding a demonstration of low ability is more important to me than learning a new skill.
12. I'm concerned about taking on a task at work if my performance would reveal that I had low ability.
13. I prefer to avoid situations at work where I might perform poorly.

Appendix B

Post-Assessment Survey

Section A: Survey ID

1. Last name/Family name
   Fill in the blank

2. First name
   Fill in the blank

3. Survey ID
   Fill in the blank

Section B: ACTFL OPIc®

Directions: Please read the following items related to taking the ACTFL OPIc® (speaking test taken on the computer) and choose the appropriate response.

Initial Instructions

1. Please indicate if you chose to read the ACTFL OPIc® instructions in English or Korean?
   1 = English
   2 = Korean

[Answer choices for these items: 1 = Strongly Disagree, 2 = Disagree, 3 = Neither Disagree nor Agree, 4 = Agree, 5 = Strongly Agree]

2. I was able to understand the instructions that preceded the ACTFL OPIc® and that explained the procedures for taking the ACTFL OPIc® in the language that I chose.
3. I found it helpful to be able to choose to read the instructions in Korean or English.

Background Survey

[Scale for these items: 1 = Strongly Disagree, 2 = Disagree, 3 = Neither Disagree nor Agree, 4 = Agree, 5 = Strongly Agree]

1. The questions on the background survey were clear and easy to understand.
2. The questions on the background survey allowed me to provide sufficient information about my background and life.
3. I felt comfortable answering the questions on the background survey.

Open-ended question:

1. Please indicate any other topics that should or should not have been included in the background survey.

Self-Assessment

1. What level of proficiency did you choose during the self-assessment?
   Responses: 4 Self Assessment options from the ACTFL OPIc®.

[Answer choices for these items: 1 = Strongly Disagree, 2 = Disagree, 3 = Neither Disagree nor Agree, 4 = Agree, 5 = Strongly Agree]

2. I clearly understood that the intention of the self assessment was for me to estimate my English speaking ability.
3. I found it difficult to select the description (one of the four), that best describes my level of English speaking ability.
4. I found it easy to select the description (one of the four) that best describes my level of English speaking ability.
5. When choosing the description (one of the four) for my level of English speaking ability, I had difficulty choosing between some of the options.

Test description/instructions

[Answer choices for these items: 1 = Strongly Disagree, 2 = Disagree, 3 = Neither Disagree nor Agree, 4 = Agree, 5 = Strongly Agree]

1. After reading the instructions, I knew how to navigate through the ACTFL OPIc®.
2. I found the part of the instructions where the functions of each of the buttons that were used in the ACTFL OPIc® were described to be helpful.
3. After reading the description of the buttons used in the ACTFL OPIc® interface, I had no problems navigating through the ACTFL OPIc® using the buttons.
4. I understood from the directions that I could not listen to or re-record my answers to questions on the ACTFL OPIc®.
5. It was only after I made a mistake and tried to stop and re-record my answer on the ACTFL OPIc® that I realized I was not able to do so.
6. After reading the instructions, I understood that I was only able to press the "Repeat" button once for each question on the ACTFL OPIc®.
7. I did not realize that I could only press the "Repeat" button once after each question on the ACTFL OPIc® after reading the instructions.
8. The test description and instructions adequately provided me with the information necessary to take the ACTFL OPIc®.

Test format tutorial/ sample

[Answer choices for these items: 1 = Strongly Disagree, 2 = Disagree, 3 = Neither Disagree nor Agree, 4 = Agree, 5 = Strongly Agree]

1. The sample test question was helpful.
2. I could hear Ava the Avatar clearly when listening to the sample test question.
3. It was helpful to see Ava the Avatar on the screen when listening to the sample test question.
4. I found it helpful to practice recording answers to the sample test questions.
5. The tutorial and sample test question prepared me to take the ACTFL OPIc®.

ACTFL OPIc®

[Answer choices for these items: 1 = Strongly Disagree, 2 = Disagree, 3 = Neither Disagree nor Agree, 4 = Agree, 5 = Strongly Agree]

1. I had no technical/computer problems while completing the ACTFL OPIc®.

2. The audio quality of the questions on the ACTFL OPIc® was sufficient for me hear the questions clearly.
3. I had difficulty with the audio quality of the test questions.
4. Ava the Avatar was a helpful visual aid.
5. I liked the "look and feel" of Ava the Avatar.
6. Ava the Avatar's voice was clear and understandable.
7. I found Ava the Avatar to be distracting and annoying.
8. Ava the Avatar made the testing experience comfortable and user friendly.
9. Ava the Avatar added to the realism of the interview for me.
10. I had problems recording answers to the test questions.
11. I used the "Repeat" button frequently when taking the ACTFL OPIc®.
12. The questions were clear and I understood what I needed to say.
13. The topics asked were related to my interests and experiences.
14. The topics asked allowed me to demonstrate my speaking proficiency.
15. The instructions for the role-play were clear.
16. I understood what was expected of me in the role-play.
17. I found the role-play format to be easier to understand and to respond to than the other questions.
18. I found the role-play format to be difficult.
19. I found the topic of the role-play to be related to my interests and my experiences.
20. I believe my performance on the ACTFL OPIc® accurately reflects my current speaking proficiency level.
21. I believe the ACTFL OPIc® is an effective way to measure English speaking proficiency.
22. I would recommend taking an ACTFL OPIc® to a friend who needs their speaking proficiency assessed.
23. My interview lasted approximately _____ minutes.
    1 = Less than 10
    2 = 10
    3 = 15
    4 = 20
    5 = 25
    6 = 30
    7 = More than 30

Open-ended questions:

1. Please describe any technical problems you encountered when taking the ACTFL OPIc®.
2. Please provide any suggestions to improve the ACTFL OPIc® experience.
3. Do you have any comments about the ACTFL OPIc® process and interface that you would like to share?

Section C: ACTFL OPI®

Directions: Please read the following items related to taking the ACTFL OPI® (speaking test taken via telephone) and choose the appropriate response.

[Answer choices for these items: 1 = Strongly Disagree, 2 = Disagree, 3 = Neither Disagree nor Agree, 4 = Agree, 5 = Strongly Agree]

1. The interviewer provided an introduction/overview to the ACTFL OPI®.
2. The interviewer was friendly and polite.
3. The interviewer's voice was clear and understandable.

4. The interviewer asked questions at the beginning of the interview about my interests and experiences.
5. I felt comfortable speaking with the interviewer.
6. The interviewer encouraged me to answer the questions.
7. I had no technical problems with the telephonic ACTFL OPI®.
8. I could hear the interviewer clearly when listening to his/her questions.
9. I had difficulty understanding the interviewer over the telephone.
10. I asked the interviewer to repeat questions frequently during the interview.
11. The questions asked provided me with an adequate opportunity to demonstrate my speaking proficiency.
12. I believe my performance on the ACTFL OPI® accurately reflects my current speaking proficiency level.
13. I believe the ACTFL OPI® is an effective way to measure English speaking proficiency.
14. I would recommend taking an ACTFL OPI® to a friend who needs their speaking proficiency assessed.
15. My interview lasted approximately _____ minutes.
    1 = Less than 10
    2 = 10
    3 = 15
    4 = 20
    5 = 25
    6 = 30
    7 = More than 30

Open-ended questions:

1. Please describe any technical problems you encountered when taking the ACTFL OPI®.
2. Please provide any suggestions to improve the ACTFL OPI® experience.
3. Do you have any comments about the ACTFL OPI® process and interface that you would like to share?

Section D: Comparison of ACTFL OPI® and ACTFL OPIc®

Directions: Please read the following items related to taking the ACTFL OPI® (speaking test taken via telephone) as compared tot taking the ACTFL OPIc® (speaking test taken on the computer) and choose the appropriate response.

[Answer choices for these items: 1 = Strongly Disagree, 2 = Disagree, 3 = Neither Disagree nor Agree, 4 = Agree, 5 = Strongly Agree]

1. I thought it was more difficult to demonstrate my speaking proficiency via the computer (ACTFL OPIc®) than with a live interviewer over the telephone (ACTFL OPI®).
2. Both the computer and telephonic interviews provided an adequate opportunity for me to demonstrate my speaking proficiency.
3. The ACTFL OPIc® was more user friendly than the ACTFL OPI®.
4. The ACTFL OPIc® provided a better opportunity for me to demonstrate my speaking proficiency.
5. I preferred the testing format with a live interviewer than with the Avatar.
6. It was easier to understand questions from a live interviewer than from the Avatar.
7. I felt more comfortable recording my answers on the computer than providing answers to a live interviewer.

8. In which format (ACTFL OPI®/ ACTFL OPIc®) did you feel you were able to demonstrate your best speaking proficiency?
   1 = ACTFL OPI®
   2 = ACTFL OPIc®
   3 = Both equally

Open-ended question:

1. If you were to take the test again, would you rather take the ACTFL OPI® or the ACTFL OPIc®? Please explain your reasons for choosing the test taking format that you chose.

Appendix C

Responses to Post-Assessment Survey

Table 1. Language of ACTFL OPIc® Instructions

Please indicate if you chose to read the ACTFL OPIc® instructions in English or Korean?		N	Percentage
	English	4	5
	Korean	76	95

Table 2. ACTFL OPIc® – Initial Instructions

				Percentage (%) of Responses				
	N	Mean	Standard deviation	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
I was able to understand the instructions that preceded the ACTFL OPIc® and that explained the procedures for taking the ACTFL OPIc® in the language that I chose.	80	4.08	.81	--	5.0	13.8	50.0	31.3
I found it helpful to be able to choose to read the instructions in Korean or English.	80	4.02	.86	2.5	2.5	12.5	55.0	27.5

Table 3. ACTFL OPIc® – Background Survey

				Percentage (%) of Responses				
	N	**Mean**	**Standard deviation**	**Strongly Disagree**	**Disagree**	**Neutral**	**Agree**	**Strongly Agree**
The questions on the background survey were clear and easy to understand.	80	3.79	.74	--	6.3	21.3	60.0	12.5
The questions on the background survey allowed me to provide sufficient information about my background and life.	80	3.37	.93	2.5	13.8	37.5	36.3	10.0
I felt comfortable answering the questions on the background survey.	80	3.39	.93	1.3	17.5	32.5	38.8	10.0

Table 4. Reported Self-assessment

What level of proficiency did you choose during the self-assessment?		N	Percentage
	Level 1	52	67.5
	Level 2	13	16.9
	Level 3	7	9.1
	Level 4	5	6.5

Table 5. ACTFL OPIc® – Self-assessment

				Percentage (%) of Responses				
	N	Mean	Standard deviation	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
I clearly understood that the intention of the self assessment was for me to estimate my English speaking ability.	77	3.62	.81	2.6	6.5	23.4	61.0	6.5
I found it difficult to select the description (one of the four), that best describes my level of English speaking ability.	77	2.92	.97	6.5	28.6	33.8	28.6	2.6
I found it easy to select the description (one of the four) that best describes my level of English speaking ability.	77	3.19	.92	2.6	20.8	36.4	35.1	5.2
When choosing the description (one of the four) for my level of English speaking ability, I had difficulty choosing between some of the options.	77	2.99	.95	6.5	22.1	41.6	26.0	3.9

Table 6. ACTFL OPIc® – Test Description/instructions

	N	Mean	Standard deviation	Percentage (%) of Responses				
				Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
After reading the instructions, I knew how to navigate through the ACTFL OPIc®.	77	3.75	.80	1.3	5.2	23.4	57.1	13.0
I found the part of the instructions where the functions of each of the buttons that were used in the ACTFL OPIc® were described to be helpful.	77	3.74	.89	2.6	3.9	28.6	46.8	18.2
After reading the description of the buttons used in the ACTFL OPIc® interface, I had no problems navigating through the ACTFL OPIc® using the buttons.	77	3.70	.90	1.3	10.4	20.8	51.9	15.6
I understood from the directions that I could not listen to or re-record my answers to questions on the ACTFL OPIc®.	77	3.58	.92	2.6	10.4	24.7	50.6	11.7
It was only after I made a mistake and tried to stop and re-record my answer on the ACTFL OPIc® that I realized I was not able to do so.	77	2.82	1.19	13.0	35.1	15.6	29.9	6.5
After reading the instructions, I understood that I was only able to press the "Repeat" button once for each question on the ACTFL OPIc®.	77	3.47	1.07	2.6	23.4	11.7	49.4	13.0
I did not realize that I could only press the "Repeat" button once after each question on the ACTFL OPIc® after reading the instructions.	77	2.81	1.19	11.7	39.0	14.3	27.3	7.8
The test description and instructions adequately provided me with the information necessary to take the ACTFL OPIc® .	77	3.64	.76	1.3	3.9	33.8	51.9	9.1

Table 7. ACTFL OPIc® – Test Format/tutorial Sample

				Percentage (%) of Responses				
	N	**Mean**	**Standard deviation**	**Strongly Disagree**	**Disagree**	**Neutral**	**Agree**	**Strongly Agree**
The sample test question was helpful.	77	3.66	.75	1.3	2.6	35.1	50.6	10.4
I could hear Ava the Avatar clearly when listening to the sample test question.	77	3.77	.76	1.3	3.9	23.4	59.7	11.7
It was helpful to see Ava the Avatar on the screen when listening to the sample test question.	77	3.35	.87	1.3	14.3	40.3	36.4	7.8
I found it helpful to practice recording answers to the sample test questions.	77	3.78	.84	1.3	5.2	24.7	51.9	16.9
The tutorial and sample test question prepared me to take the ACTFL OPIc®.	77	3.62	.81	1.3	5.2	35.1	46.8	11.7

Table 8. ACTFL OPIc® Evaluation

	N	Mean	Standard deviation	Percentage (%) of Responses				
				Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
I had no technical/computer problems while completing the ACTFL OPIc®.	75	3.91	1.02	2.7	10.7	8.0	50.7	28.0
The audio quality of the questions on the ACTFL OPIc® was sufficient for me hear the questions clearly.	75	4.04	.80	1.3	2.7	13.3	56.0	26.7
I had difficulty with the audio quality of the test questions.	75	2.19	1.04	21.3	57.3	8.0	8.0	5.3
Ava the Avatar was a helpful visual aid.	75	3.19	.78	2.7	13.3	48.0	34.7	1.3
I liked the "look and feel" of Ava the Avatar.	75	2.60	.92	14.7	24.0	49.3	10.7	1.3
Ava the Avatar's voice was clear and understandable.	75	3.75	.74	1.3	2.7	26.7	58.7	10.7
I found Ava the Avatar to be distracting and annoying.	75	2.44	.81	4.0	61.3	24.0	8.0	2.7
Ava the Avatar made the testing experience comfortable and user friendly.	75	3.04	.69	2.7	12.0	65.3	18.7	1.3
Ava the Avatar added to the realism of the interview for me.	75	2.95	.84	4.0	24.0	46.7	24.0	1.3
I had problems recording answers to the test questions.	75	2.80	1.01	4.0	46.7	18.7	26.7	4.0
I used the "Repeat" button frequently when taking the ACTFL OPIc®.	75	3.05	1.20	12.0	22.7	22.7	33.3	9.3

Table 8. ACTFL OPIc® Evaluation (continued)

	N	Mean	Standard deviation	Percentage (%) of Responses				
				Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
The questions were clear and I understood what I needed to say.	75	3.40	.85	2.7	9.3	40.0	41.3	6.7
The topics asked were related to my interests and experiences.	75	3.25	.86	4.0	12.0	41.3	40.0	2.7
The topics asked allowed me to demonstrate my speaking proficiency.	75	3.08	.85	4.0	17.3	48.0	28.0	2.7
The instructions for the role-play were clear.	75	3.29	.82	2.7	9.3	49.3	33.3	5.3
I understood what was expected of me in the role-play.	75	3.21	.93	5.3	14.7	37.3	38.7	4.0
I found the role-play format to be easier to understand and to respond to than the other questions.	75	2.89	.89	4.0	29.3	44.0	18.7	4.0
I found the role-play format to be difficult.	75	3.15	.88	1.3	22.7	41.3	29.3	5.3
I found the topic of the role-play to be related to my interests and my experiences.	75	3.05	.84	2.7	21.3	46.7	26.7	2.7
I believe my performance on the ACTFL OPIc® accurately reflects my current speaking proficiency level.	75	3.20	.87	4.0	12.0	49.3	29.3	5.3
I believe the ACTFL OPIc® is an effective way to measure English speaking proficiency.	75	3.35	.89	2.7	12.0	40.0	38.7	6.7
I would recommend taking an ACTFL OPIc® to a friend who needs their speaking proficiency assessed.	75	3.31	.87	2.7	13.3	40.0	38.7	5.3

Table 9. Length of ACTFL OPIc® Interview

My interview lasted approximately _____ minutes.	N	Percentage
Less than 10	10	13.3
10	7	9.3
15	16	21.3
20	24	32.0
25	9	12.0
30	4	5.3
More than 30	5	6.7

Table 10. ACTFL OPI® Evaluation

				Percentage (%) of Responses				
	N	**Mean**	**Standard deviation**	**Strongly Disagree**	**Disagree**	**Neutral**	**Agree**	**Strongly Agree**
The interviewer provided an introduction/overview to the ACTFL OPI®.	70	3.76	.91	4.3	5.7	12.9	64.3	12.9
The interviewer was friendly and polite.	70	4.00	.85	4.3	--	10.0	62.9	22.9
The interviewer's voice was clear and understandable.	70	3.94	.92	4.3	1.4	14.3	55.7	24.3
The interviewer asked questions at the beginning of the interview about my interests and experiences.	70	3.83	.82	4.3	--	17.1	65.7	12.9
I felt comfortable speaking with the interviewer.	70	3.54	.91	4.3	4.3	35.7	44.3	11.4
The interviewer encouraged me to answer the questions.	70	3.70	.92	4.3	5.7	18.6	58.6	12.9
I had no technical problems with the telephonic ACTFL OPI®.	70	3.80	.91	4.3	2.9	18.6	57.1	17.1
I could hear the interviewer clearly when listening to his/her questions.	70	3.63	1.00	5.7	4.3	27.1	47.1	15.7
I had difficulty understanding the interviewer over the telephone.	70	2.63	1.01	8.6	45.7	24.3	17.1	4.3
I asked the interviewer to repeat questions frequently during the interview.	70	2.89	1.21	15.7	24.3	22.9	30.0	7.1

Table 10. ACTFL OPI® Evaluation (continued)

				Percentage (%) of Responses				
	N	**Mean**	**Standard deviation**	**Strongly Disagree**	**Disagree**	**Neutral**	**Agree**	**Strongly Agree**
The questions asked provided me with an adequate opportunity to demonstrate my speaking proficiency.	70	3.37	.89	4.3	5.7	47.1	34.3	8.6
I believe my performance on the ACTFL OPI® accurately reflects my current speaking proficiency level.	70	3.46	.94	4.3	8.6	34.3	42.9	10.0
I believe the ACTFL OPI® is an effective way to measure English speaking proficiency.	70	3.66	1.00	5.7	5.7	21.4	51.4	15.7
I would recommend taking an ACTFL OPI® to a friend who needs their speaking proficiency assessed.	70	3.64	1.00	5.7	5.7	22.9	50.0	15.7

Table 11. Length of ACTFL OPI® Interview

My interview lasted approximately _____ minutes.	**N**	**Percentage**
Less than 10	8	11.4
10	8	11.4
15	12	17.1
20	23	32.9
25	9	12.9
30	5	7.1
More than 30	5	7.1

Table 11. Comparison of ACTFL OPI® and OPIc®

				Percentage (%) of Responses				
	N	**Mean**	**Standard deviation**	**Strongly Disagree**	**Disagree**	**Neutral**	**Agree**	**Strongly Agree**
I thought it was more difficult to demonstrate my speaking proficiency via the computer (ACTFL OPIc®) than with a live interviewer over the telephone (ACTFL OPI®).	69	3.41	1.05	5.8	14.5	23.2	46.4	10.1
Both the computer and telephonic interviews provided an adequate opportunity for me to demonstrate my speaking proficiency.	69	3.25	.76	1.4	13.0	46.4	37.7	1.4
The ACTFL OPIc® was more user friendly than the ACTFL OPI®.	69	2.70	.96	7.2	39.1	34.8	14.5	4.3
The ACTFL OPIc® provided a better opportunity for me to demonstrate my speaking proficiency.	69	2.90	.91	7.2	21.7	47.8	20.3	2.9
I preferred the testing format with a live interviewer than with the Avatar.	69	3.70	.91	2.9	7.2	21.7	53.6	14.5
It was easier to understand questions from a live interviewer than from the Avatar.	69	3.30	.91	2.9	14.5	39.1	36.2	7.2
I felt more comfortable recording my answers on the computer than providing answers to a live interviewer.	69	2.71	.94	8.7	33.3	39.1	15.9	2.9

Table 12. Demonstrating Speaking Proficiency

In which format (ACTFL OPI®/ACTFL OPIc®) did you feel you were able to demonstrate your best speaking proficiency?	N	Percentage
ACTFL OPI®	44	63.8
ACTFL OPIc®	10	14.5
Both equally	15	21.7

Table 13. Responses to Open-ended Question 1

Question: Please indicate any other topics that should or should not have been included in the background survey.
I think it would be good to include past memories, such as of college or memorable things that happened throughout life. For example if you have memories of having been involved in a particular activity in college, or if you had felt the joy of achievement at your previous employer after working on something.
Certain specific issues, such as those pertaining to political, economical or social issues should not be included
Selecting items of interest was difficult.
A self-assessment of one's English proficiency should most certainly be included, and perhaps by asking the participants to disclose their TOEIC or TOEFL scores, which could be objective measurements of their English skills, a level-appropriate test could be provided?
How much time do you invest into your interests?
Necessary - Family members, hobby, talents
I do not remember what kind of items there were.
Necessary items : Sex distinction, Age, level of English proficiency Unnecessary items : Such personal and private items such as where you live and what company you work for
Include : Experience in taking other English tests and average score
The specific nature of your work and interests/hobbies should be included
There are too many items
Include : English speaking proficiency Do not include : Education
It's been such a long time I don't remember
It is not necessary to ask whether or not you are employed/working or not, since most test-takers are obviously company employees. Also, as most work environments are within the company offices, asking whether or not you work at home is also unnecessary. It seems that there should be an extra row (column) indicating 'other' following leisurely activities, interests and sports.
Name, age, education, major, hobby, talent, married/unmarried, interests, etc. should be included

Note. Not all participants provided comments. Some participants provided open-ended comments that were not transmitted properly and therefore are not included in this table.

Table 13. Responses to Open-ended Question 1 (continued)

Question: Please indicate any other topics that should or should not have been included in the background survey.
All of the necessary items seem to have been included
good as is
I don't know (do not remember contents of the survey)
Not sure, haven't thought about it.
None (Note: Comment made by 12 participants).
Nothing in particular. (Note: Comment made by 2 participants).
I don't know (Note: Comment made by 2 participants).
Do not remember. (Note: Comment made by 2 participants).

Note. Not all participants provided comments. Some participants provided open-ended comments that were not transmitted properly and therefore are not included in this table.

Table 14. Responses to Open-ended Question 2

Question: Please describe any technical problems you encountered when taking the ACTFL OPIc®.
Sometimes the screen would freeze and so it was uncertain whether or not my reply was being recorded or not, and there were times I had to close the current window and log-in again.
No particular technical problems
Uncomfortable because there is no reply or re-record function
A virus check would appear every three minutes, making it uncomfortable because I had to continuously close pop-up windows.
Wrong ID
It was difficult to understand the accent and I was not able to ask if there was vocabulary that I was not aware of
1. I thought I was able to listen to the question again but only one extra chance was given. 2. It was uncomfortable because when I wanted to immediately reply after the question was over the recording function would not activate immediately.
The avatar is unsatisfactory
The system froze during the test
Only one chance was given to listen again
No extra chance to listen again
There was no way to verify whether or not the volume of my voice was appropriate. Felt uncomfortable speaking loudly by myself.
There was no interaction; the question that should ensure the previous question was not presented, therefore taking up more time. It is important to emphasize prior to beginning the test that you are being tested on your proficiency, and not on the actual contents of the questions.
I could not find out if my voice volume was appropriate or not in recording.
It was the first test, a level was not selected and the test started from number 2
Operating the headset
The lip movement and actual pronunciation did not seem to match

Note. Not all participants provided comments. Some participants provided open-ended comments that were not transmitted properly and therefore are not included in this table.

Table 14. Responses to Open-ended Question 2 (continued)

Question: Please describe any technical problems you encountered when taking the ACTFL OPIc®.
Automatic log-out during the test.
The buttons, including the play button, are too small.
None (Note: Comment made by 16 participants).
No particular technical problems, (Note: Comment made by 3 participants).
Nothing in particular (Note: Comment made by 2 participants).

Note. Not all participants provided comments. Some participants provided open-ended comments that were not transmitted properly and therefore are not included in this table.

Table 15. Responses to Open-ended Question 3

Question: Please provide any suggestions to improve the ACTFL OPIc® experience.
When I started to take the test, I couldn't find the stop button. When I realized I had to press the next button instead of the stop button, I was surprised.
The avatar, the re-listening function and the sound quality were all pretty good. When the stop button was pressed after recording, the screen would freeze and the test would not proceed. That is the only error that needs improvement.
In the OPI, you can ask to slow down a bit when you are not feeling sure about the questions at hand but with OPIc, since you are talking with a computer you cannot make such a request. If there were a 'slower' or 'speed' function then participants could adjust the speed on their own.
It would be nice to have a re-recording function
It was difficult because I didn't feel like I was talking with an actual person.
It would be nice to get more than two chances to listen. There should be a limit to the overall timing, but it would be better if there were multiple chances for listening.
A problem was that some questions required too much talking at once. If possible, even if you have to increase the number of questions perhaps you could change the questions so that answers could be less than 1 minute long. Talking alone into a computer is a little uncomfortable in the first place and therefore it is hard to open up and start talking. In the OPI, you are talking with an actual tester and such characteristic should be applied to the OPIc.
Avoid repetitive questions.
The waiting time after listening to the questions was too long. It would have been nice to be able to proceed with recording by clicking on a button.
Increase the number of times for repetition
Include more elements that can emphasize the distinctive character of the test (like describing images, etc)
Inevitably speaking to a computer causes less tension than talking with someone on the phone.
Compose the screen to be more active and less static, right now it is too static and boring.
Diversify the question pool, and make it into a system that can ask questions pertinent to the opinions presented by test-takers
Improvements could be made to the character (avatar)

Note. Not all participants provided comments. Some participants provided open-ended comments that were not transmitted properly and therefore are not included in this table.

Table 15. Responses to Open-ended Question 3 (continued)

Question: Please provide any suggestions to improve the ACTFL OPIc® experience.

Make multiple times of re-listening possible

How about accompanying questions with supplmentary questions that could help test-takers in answering?

Compared with the OPI which is carried out in a conversation format, the evaluator's reaction is weak. For example, when the testee is not able to answer appropriately to a question, in the OPI, the tester repeats the questions in a slower speed or in an easier format. The opic is more difficult because you can only listen to the same question twice. Also, it feels like you are recording rather than having a conversation and therefore it feels like an entirely different test from the OPI.

Change the Avatar

Improve the screen composition.

It would look more realistic if the images were actual pictures rather than illustrations.

The OPI test seems to be better. It should be improved to give the feeling of talking with an real person.

Modify the screen composition, improve how pilot tests are conducted.

Improve the quality of design

An improvement could be to change it into an interactive conversational format..

Preset the volume of what is already recorded.

The alignment of the questions were weak because of the low level of interactivity with the participant.

Soundproofing between participants

Stabilize the system

The avatar made me feel uncomfortable

I wish there was a time-keeping device while recording

None (Note: Comment made by 10 participants).

Note. Not all participants provided comments. Some participants provided open-ended comments that were not transmitted properly and therefore are not included in this table.

Table 16. Responses to Open-ended Question 4

Question: Do you have any comments about the ACTFL OPIc® process and interface that you would like to share?
It would be better to have a better design. I feel the test is well organized but the user Interface looks a little cheap, tacky and old.
It could be confusing at first if you don't know your way around but the testing process, the practice and explanations were helpful and satisfactory. Another satisfactory point was that the test was composed so that the test-taker can skip the explanation part and go straight into the test. The overall process looks pretty good.
Looks good at present.
It is possible to grasp the interface after one testing experience; questions could be predicted for future tests and therefore it is possible to memorize predicted answers beforehand and answer with them during the test.
The waiting time after listening to the questions was too long. It would have been nice to be able to proceed with recording by clicking on a button.
A blond character appears in the test screen, which looks a little cheap for a test that requires to appear credible.
In order to aid the test-takers to provide more various and colorful answers the interface should be strengthened by inserting illustrations or pictures.
Although there were no particular problems in the procedures of the tests, the UI requires some improvement. Evaluation and AI application seem to be crucial.
The screen composition reflects no sense of reality.
How about using a video-conference style format? That would make it feel more like a test.
The contents of the questions and testing time should be adjusted to accommodate different levels of speaking proficiency.
There was some misunderstanding in understanding the functions of the different types of buttons. An interface design that makes possible a more general understanding is needed.
It should have a more sophisticated design and the screen should be composed to look easier.. In terms of screen composition, you could collect opinions from the company's contents development team.
faster instructions
I wish there was a remaining time display or a recording time display function
I don't know.
None (Note: Comment made by 23 participants).

Note. Not all participants provided comments. Some participants provided open-ended comments that were not transmitted properly and therefore are not included in this table.

Table 17. Responses to Open-ended Question 5

Question: Please describe any technical problems you encountered when taking the ACTFL OPI®.
I am not sure because I have never taken the OPI.
It is uncomfortable because you can not listen again nor re-record
A virus-checking pop-up window kept appearing every three minutes which made it uncomfortable.
Now..
Because of the cultural difference it is difficult to explain things that I don't know.
The avatar is not satisfactory.
It could be tiring to have to listen on a phone for a long time.
Sometimes it was hard to get connected via phone.
Automatic log-out during the test
Operating the headset.
No specific errors
Already answered. (Note: Comment made by 2 participants).
There were no technical problems. (Note: Comment made by 5 participants).
No particular technical problems. (Note: Comment made by 3 participants).
None (Note: Comment made by 17 participants).

Note. Not all participants provided comments. Some participants provided open-ended comments that were not transmitted properly and therefore are not included in this table.

Table 18. Responses to Open-ended Question 6

Question: Please provide any suggestions to improve the ACTFL OPI® experience.
I am not sure because I have never taken the OPI.
I got the impression that the conversation was like a Q&A session. It would have been better to have the test to be carried out in a comfortable conversational format but the questions were awkward and I felt pressured to answer when asked a question.
Re-recording should be added.
There should be encouragements and hints to help test-takers answer.
There is nothing to improve in OPI.
The conversation should be more realistic
Questions should be more diversified.
It is more difficult compared with OPIc because you have to repeatedly listen to the questions when you have not understood them.
It should look more active. It looks boring because it's so static.
Make re-listening possible.
Test-takers should be given chances to have more active questions and chances to select the theme.
The screen composition should be changed to look more sophisticated.
Connection is crucial.
I wish I could get access to information on who is making the phone call.
A diverse pool of questions should be gathered, and personalized questions should be given.
Soundproofing between participants.
The volume of the recorded parts should be preset.
Interviewers might have different speed, tone, intonation of spoken English, which may have influence on the result of the test. Male interviewers have such a deep voice that may be difficult to novice speakers.

Note. Not all participants provided comments. Some participants provided open-ended comments that were not transmitted properly and therefore are not included in this table.

Table 18. Responses to Open-ended Question 6 (continued)

Question: Please provide any suggestions to improve the ACTFL OPI® experience.
Stabilize the system
The avatar makes me feel uncomfortable
Already answered (Note: Comment made by 2 participants).
None (Note: Comment made by 15 participants).

Note. Not all participants provided comments. Some participants provided open-ended comments that were not transmitted properly and therefore are not included in this table.

Table 19. Responses to Open-ended Question 7

Question: Do you have any comments about the ACTFL OPI® process and interface that you would like to share?
I am not sure as I have never taken the OPI test.
The new kind of over-the-phone talking method was good..
If the OPI tester resides in the country then the scheduling the test might be easier
It is good to be able to receive immediate feedback after talking directly with the foreign instructor.
How about introducing a video-conferencing format to OPI.
It would be nice to have a remaining time display, or recording time display function
Although it is a test since it is carried out in the format of a conversation with another person it feels much more comfortable.
Nothing in particular
Already answered. (Note: Comment made by 2 participants).
None (Note: Comment made by 28 participants).

Note. Not all participants provided comments. Some participants provided open-ended comments that were not transmitted properly and therefore are not included in this table.

Table 20. Responses to Open-ended Question 8

Question: If you were to take the test again, would you rather take the ACTFL OPI® or the ACTFL OPIc®? Please explain your reasons for choosing the test taking format that you chose.
I would select the OPIC. I have never taken the OPI, but I would be more nervous if I had to talk on the phone and I don't think I would be able to hear well. It would feel strange to ask to repeat questions and so I don't think it would be more efficient. I think it would be more efficient to have some time to think and then record when you are ready.
OPI. Because if you talk with a real person then you can talk naturally and connect with different topics.
OPI. Talking with a real person lessens the emotional pressure coming from being tested; being able to have a conversation seems to be good.
OPI. It was more comfortable, and I would be guided when I was lost or confused.
OPIc. Less pressure.
OPI, A sense of trust is built when someone is listening to me and we are having a real-time, interactive conversation (although the tester does evaluate me while listening) and so it feels much more comfortable.
OPI felt more comfortable. I didn't like talking with the Avatar.
Opi. Talking with a real person made me feel more comfortable and when I didn't know something the tester would provide explanations so that I could answer.
I would select the OPI. It is good to receive direct feedback in the conversation. But I would recommend the OPIc for people who don't like talking directly with the tester or who find it difficult to take a test within a limited amount of time.
OPI. Although it is a bit more stressful it was still a little more fun. Whether or not it can objectively evaluate is the most important thing.
OPI (The test isn't boring since you are having a conversation and you can have a real conversation.)
OPI: In the OPIc, since there is no interaction between the tester and the test-taker, it is less interesting and hard to take seriously
OPI. Because it is real communication with a person it is more natural and when you don't understand easier language is provided and therefore less stressful in answering questions.
I would select the OPI. Because I felt that the conversation unfolded with more ease through the interactions with the tester.
OPI-Because it would be a more accurate assessment
opic Because it's less stressful

Note. Not all participants provided comments. Some participants provided open-ended comments that were not transmitted properly and therefore are not included in this table.

*Table 2*0. Responses to Open-ended Question 8 (continued)

Question: If you were to take the test again, would you rather take the ACTFL OPI® or the OPIc®? Please explain your reasons for choosing the test taking format that you chose.
OPI. The conversation is more natural because you're talking with a real person.
OPI, You can attend to the test in a more active way.
OPI. Since you can interact with the tester you can provide answers more naturally.
Having a conversation with a person is more comfortable and there is more understanding of mistakes; questions are sometimes rephrased to sound easier.
OPI. Exchanging conversations is more comfortable.
OPIc - When talking directly with a foreigner on the phone one could get excessively frustrated when an answer is delayed; such problems are deterrents to accurately measuring one's real ability.
OPI . It has the advantage that even if you only say one word the person can understand you.
opic it's more realistic
OPI is a little better because the testers continue to encourage test-takers, gives related questions, and also provides explanations to parts that are not understood. The OPIC should also have such a function.
Both should be used.
opi, more interactive interview responding to different situations
OPI. For beginners like me; because the questions are asked by a real person and not a machine, I can ask to repeat if I didn't understand the intention of the questions and can also ask the question to be rephrased in an easier way.
OPI. The questions are more realistic.
→ opi! I don't believe one's English proficiency can be assessed with a limited amount of questions. It is better to assess a person's English proficiency through conversation.
Opi. Because it's not merely an unconditional test
OPI (Note: Comment made by 26 participants).
OPIc (Note: Comment made by 10 participants).

Note. Not all participants provided comments. Some participants provided open-ended comments that were not transmitted properly and therefore are not included in this table.

