Brief Reliability Report 5:

Test-Retest Reliability and Absolute Agreement Rates of English ACTFL OPIc® Proficiency Ratings for Double and Single Rated Tests within a Sample of Korean Test Takers

Prepared by

SWA Consulting Inc.

301 Glenwood Avenue
Suite 220
Raleigh, NC 27603
www.swa-consulting.com

Prepared for

Language Testing International

3 Barker Avenue
Suite 300
White Plains, NY 10601

Executive Summary

This report discusses the stability of final English ACTFL OPIc® ratings provided to a sample of Korean test takers. To complete this report, Language Testing International (LTI) provided data that contained 1 to 13 final ACTFL OPIc® ratings for a sample of individuals (N=2934). This dataset allowed SWA Consulting to assess the stability and agreement of the final ratings obtained by individual test takers over the course of consecutive ACTFL OPIc® administrations. To accomplish this goal, the test-retest reliability and rates of absolute agreement of all available ACTFL OPIc® final ratings were calculated. The results from these calculations revealed that the final ratings of the first two ACTFL OPIc® administrations were highly stable even when taking into account the time elapsed between administrations (Pearson's *r* values ranging from .90 to .93, Spearman's *R* values ranging from .90 to .94, and rates of absolute agreement ranging from 85% to 92%). Furthermore, a series of similar analyses focusing on final ratings provided by single raters indicated that these ratings were also highly stable (Pearson's *r* values ranging from .96 to .99, and Spearman's *R* values ranging from .97 to .99) even when taking into account time elapsed between administrations. To provide additional context to these results, conceptual discussions are offered of reliability in general, as well as the use of test-re-test reliability and rates of absolute agreement as indicators of the reliability of the ACTFL OPIc®. These findings provide evidence supporting the stability of final ratings attained on the English ACTFL OPIc® within a 30-day period. Test-retest reliability and absolute agreement were high and exceeded traditionally accepted minimum levels for all ACTFL OPIc®s, including single rated tests.

Table of Contents

The ACTFL OPIc® 4

Goals of this Report 4

Reliability in General Terms 5

Test-Retest Reliability of Final ACTFL OPIc® Ratings 6

Absolute Agreement between ACTFL OPIc® Final Ratings 6

Empirical Evidence for Test-Retest Reliability and Absolute Agreement 7

- Sample 7
- Procedures 7
- Results 8

Conclusion 11

References 12

The ACTFL OPIc®

Language Testing International (LTI) uses the American Council on the Teaching of Foreign Languages (ACTFL) Oral Proficiency Interview (OPI®) as a standardized procedure to assess the functional speaking ability of individuals around the world. The OPI® is most accurately characterized as an assessment that measures how well individuals speak a particular language. All individuals that complete an OPI are assessed in terms of ten proficiency criteria specified by ACTFL in the ACTFL Revised Proficiency Guidelines—Speaking Revised 1999 (Breiner-Sanders, Lowe, Miles, & Swender, 2000)

Obtaining a rating on the OPI® involves a person engaging in a structured interview with a single certified LTI interviewer. During this interview a ratable speech sample is elicited from an interviewee by an interviewer who follows a series of structured questions and comments as specified by the ACTFL protocols for determining levels of language proficiency (LTI, 2008).

Inherently, the OPI® does not involve a comparison between different interviewees. All ratings are highly individualized and done on a person-by-person basis, with at least one interviewer and one interviewee participating in the rating process (LTI, 2004).

The OPI® is also available in a computerized version, known as the OPIc®, with the "c" representing the computerized nature of the assessment. 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.

Goals of this Report

The current report was compiled with four main goals in mind. The first goal is to provide an overview of the test-retest reliability of consecutive final ACTFL OPIc® ratings, while the second goal is to provide information regarding the rate of absolute agreement (i.e. concordance) between final ACTFL OPIc® ratings across consecutive administrations. The third and fourth goals are to examine the test-retest reliability and the rate of absolute agreement of final ACTFL OPIc® ratings in instances where only one rater's rating determined test taker's final rating.

To accomplish these goals, a conceptual overview of reliability is offered below. This overview is followed by detailed discussions of the use and implications of test-retest reliability as well as rates of absolute agreement. These discussions serve as the theoretical basis for the empirical inquiry that was performed to establish evidence for the stability of ACTFL OPIc® final ratings.

Reliability in General Terms

The term reliability can be used to describe the consistency and stability of the measurement of characteristics of people and things (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). This general definition also applies to the testing of human attributes such as language proficiency. Therefore, in terms of psychometric measurement, reliability is synonymous with the consistency, stability, replicability, and repeatability of a measurement across locations, times, or populations (Anastasi, 1988; Cattell, 1988; Feldt & Brennan, 1989; Flanagan, 1951; Stanley, 1971; Thorndike, 1951; Traub, 1994). In other words, the reliability of a measurement indicates the degree to which it measures an attribute of a person in a systematic and repeatable way (Walsh & Betz, 2000). A common way to conceptualize reliability is to refer to the use of a ruler and a tape measure, both instruments which will yield highly similar results consistently if they are used accurately. Thus, both instruments can be described as being highly reliable (Walsh & Betz, 2000).

This conceptualization of reliability applies equally well to psychometric measurement. In classic psychometric testing theory, it is assumed that individuals have a specific or a "true" amount of an attribute, which is referred to as the person's *true score* (reflected in part in the individual's score on a psychometric measurement). However, this true score is only a component of the *observed score* (the score received on the psychometric measurement). This is due to the notion that every psychometric measure has an inherent amount of *error* that takes place with every measurement. In other words the observed score (score on the psychometric measure) is equal to the true score (how much of an attribute the person actually has) plus error (Traub, 1994). This relationship is summarized in the following equation:

True Score = Observed Score + Error

To place the concepts of *observed score*, *true score*, and *error* into more concrete terms, it is useful to reference the imagery of using a ruler or tape measure to assess the dimensions of a physical object. It can be said that the *true* dimensions of an object are its actual dimensions, whereas its *observed* dimensions are those determined through the use of a ruler or a tape measure. These measurements contain the true score as well as a certain amount of *error inherent to observation*. The error contained within the measurement, however, is negligible if the ruler or the tape measure was used correctly and accurately, but is nonetheless present. Consequently, if only a minimal amount of error was involved in using a ruler and the tape measure, both instruments can be deemed to be reliable methods of determining the dimensions of a physical object.

Similarly, a psychometric measure can only be deemed reliable if it has a small amount of error contained within the measurements that it makes. This relationship between true score and error obtained by a psychometric measure is known as a reliability coefficient (the higher the reliability coefficient, the smaller the error in the observed score). Consequently, the reliability coefficient of a measure indicates, at least, partially how useful that measure is (Walsh & Betz, 2000).

Test-Retest Reliability of Final ACTFL OPIc® Ratings

Test-retest reliability is calculated when the same test is administered to the same person on two or more different occasions. Test-retest reliability therefore is considered to be an indication of the stability with which a particular test measures the same phenomenon over time (Trochim, 2006; Walsh & Betz, 2000), and can be assessed through the calculation of what is known as Pearson's correlation (*r).*

Pearson's correlation (*r*), sometimes called a *product–moment* correlation, is one of the most widely used methods for assessing test-retest reliability. This correlation assesses the degree to which ratings covary or display rank-order stability across multiple administrations. Spearman's rho correlation (*R*) is computationally identical to Pearson's *r*; however, Spearman's *R* rank-orders all data for each variable prior to calculating the correlation. Spearman's *R* will be computed in addition to Pearson's *r* because *R* is less affected by unusual (or outlying) cases and nonlinearity in the relationship of interest compared to *r*. In this sense, reliability can be depicted in the classical test theory 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* (reflected by the observed score and error of ratings; Surface, et al., 2008). Interpreting these correlations is intuitive, with higher correlations (i.e., those closer to 1.00) suggesting more stability between ratings than lower correlations.

Generally, indices of reliability at or above .70 are traditionally considered to be adequate (LeBreton, Burgess, Kaiser, Atchley, & James 2003), with higher levels being recommended for high-stakes testing purposes.

Absolute Agreement between ACTFL OPIc® Final Ratings

Another way to assess the stability of tests is to determine the rate of agreement between scores attained by test takers across multiple administrations. Traditionally, rates of absolute agreement are used to estimate the concordance of ratings provided by multiple raters regarding a single event (i.e. interrater agreement). However, it also has great utility in the current context in the sense that it can be used to provide an overview of the stability of final ratings provided of individual test takers that have completed multiple administrations of the ACTFL OPIc®.

Higher rates of absolute agreement (i.e. the percentage of ratings that are exactly the same for each test taker) indicate greater degrees of stability between ratings across multiple administrations. Therefore, in the current context, higher rates of agreement indicate the extent to which test takers receive the same scores across different administrations of the ACTFL OPIc®.

Generally, when two ratings are utilized, absolute agreement rates at or in excess of 70% are seen to be acceptable (Surface et al., 2008). Absolute agreement requires perfect stability; this however, is a very stringent requirement to meet. Adjacent agreement refers to the rate of agreement between ratings provided to a single test taker on consecutive tests that are only 1 level removed from one another.

This is a less stringent measure of concordance than absolute agreement, but it does provide a good indication of test stability across multiple administrations.

Empirical Evidence for Test-Retest Reliability and Absolute Agreement

To achieve the goals of the current inquiry, test-retest reliability and rates of absolute agreement were calculated for the final ratings attained on the ACTFL OPIc® for a sample of Korean test takers who completed this assessment at least twice in a 30-day period. The following sections describe the sample, the procedures involved in the administration of the ACTFL OPIc®, the data utilized, and the results that were obtained for both sets of calculations.

Sample

The sample (N =2934) included Korean test takers who have completed multiple administrations (ranging from 1 to 13 administrations) of the ACTFL OPIc® to assess their English language proficiency. All of these test takers completed two administrations of the ACTFL OPIc® within at least one 30-day period prior to 20 January, 2009.

A small portion of the initial data points were removed for one of the following two reasons. First, if a test taker only completed one ACTFL OPIc®, his/her data were removed from the final dataset. Second, only final ratings from ACTFL OPIc®s that were administered within 30 days of another ACTFL OPIc® were included in the final dataset. Since true change in actual speaking ability would be treated as error in the calculation of test-retest reliability and agreement, ACTFL OPIc®s taken more than 30 days from the previous administration were not considered.

Employing this approach avoids bias in the reliability analyses caused by inconsistency in ratings that may reflect real changes in test takers' language proficiency rather than instability in the ACTFL OPIc® rating process. Additional analytical steps, outlined below, were also employed to further account for changes in individuals' language proficiency.

The final dataset contained ratings from 2852 individual test takers, which constituted 97% of those from the initial sample. The number of ACTFL OPIc® administrations (and final ratings) for any individual in the final dataset ranged from two to eight ACTFL OPIc®s.

Procedures

Test takers that were dissatisfied with their initial rating, completed additional administrations of the ACTFL OPIc® within certain parameters. Generally, test takers are allowed to complete the ACTFL OPIc® with 30 days between administrations. This rule, however, has two notable exceptions: All test takers are entitled to a one-time only retest within the customary retest timeframe and waivers can be granted to test takers who experienced technical difficulties during their last assessment or for other reasons, thereby allowing them to complete another ACTFL OPIc® within the customary retest timeframe. All final ratings were assigned using standard ACTFL rating protocols.

Results

Analytical Process

To accurately assess the stability of the ACTFL OPIc®, ratings were grouped according to the length of time that had elapsed between the first and second administration of the test. All cases in which the first two consecutive ACTFL OPIc® administrations were completed within a week (up to 7 days) were grouped together. The same procedure was used to group cases together for which the first two consecutive administrations were completed between one and three weeks (8-14 days), two and four weeks (15-21 days), and more than 3 weeks apart (22-30 days) apart. This strategy was designed to account for the potential impact of actual change in test takers' language proficiency over time on the estimates of test-retest reliability and absolute agreement of final ratings.

Following this grouping procedure, two separate sets of analyses were performed. The first set of these analyses considered ratings irrespective of the number of raters that were involved in providing final ratings to test takers. The second set of analyses considered only ratings that were provided by single raters. This second set of analyses was deemed necessary due to suspicions of instability regarding ratings provided by single raters.

In both sets of analyses, test-retest reliability (Pearson's *r* and Spearman's R_s) and rates of absolute agreement were calculated for each of the groups based on the first two ACTFL OPIc® administrations per test taker as well as all consecutive final ratings available per test taker.

Stability of all final ratings

The results of these computations revealed that the final ratings of all ACTFL OPIc® tests included in the final dataset are highly stable across the first two consecutive administrations (*r* values from .90 to .93, *R* values from .90 to .94, for all of which $p \leq .01$). Similarly, rates of absolute agreement were also computed for each of the groups based on the first two ACTFL OPIc® administrations per test taker included in the final dataset. The results of the absolute agreement computations mimicked the pattern of the test-retest results, indicating high rates of agreement (85% to 92%) between the first and second final ratings attained on the ACTFL OPIc®.

Both sets of statistics, provided in *Table 1*, trended downward slightly as elapsed time between the initial and the subsequent administration of the ACTFL OPIc® increased. A statistical comparison (using Fisher's *r-to-z* transformation) indicated the decrease in Pearson *r* value from .93 (for a retest within 7 days) to .90 (for a retest during the 8-14, 15-21, and 22-30 day time intervals) was statistically significant ($p < .05$). That is, test-retest reliability for retests that occurred more than one week after the first test was significantly lower than that of retests that occurred within seven days. However, the absolute levels of these indices at each of the four time intervals remained well above minimally acceptable levels, and .90 is still a very high reliability coefficient. Thus, these results indicate a high degree of stability in ratings across the first and second administrations conducted within 30 days. The slight downward trend may be a result of additional student learning and preparation which would be considered error in this situation. This may also be an artifact of high sample size (i.e. a statistically significant but practically meaningless difference).

Number of Days Since 1st OPIc When Retest Occurred		Relationship with 1st OPIc Final Rating
1 to 7 days	*r*	.93*
	R	.94*
	Agreement	92%
	N	776
8-14 days	*r*	.90*
	R	.90*
	Agreement	89%
	N	900
15-21 days	*r*	.90*
	R	.91*
	Agreement	*85%*
	N	517
22-30 days	*r*	.90*
	R	.91*
	Agreement	*85%*
	N	659

*Note: * p*≤.01. *r* = Pearson correlation. *R* = Spearman correlation.

Table 1. Test-retest and absolute agreement of the ACTFL OPIc® final rating

To place these results in a larger context, *Table 2* illustrates the direction and percentage of change for all consecutive ACTFL OPIc® final ratings that occurred within a 30 day period. This analysis, therefore, presents the observed changes in all available consecutive ratings (e.g. first and second ratings, second and third ratings, third and fourth ratings, and so on) for each test taker within the final pared down sample used for the previous analysis. For example, a change in final rating from Intermediate Low to Intermediate Mid is considered an "Increase by 1 level" in *Table* 2. Results presented in *Table 2* indicate the vast majority of final ratings (88%) remain the same from one administration to the next. In addition, more than 99% of all final ratings remained unchanged or changed by one level from one administration to the next. Therefore, the adjacent agreement would be 99%.

Amount of Change in Final Rating	Frequency of Rating Change	Percentage of All Ratings
Decrease by 2 levels	1	< 1%
Decrease by 1 levels	55	2%
No change	2715	88%
Increase by 1 levels	292	9%
Increase by 2 levels	16	< 1%

Table 2. Change in final rating of any two ACTFL OPIc® administrations taken within 30 days

Stability of ratings provided by single raters

For 1652 test takers, the first and second (within 30 days) test was rated by a single rater, which enabled the calculation of test-retest reliability and rates of absolute agreement for single-rated ACTFL OPIc®s.

For all of these single-rated ACTFL OPIc®s, test-retest reliability for the first and second administrations was high (r = .98, p < .01; R = .98, p < .01) as well as absolute agreement (97%).

Number of Days Since 1st OPIc When Retest Occurred		Relationship with 1st OPIc Rating
1 to 7 days	*r*	.99*[a]
	R	.99*
	Agreement	99%
	N	458
8-14 days	*r*	.97*[b]
	R	.97*
	Agreement	97%
	N	535
15-21 days	*r*	.97*[b,c]
	R	.99*
	Agreement	96%
	N	294
22-30 days	*r*	.96*[c]
	R	.97*
	Agreement	96%
	N	365

Note: * p < .01. r = Pearson correlation. R = Spearman correlation. Pearson correlations that do not share the same letter are significantly different (p < .05)

Table 3. Test-retest and absolute agreement of single-rated ACTFL OPIc®s

Both sets of statistics, provided in *Table 3*, trended slightly downward as elapsed time between the first and second single-rated administrations of the ACTFL OPIc® increased. A statistical comparison (using Fisher's *r-to-z* transformation) indicated the decrease in Pearson r value from .99 (for a retest within 7 days) to .97 (for a retest within 8-14 days) was statistically significant (p < .05). The decrease in Pearson r value from .97 (for a retest within 8-14 days) to .96 (for a retest within 22-30 days) was also statistically significant (p < .05). However, the absolute levels of these indices at each of the four time intervals remained well above minimally acceptable levels. As explained before, this is likely an artifact of large sample size and is statistically significant, but practically meaningless. Reliability coefficients above .90 are extremely high. Thus, these results indicate a high degree of stability in ratings across the first and second single-rated administrations.

Conclusion

Collectively, these findings provide evidence for the stability of final ratings attained on the English ACTFL OPIc®.

Both the test-retest reliability coefficients and absolute agreement indices were high and above traditionally accepted minimum levels irrespective of whether final ratings were provided by single ratings or derived from the ratings of multiple raters. These results indicate that test takers received very consistent results across consecutive administrations of the ACTFL OPIc®. Furthermore, it was found that the vast majority of test takers' final ratings (more than 99%) remained unchanged or only changed by one level from any given administration of the ACTFL OPIc® to the next, providing additional evidence for the stability of final ratings. To summarize, the stability of ACTFL OPIc® final ratings was high and exceeded minimum professional standards.

Based on these results, ACTFL OPIc® final ratings demonstrated adequate levels of test-retest reliability and agreement to justify their use as a standard assessment of speaking proficiency. Users of the ACTFL OPIc® can be confident in the stability of final ratings, regardless of the use of double or single-rated protocols.

References

LTI (2004). ACTFL certified proficiency testing program: Oral and writing proficiency testing for the state of Florida prospective teachers. White Plains, NY. Available at www.languagetesting.com/download/send_file.cfm?filename=FLTeacherProfAssessmentbrochureFinal1.doc, accessed 20 January, 2009.

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.

Breiner-Sanders, K.E., Lowe, P., Miles, J., & Swender, E. (2000). ACTFL Proficiency Guidelines—Speaking Revised 1999. *Foreign Language Annals, 33(1)*, 13-17.

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.

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.

LeBreton, J. M., Burgess, J. R. D., Kaiser, R. B., Atchley, E. K., & James, L. R. (2003). The restriction of variance hypothesis and interrater reliability and agreement: Are ratings from multiple sources really dissimilar? *Organizational Research Methods, 6*, 80-128.

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

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

Traub, R. (1994). *MMSS Reliability for the social sciences: Theory and Applications*. Sage Publications. Newbury Park, CT.

Trochim, W.M.K. (2006). Research Methods Knowledge Base. Available at http://www.socialresearchmethods.net/kb/index.php, accessed 20 January, 2009.

Surface, E.A., Poncheri, R.M., & Bhavsar, K.S. (2008, March). *Two studies investigating the reliability and validity of the English ACTFL OPIc® with Korean test takers*. Raleigh, NC: SWA Consulting Inc.

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

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

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

Walsh, W.B., Betz, N.E. (2000). *Tests and assessment*. Prentice Hall, Upper Saddle River, NJ.

