american council on the teaching of foreign languages
ACTFL
6 Executive Plaza, Yonkers, NY 10701-6801
914.963.8830 • FAX 914.963.1275 • headquarters@actfl.org • http://www.actfl.org

Preliminary Reliability and Validity Findings for the ACTFL Writing Proficiency Test

SWA Technical Report 2004-C04-R01

Prepared for

American Council on the Teaching of Foreign Language
Washington, D.C.

Language Testing International
White Plains, NY

Prepared by:
Surface, Ward & Associates

Dr. Eric A. Surface
Principal, SWA

Dr. Erich C. Dierdorff
Visiting Professor, DePaul University
Senior Associate, SWA

Executive Summary

This technical report provides preliminary reliability and validity information about the Writing Proficiency Test (WPT) developed by the American Council on the Teaching of Foreign Languages (ACTFL). The ACTFL WPT is a standardized global assessment of functional writing ability in a language and measures the ability of the test taker to write spontaneously in a language without the opportunity to revise his or her responses and/or use reference or editing tools. The performance of the test taker is compared with the criteria stated in the *ACTFL Proficiency Guidelines—Writing (revised 2001)* to assign a rating. A total of 509 writing proficiency tests, conducted and rated by experienced ACTFL-certified testers using the ACTFL WPT assessment procedure, were included in this study. Measures of interrater consistency (interrater reliability and agreement) were calculated in order to assess the quality of judgments made by those who score the WPT, for both a full sample and a Spanish-only sample. Interrater consistency was found to be well above acceptable levels for applied settings (e.g., r = .94 and .92 for full and Spanish-only samples, respectively). Measures of interrater agreement indicated that for the full sample the majority of judges provided identical scores to (80% perfect agreement). Similar results were found for the Spanish-only sample as well (78% perfect agreement). Interrater agreement is also provided within each proficiency category (e.g., Advanced) and levels (e.g., Advanced-Mid) by this report. Finally, longitudinal interrater reliability was assessed for the more than three years that the revised WPT guidelines have been in use. The longitudinal reliability trends indicate that the interrater reliability has generally increased during the time the revised procedures have been in place. Limited evidence of validity can be provided by relating the assessment to other variables that measure the same, similar or different constructs. For a subgroup of the cases (n = 460), ACTFL Oral Proficiency Interview

(OPI) scores were also available. The relationship between the OPI and WPT scores was found to be robust ($r = .81$; $p < .001$), suggesting that both the OPI and WPT are assessing related, overlapping constructs. This finding provides limited evidence of validity because one would expect measures of language skill in the same language using the same assessment method to be at least moderately related, especially since writing and speaking are both productive skills. In other words, we found what was expected given our data. If a non-significant relationship had been found, then we would have had a potential validity problem. More data and a well-designed experiment are needed to assess the validity of the WPT thoroughly.

Overall, the results of this preliminary study of reliability and validity are positive for the WPT. As more data becomes available, future research should examine the reliability and validity of the WPT in more depth.

Navigating this Report

This report is organized similar to an APA-styled article in order to help the reader understand, find, and evaluate our research and its findings. This will allow the reader to make an informed decision about the WPT. This document is organized into five sections: (a) introduction and purpose; (b) research background and rationale; (c) methods; (d) results; and, (e) discussion. The table of contents below should help the reader to navigate the document more effectively.

Table of Contents

Preliminary Reliability and Validity Findings for the ACTFL Writing Proficiency Test

Introduction & Purpose

In 2002, the American Council on the Teaching of Foreign Languages (ACTFL) published their revised guidelines for writing proficiency (Breiner-Sanders, Swender & Terry, 2002). These guidelines were used to create the ACTFL writing proficiency test (WPT). This technical report provides preliminary reliability and validity information about the WPT. Our justifications and guide for presenting this reliability and validity information can be found in the *Standards for Educational and Psychological Testing*, published by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) (1999).

The *Standards* provide evaluative guidelines for the users, developers, and publishers of tests, which 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" (p. 3, AERA et al., 1999), not simply restricted to paper-and-pencil assessments. Test publishers and developers have a responsibility to provide validity and reliability data about their assessments. Validity refers "to the degree to which evidence and theory support the interpretations of the test scores entailed by proposed uses of tests" (p. 9), whereas reliability indicates the ability of the testing procedure to provide a consistent measure of the specified domain when repeated. As a new assessment, ACTFL and Language Testing International (LTI) could not provide preliminary psychometric information about the WPT until a minimal number of tests had been conducted.

To assess reliability, this report provides indices of consistency and agreement for the WPT as applied to five languages (French, German, Italian, Russian, and Spanish). The findings are presented for the overall sample and for Spanish-only because Spanish is the only language with sufficient test cases to justify a separate analysis. Some limited validity evidence is provided in the form of the relationship between the WPT and the Oral Proficiency Interview (OPI). Because these assessments measure two different skill modalities of language proficiency (i.e., writing and speaking) and use the same method (i.e., trained raters using the ACTFL testing protocol), a strong relationship was expected. The remainder of this document presents the rationale, methodology, findings, and interpretation of our analysis of the reliability and validity of the WPT.

Research Background and Rationale

This section reviews the ACTFL writing proficiency guidelines (Breiner-Sanders et al., 2002) and provides an overview of reliability and validity as it applies to this context before discussing previous research and presenting our research objectives.

Writing Proficiency Guidelines

The ACTFL proficiency guidelines were first published in 1986 and are global characterizations of integrated performance in each of the four language skill modalities, including writing (ACTFL, 1986; Breiner-Sanders et al., 2002). These guidelines, which were based on the language skill level descriptions used by the Interagency Language Roundtable and adapted for the academic context, have been revised since 1986. The writing proficiency guidelines were revised in 2001 following the precedent set by revising the speaking guidelines in 1999 (see Breiner-Sanders et al., 1999).

As with the speaking guidelines, the writing proficiency guidelines specify four major levels of writing proficiency (i.e., *Superior*, *Advanced*, *Intermediate*, and *Novice*) that are divided into 10 sublevels. According to the *ACTFL Writing Proficiency Test Familiarization Guide* (ACTFL, 2002a), the four major levels differentiate proficiency according to “a hierarchy of global tasks” that span the full range of writing proficiency. The levels are hierarchical and are presented in descending order from *Superior* to *Novice* with each level subsuming the levels below it. The 10 sublevels in descending order are: *Superior, Advanced High, Advanced Mid, Advanced Low, Intermediate High, Intermediate Mid, Intermediate Low, Novice High, Novice Mid, and Novice Low*.

According to *ACTFL Writing Proficiency Test Familiarization Guide* (ACTFL, 2002a), *Superior* writers “can produce informal and formal writing on practical, social and professional topics treated both abstractly and concretely,” “can present well-developed ideas, opinions, arguments, and hypotheses through extended discourse,” and “can control structures, both general and specialized/professional vocabulary, spelling, punctuation, cohesive devices and all other aspects of written form and organization with no pattern of error to distract the reader” (p. 5). *Advanced* writers are less proficient. They “can write routine, informal and some formal correspondence, narratives, descriptions, and summaries of a factual nature in all major time frames in connected discourse of a paragraph length,” and their writing “is comprehensible to all native speakers due to breadth of generic vocabulary and good control of the most frequently used structures” (p. 5). The proficiency of *Intermediate* writers decreases further. *Intermediate* writers “can meet a range of simple and practical writing needs” as well as “communicate simple facts and ideas” (p.5). However, their writing is only “comprehensible to those accustomed to the writing of non-natives” (p.5). *Novice* writers “can produce lists and notes and limited formulaic information on simple forms and documents," and their writing is "typically limited to words, phrases and memorized material" (p.5).

The *ACTFL Proficiency Guidelines—Writing (Revised 2001)* (Breiner-Sanders et al., 2002) are the basis for the ACTFL WPT. The WPT is a standardized global assessment of functional writing ability in a language and measures the ability of the test taker to write spontaneously in a language without the opportunity to revise his or her responses and/or use reference or editing tools. The performance of the test taker is compared with the criteria stated in the guidelines to assign a rating. As a new assessment, preliminary reliability and validity information need to be presented for the WPT in accordance with the *Standards* (AERA et al., 1999). Before presenting the relevant data from our research, some basic information about reliability and validity in this context is provided.

Reliability and Interrater Consistency

Consistency defined by 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 raters. One of the most commonly used forms of rater reliability estimation is interrater reliability, which portrays the overall level of consistency among the sample of raters involved in a particular judgment process. When interrater reliability estimates are high, the interpretation has a large degree of consistency across sample raters.

Another common approach to examining interrater consistency is to use measures of 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, test takers, etc.). 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.

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 et al., 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* (AERA et al., 1999) 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).

Given the nature of the ACTFL WPT and our study, the following *Standards* (AERA et al., 1999) are particularly noteworthy: (1) reliability estimates should be reported for each test score, subscore, or combination of scores (Standard 2.1); (2) reliability coefficients from similar assessments are not interchangeable unless their implicit definitions of measurement error are equivalent (Standard 2.5); (3) evidence of both interrater consistency and within examinee consistency over repeated measurements should be provided for assessments when subjective judgment enters into the scoring process (Standard 2.10); (4) test developers should document the process for the selection and training of raters as well as scorer reliability and drift over time (Standard 3.23); and, (5) test developers and publishers are responsible for amending, revising, or withdrawing a test as new research data becomes available (Standard 3.25). Taken together, providers of test/assessment procedures have the responsibility to report and periodically update the reliability data for their procedures. Thus, the *Standards* provide a strong justification for the research in this study.

Validity

Validity refers “to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests” (p. 9; AERA et al., 1999). In other words, a test or assessment must be valid for its intended use. If it is not, then the test should not be used for that purpose. Validity is the most important psychometric characteristic of a test and must be demonstrated through the accumulation of empirical, scientific evidence that the scores can be appropriately interpreted and used for a specified purpose. Evidence supporting the test for one purpose does not automatically make it valid for another purpose. The *Standards* provide guidelines for assessing and reporting evidence of validity. The *Principles for the Validation and Use of Personnel Selection Procedures* (Society for Industrial and Organizational Psychology [SIOP], 2003) provide additional guidance in situations were a test is used for the purpose of personnel selection.

Validity refers to a unitary concept and is "the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed purpose" (p. 11; AERA et al., 1999). In the past, types of validity evidence were given specific labels (e.g., construct validity or criterion validity). This practice was confusing; therefore, the current version of the *Standards* drops the traditional nomenclature in favor of descriptions of the evidence types. There are five categories of evidence: (a) evidence based on test content; (b) evidence based on response processes; (c) evidence based on internal structure; (d) evidence based on relations to other variables; and (e) evidence based on the consequences of testing. Although a complete review of all categories of evidence is beyond the scope of this document, the category, evidence based on relations to other variables, is reviewed because it provides the basis for the limited validity data presented in this report.

The statistical relationship (e.g., correlation) of a test to established measures of the same construct, related constructs, or different constructs can provide evidence of validity. According to the *Standards*, "evidence based on relationships with other variables addresses questions about the degree to which these relationships are consistent with the construct underlying the proposed test interpretations" (p. 13; AERA et al., 1999). The relationship between scores on a test and scores on measures that assess the same or similar constructs provides convergent evidence of validity. The relationship between scores on the target assessment and scores on tests that measure different constructs provide discriminant evidence. In general, a test should be correlated more highly with measures assessing the same or similar constructs, than with tests assessing different or dissimilar constructs. In the case of this study, a strong correlation between the ACTFL WPT and OPI would provide some limited validity evidence because one would expect measures of language skill in the same language using the same assessment method to be at least moderately related, regardless of the skill modality difference (i.e., writing versus speaking). Additionally, since writing and speaking are both productive skills, one might expect a significant relationship. However, the evidence would be more powerful if the other skill modalities (i.e., listening and reading), multiple languages, and/or multiple test administrations were included in the study. Therefore, a strong relationship between the WPT and OPI will only provide limited validity evidence, whereas, the lack of a statistically significant relationship will call the tests into question.

Previous Research

Although the WPT is a new assessment and no previous research related directly to its reliability and validity exists, evidence from two previous studies using writing proficiency measures developed according the ACTFL proficiency guidelines (ACTFL, 1986) does exist. However, before presenting the results from the two studies, the reader should know that very limited information was given as to the development and nature of the writing proficiency assessments used in these studies. Therefore, these findings should be viewed with great caution when making comparison with the current study.

In 1990, Dandonoli and Henning presented the results of a multitrait-multimethod validation study of the OPI, which included tests of speaking, writing, listening and reading in French and English as a Second Language (ESL). The study sample included French students at Northwestern University and ESL students primarily from Brandeis University. Completely new tests were created for this study, but little information was presented about the writing test. The interrater reliability (Pearson *r*) for the writing test for the English sample (n = 59) was reported as .87. The interrater reliability (Pearson *r*) for the writing test for the French sample (n = 60) was reported as .89. Interestingly, four correlations between writing and speaking proficiencies were reported for each sample—correlations for speaking Rater A and Rater B with writing Rater A and Rater B (i.e., a 2x2 matrix). The writing and speaking tests were not rated by the same people. For the English sample, the four correlations between the speaking raters and writing raters were .85, .86, .92, and .88. For the French sample, the four correlations between the speaking raters and writing raters were .85, .80, .84, and .80.

Thompson (1996) presented results from an assessment of speaking, reading, listening and writing proficiency of students of Russian who had varying years of study using tests based on the ACTFL proficiency guidelines (ACTFL, 1986) and scale. The writing, listening, and reading tests were developed for the study. Speaking was measured using ACTFL OPI testers. The writing test consisted of five prompts at varying difficulty levels on the ACTFL scale. Each participant was given 45 minutes to complete the writing test, and ACTFL-certified OPI testers familiar with the writing proficiency guidelines scored each test. The interrater reliability (Pearson *r*) for the pilot was reported as .88 (df = 18; p < .05). The actual study used a sample of students from the University of Iowa and one from the Middlebury Russian Summer program. The interrater reliability (Pearson *r*) for the Iowa sample was reported as .91 (df = 25; p < .001). The interrater reliability (Pearson *r*) for the Middlebury sample was reported as .72 (df = 28; p < .001). For the combined samples, the raters had absolute agreement for 27% of the cases. The relationship between writing and speaking was found to be .64 (p < .001).

Research Objectives

In order to provide preliminary psychometric data on the ACTFL WPT, this technical report addresses the following general research objectives or questions:

- What is the reliability of the WPT across the five languages in the sample?
- What is the reliability of the WPT for Spanish?
- Has the reliability of the WPT changed since its inception in 2002?
- What is the relationship between WPT and OPI scores for all languages in the study and for Spanish only?

Methods

Participants

A total of 509 writing proficiency tests, conducted and rated by experienced ACTFL-certified testers using the ACTFL WPT assessment procedure, were included in this study. The majority cases were completed between January 2002 and April 2004 and were for the purpose of teacher certification in two states. This study used data from tests in five different languages: French (*n* = 81), German (*n* = 8), Italian (*n* = 22), Russian (*n* = 3), and Spanish (*n* = 395). Spanish was the only language separately analyzed because of its substantial number of cases. 460 cases also had OPI scores as well as WPT scores, allowing us to assess the correlation between the two skill modalities. All data were made available by Language Testing International (LTI), the ACTFL testing affiliate. No demographic data related to the cases were available.

WPT Rating Procedure

The ACTFL WPT assessment procedure, as described in the *ACTFL Writing Proficiency Test Familiarization Guide (ACTFL, 2002a),* was used to assess writing proficiency. The

ACTFL WPT is a standardized global assessment of functional writing ability in a language and measures the ability of the test taker to write spontaneously in a language without the opportunity to revise his or her responses and/or use reference or editing tools. The performance of the test taker is compared with the criteria stated in the guidelines to assign a rating. The WPT consists of four prompts for written responses dealing with practical, social, and professional topics that are encountered in informal and formal contexts. The test taker is presented with writing tasks and contexts that represent the full range of proficiency levels from *Novice* to *Superior*. The WPT is "not an achievement test assessing a writer's acquisition of specific aspects of course and curriculum content, nor is it tied to any specific method of instruction" (p.3). The WPT assesses writing proficiency in terms of real-life writing tasks.

The WPT is a proctored 90-minute test that consists of an introduction and warm-up, followed by four requests for a variety of writing tasks. All directions and prompts are in English, and all responses are open-ended and written in the target language. The test can be administered via paper and pencil or computer. Each task covers multiple tasks (e.g., descriptive, narrative, etc.) and specifies the audience, context, purpose of the prompt, the suggested length of the response, and the suggested time allotment. In evaluating the test taker's writing, judges consider the following criteria from the wider perspective of how they contribute to the overall writing sample: (a) the functions or global tasks the writer performs, (b) the social contexts and specific content areas within which the writer performs the tasks, (c) the accuracy of the writing, and, (d) the length and organization of the written discourse the writer produces. All judges must go through a rigorous training program to become certified.

OPI Rating Procedure

The ACTFL OPI assessment procedure, as described in the *ACTFL Oral Proficiency Interview Tester Training Manual* (Swender, 1999), consists of four phases (Warm Up, Level Checks, Probes, and Wind Down) that are designed to efficiently elicit a ratable sample. As stipulated by the procedure, a pair of judges rated each case. Some cases required a third tester to serve as a "tie-breaker" in situations of discrepancy between the pair's proficiency ratings. In all cases, the first rater conducted and audiotaped the interviews. Subsequently, this rater judged the interviewee's speaking proficiency from the tape at some later time. Next, the taped interviews were independently rated by the second rater. All raters used the ACTFL rating scale described in the *ACTFL Proficiency Guidelines—Speaking Revised* (Breiner-Sanders et al., 1999) to describe the proficiency levels of the interviewees. If the independent ratings provided by the rating pair disagreed, a third rater was assigned as an arbitrator to rate the interview tape and provide a rating. This rater did not know the previously assigned scores, nor that he or she was the third rater. All raters were ACTFL-certified, meaning that they had completed the ACTFL OPI tester certification process as described in the *ACTFL OPI Tester Certification Information Application Packet* (ACTFL, 2002b). The ACTFL OPI has been found to be highly reliable across 19 languages (Surface & Dierdorff, 2003). For example, the Spanish OPI was reported to have an interrater reliability coefficient of .978 (Pearson *r*).

Analytic Procedure

In order to more accurately assess the extent of interrater consistency, we used a multimethod approach. Interrater consistency can be conceptualized from several perspectives (e.g., interrater reliability, interrater agreement, and so forth) and, thus, a multimethod approach allows for a more complete picture of the level of rating consistency. The overall rationale was to expand the breadth of rater consistency assessment. Interrater consistency measures were calculated for both the full sample and the Spanish version of the WPT in order to facilitate relative comparisons of rater consistency.

Pearson correlation. Sometimes called a *product–moment* correlation, Pearson 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 rank–order correlation (R). This is another commonly used correlation for assessing interrater reliability, particularly in situations involving ordinal variables. Spearman rank–order correlation *(R)* has a 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 the WPT data in that the proficiency levels are ordinal in nature.

Kendall's tau. Tau is equivalent to Spearman's *R* with regard to the underlying assumptions. However, tau and *R* carry different interpretations. *R* is a correlation and thus represents a proportion of variability accounted for, whereas tau is a measure of agreement and represents the difference between two probabilities. Tau is the difference between the probability that the cases are rated in the same order by the two raters and the probability that the cases are rated in different orders by the two raters.

Goodman and Kruskal's gamma. Similar to tau, gamma is a probability-based measure of agreement. However, unlike tau, gamma does not penalize for ties in that they are computationally ignored. As it is desirable to have high interrater consistency (i.e., a large number of tied ratings), gamma can provide useful information beyond that given by tau in terms of interrater consistency. As tied ratings are computationally ignored, the result is that gamma is typically higher in magnitude than tau.

Cohen's kappa. Cohen's kappa is another commonly used measure of agreement, which compares the observed agreement to the agreement expected by chance. Kappa values range from 1.00, when agreement is perfect, to 0.00, when agreement is at the chance level. Kappa does not take into account the degree of disagreement between raters as all disagreements are considered to contribute equally to the total level of disagreement. Therefore, if rating categories are ordered, it is preferable to use a weighted version of kappa, which assigns different weights to ratees for whom the raters differ by *i* categories. Thus, different levels of disagreement can contribute proportionally to the overall value of kappa. Weighted kappa was used in this study.

Raw percentages of agreement. This agreement method assesses the extent to which raters display perfect agreement. It serves as an absolute agreement estimate of interrater consistency and is calculated as the number of identical ratings divided by the number of total rating opportunities. As some disagreements can be expected, it is important to assess percentages of partial agreement as well. Thus, we estimated three separate partial agreement percentages: (1) interrater agreement within plus or minus one proficiency level (e.g., Novice-Low versus Novice-Mid); (2) interrater agreement within plus or minus two proficiency levels (e.g., Intermediate-Low versus Intermediate-High); and, (3) interrater agreement within plus or minus three proficiency levels (e.g., Advanced-Low versus Superior).

Relationship between writing and speaking. In order assess the relationship between writing and speaking proficiency and provide some limited evidence of validity, a Pearson correlation (*r*) and a Spearman rank–order correlation *(R)* were computed.

Results

Table 1 presents the results of the interrater consistency analyses for both the full sample and the Spanish-only sample. All consistency estimates were statistically significant ($p < .05$) and well within desirable levels. Consistency estimates were slightly higher for each test statistic when comparing the full sample and the Spanish-only sample, albeit the differences were minimal in magnitude.

Table 2 displays the results from the interrater agreement analyses. Again, estimates for the full and Spanish-only samples were very similar. For the full sample, a large majority of rater pairs provided identical proficiency judgments (80.1%). Raters judging the Spanish version of the WPT were similarly in absolute agreement in the majority of instances (77.7%). For both samples, when raters were in disagreement most of the discrepancies fell only within a single proficiency level (e.g., Novice -Low to Novice-Mid).

Tables 3 and 4 show the agreement percentages by each proficiency category (Novice, Intermediate, Advanced, and Superior) for the full sample and for the Spanish-only sample, respectively. In the full sample, the proficiency category that displayed the highest level of absolute agreement was the Advanced category (50.1%). The Advanced category also included the largest number of test takers. This pattern held for the Spanish WPT as well (49.6%). No test takers were judged to be in the Novice proficiency category.

Tables 5 and 6 provide a more detailed breakdown of the agreement levels across written language proficiency. These tables display agreement percentages by each proficiency level (e.g. Intermediate-Low, Intermediate-Mid, Intermediate-High, etc.). Of the 408 raters displaying absolute agreement in the full sample, most of these identical judgments fell in the Advanced-

Low and Advanced-Mid proficiency levels (49%). This pattern was similar within the Spanish-only sample as well (50%).

In order to more fully assess the interrater reliability of the Spanish-version WPT, the final set of analyses focused on the longitudinal pattern of reliability. Using Pearson correlations, the first of these analyses examined interrater reliability across yearly categories (2002, 2003, and 2004). The interrater reliabilities for the annual categories are graphically displayed in Figures 1 and 2. To allow comparisons, the reliability estimates were corrected to an equal number of rater pairs using the Spearman-Brown formula. The estimates were adjusted to levels of reliability for 180 rater pairs. The corrected interrater reliabilities showed an upwardly trending pattern across the three years of WPT application. Figures 3 and 4 display the results of similar longitudinal reliability analyses using the bi-annual categories. In these analyses, the general upward trend was again found, although a slight downward shock can be seen in the second half of 2003.

Finally, to provide some limited validity evidence (i.e., the relationship to an established measure of a similar and/or related construct), Pearson correlations were computed between WPT and OPI for all languages ($n = 460$, $r = .81$, $p < .001$) and for Spanish-only ($n = 358$, $r = .81$, $p < .001$) and Spearman rank–order correlations *(R)* were computed between the WPT and OPI for all languages ($n = 460$, $R = .81$, $p < .001$) and for Spanish only ($n = 358$, $R = .81$, $p < .001$).

Discussion

Taken collectively, the results of this study offer strong evidence of favorable interrater reliability for judges scoring the ACTFL WPT, especially for the Spanish WPT. The consistency estimates shown in Table 1 all fall within "acceptable" ranges as described by the relevant literature. For example, the interrater Pearson and Spearman reliability estimates were above the .90 levels, which have been recommended for applied research by Kaplan and Saccuzzo (1982) and Nunnally and Bernstein (1994). Moreover, the weighted Kappa coefficients were in the mid-.80s, which are levels generally accepted to be very high (Landis & Koch, 1977; Gardner, 1995). These favorable reliability results were likewise mirrored in the percentages of agreement analyses, in which the majority of agreements were absolute. In other words, the majority of rater pairs were making identical proficiency level judgments when scoring the WPT.

An additional implication of this study's findings stems from the longitudinal reliability analyses. As the current WPT rating process is a relatively new program, begun in 2002, the generally high interrater reliability levels are even more impressive. Moreover, the annual and bi-annual trends are generally progressing upward, suggesting that the raters are becoming more consistent in relation to one another's judgments and, perhaps, more comfortable with the scoring process.

Finally, a strong relationship between writing and speaking proficiency measures was found as expected. This finding provides limited validity evidence because one would expect measures of language skill in the same language using the same assessment method to be at least moderately related, especially since writing and speaking are both productive skills. As noted earlier, the evidence would be more powerful if the other skill modalities (i.e., listening and reading), multiple languages, multiple measurement methods, and/or multiple test administrations were included in the study. Because this study used archival data, evidence of this nature was not available. Therefore, our finding of a robust relationship between writing and speaking provides only limited validity evidence suggesting that both the OPI and WPT are assessing related, overlapping constructs. In other words, we found what was expected. If a non-significant relationship had been found, then we would have had a potential validity problem. More data and a well-designed experiment are needed to assess the validity of the WPT thoroughly.

Overall, the results of this preliminary study of reliability and validity are positive for the WPT, especially considering it is a new assessment. As more data becomes available, future research should examine the reliability and validity of the WPT in more depth.

References

American Council on the Teaching of Foreign Language (1986). *ACTFL proficiency guidelines*. Yonkers, NY: Author.

American Council on the Teaching of Foreign Language (2002a). *ACTFL writing proficiency test familiarization guide.* Yonkers, NY: Author.

American Council on the Teaching of Foreign Language (2002b). *ACTFL oral proficiency interview tester certification information application packet.* Yonkers, NY: Author.

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. (1999). ACTFL Proficiency Guidelines—Speaking revised 1999. *Foreign Language Annals, 33*, 13–17.

Breiner-Sanders, K.E., Swender, E., & Terry, R.M. (2002). Preliminary proficiency guidelines—Writing revised 2001. *Foreign Language Annals, 35*, 9–15.

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.

Gardner, W. (1995). On the reliability of sequential data: measurement, meaning, and correction. In John M. Gottman (Ed.), *The analysis of change*. Mahwah, N.J.: Erlbaum.

Kaplan, R. W., & Saccuzzo, D. P. (2001). *Psychological testing: Principles, applications, and issues.* Belmont, CA: Brooks and Cole.

Landis, J. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159-174.

Murphy, K. R., & Davidshofer, C. O. (1994). *Psychological testing: Principles and applications.* Englewood Cliffs, NJ: Prentice-Hall.

Nunnally, J. C. (1978). *Psychometric Theory*, 2nd ed. New York, NY: McGraw Hill Book Company.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory*, 3rd ed. New York, NY: McGraw Hill Book Company.

Society for Industrial and Organizational Psychology, Inc. (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., & Dierdorff, E.C. (2003). Reliability and the ACTFL Oral Proficiency Interview: Reporting indices of interrater consistency and agreement for 19 languages. *Foreign Language Annals, 36*, 507-519.

Swender, E. (ed.) (1999). *ACTFL oral proficiency interview tester training manual.* Yonkers, NY: ACTFL.

Thompson, I. (1996). Assessing foreign language skills: Data from Russian. *Modern Language Journal, 80*, 47-65.

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

Table 1: Interrater Consistency for the WPT

Data Type	*N*	*r*	*R*	□	□	$□_{wt}$
Full sample	509	.935	.935	.959	.890	.865
Spanish only	395	.921	.921	.949	.870	.842

Note. *r* = Pearson correlation; *R* = Spearman rank-order correlation; □ = Goodman-Kruskal gamma; □= Kendall's tau; $□_{wt}$ = Cohen's weighted kappa coefficient; all correlations are significant ($p < .05$).

Table 2: Percentages of Interrater Agreement

Data Type	Agreement	Disagreement Distance		
	Absolute	1 Step	2 Steps	3 Steps
Full sample	80.16	16.90	2.75	0.20
	(408)	(86)	(14)	(1)
Spanish only	77.72	18.73	3.29	0.25
	(307)	(74)	(13)	(1)

Note. Samples sizes are shown in parentheses below each percentage.

Table 3: Full Sample WPT Interrater Agreement by Proficiency Category

Proficiency Category	Agreement	Disagreement Distance		
	Absolute	1 Step	2 Steps	3 Steps
Novice	.	.	.	.
Intermediate	14.73	6.48	0.98	.
	(75)	(33)	(5)	
Advanced	50.10	10.02	1.57	0.20
	(255)	(51)	(8)	(1)
Superior	15.32	0.39	0.20	.
	(78)	(2)	(1)	

Note. Samples sizes are shown in parentheses below each percentage.

Table 4: Interrater Agreement by Proficiency Category for Spanish WPT

Proficiency Category	Agreement	Disagreement Distance		
	Absolute	1 Step	2 Steps	3 Steps
Novice	.	.	.	.
Intermediate	11.65	6.58	1.26	.
	(46)	(26)	(5)	
Advanced	49.62	11.90	1.77	0.25
	(196)	(47)	(7)	(1)
Superior	16.46	0.25	0.25	.
	(65)	(1)	(1)	

Note. Samples sizes are shown in parentheses below each percentage.

Table 5: Full Sample WPT Interrater Agreement by Proficiency Level

Proficiency Level	Agreement	Disagreement Distance		
	Absolute	1 Step	2 Steps	3 Steps
Novice				
Low	.	.	.	.
Mid	.	.	.	.
High	.	.	.	.
Intermediate				
Low	0.25 (1)	.	.	.
Mid	6.86 (28)	9.30 (8)	14.29 (2)	.
High	11.27 (46)	29.07 (25)	21.43 (3)	.
Advanced				
Low	25.74 (105)	30.23 (26)	.	.
Mid	23.04 (94)	20.93 (18)	57.14 (8)	100 (1)
High	13.73 (56)	8.14 (7)	.	.
Superior	19.12 (78)	2.33 (2)	7.14 (1)	.

Note. N = 408 for Absolute agreement; N = 86 for 1 Step; N = 14 for 2 Steps; and N = 1 for 3 Steps.

Table 6: Interrater Agreement by Proficiency Level for Spanish WPT

Proficiency Level	Agreement	Disagreement Distance		
	Absolute	1 Step	2 Steps	3 Steps
Novice				
Low	.	.	.	.
Mid	.	.	.	.
High	.	.	.	.
				.
Intermediate				
Low	.	.	.	.
Mid	3.91	9.46	15.38	.
	(12)	(7)	(2)	
High	11.07	25.68	23.08	.
	(34)	(19)	(3)	
Advanced				
Low	28.01	33.78	.	.
	(86)	(25)		
Mid	22.15	20.27	53.85	100
	(68)	(15)	(7)	(1)
High	13.68	9.46	.	.
	(42)	(7)		
Superior	21.17	1.35	7.69	.
	(65)	(1)	(1)	

Note. N = 307 for Absolute agreement; N = 74 for 1 Step; N = 13 for 2 Steps; and N = 1 for 3 Steps.

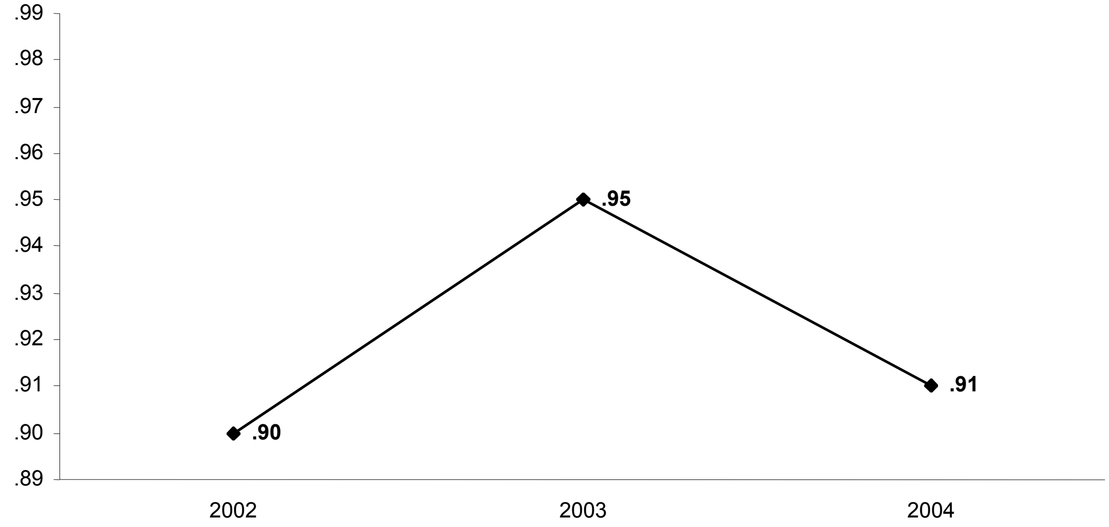

Figure 1: Uncorrected interrater reliabilities; N = 180 in 2002; N = 181 in 2003; N = 34 in 2004.

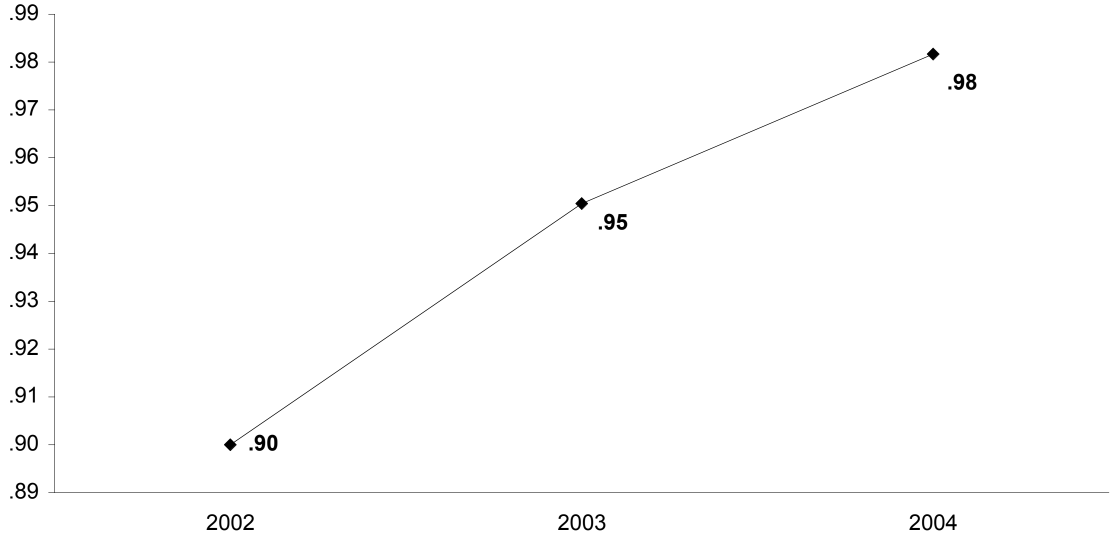

Figure 2: Corrected interrater reliabilities; all estimates were adjusted to 180 rater pairs using the Spearman-Brown formula.

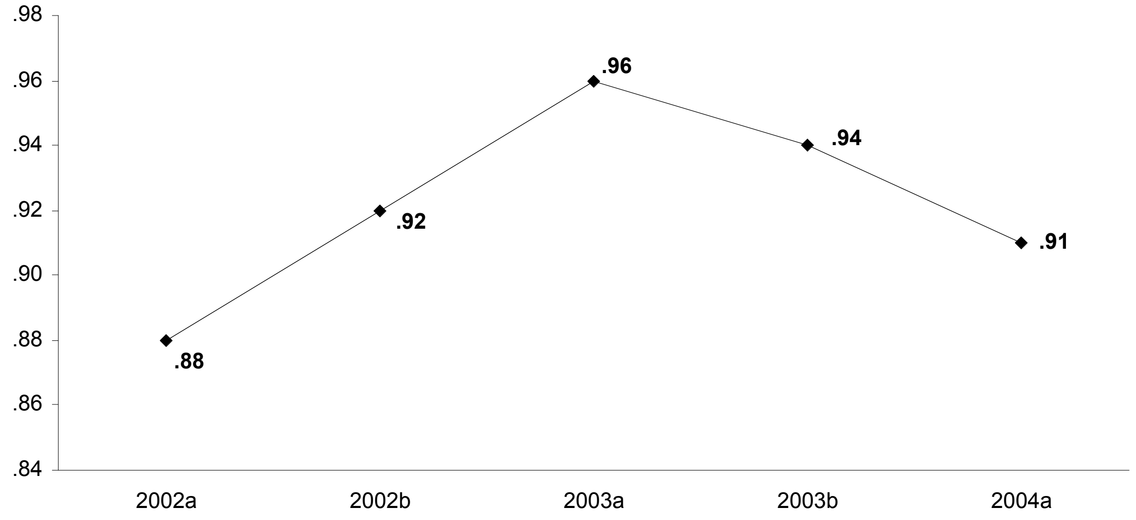

Figure 3: Uncorrected interrater reliabilities using bi-annual categories; N = 70 in 2002a; N = 110 in 2002b; N = 81 in 2003a; N = 100 in 2003b; N = 34 in 2004a.

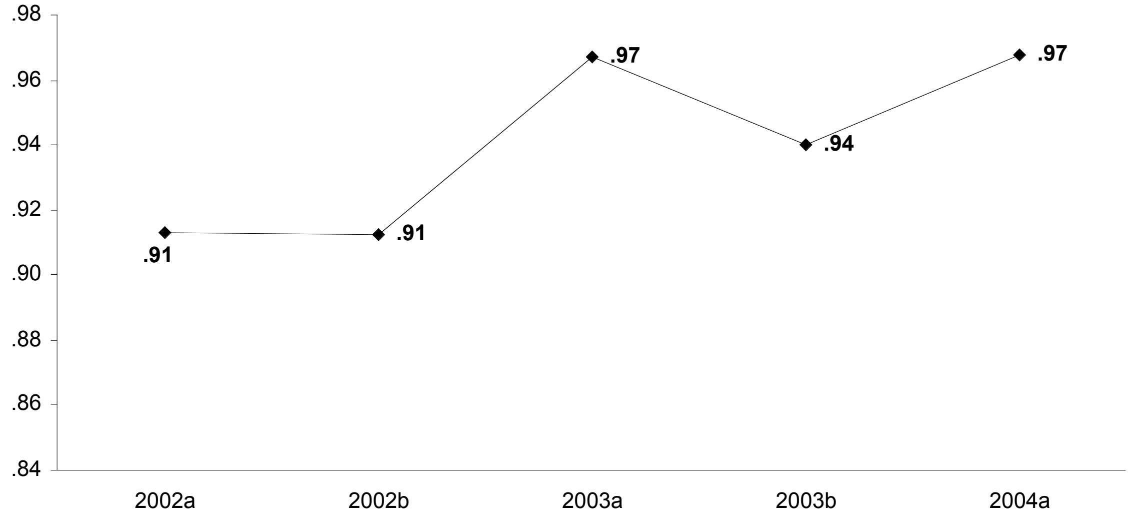

Figure 4: Corrected interrater reliabilities using bi-annual categories; all estimates were adjusted to 100 rater pairs using the Spearman-Brown formula.

About Surface, Ward & Associates

Surface, Ward & Associates (SWA) is an organizational consulting and contract research firm based in Raleigh, NC, and has been in business since 1997. SWA applies the principles, research, and methods of industrial/organizational psychology to assist organizations and their employees in enhancing their performance, solving work-related problems, and addressing workplace issues. SWA consults and conducts research in areas related to training and development, performance measurement and management, organization effectiveness and development, personnel selection, management and leadership, and human resources development and management. SWA is structured as a consulting and research network, allowing our core personnel to utilize numerous associates around the country with specialized expertise and world-class reputations as needed on a project-by-project basis. Our clients have included: Building Construction Products Division, Caterpillar, Inc; North Carolina Cooperative Education Association; seven divisions and the North American Staffing organization of IBM; American Council on the Teaching of Foreign Languages (ACTFL); Special Operations Forces Language Office (SOFLO), and the United States Special Operations Command (USASOC).

Contact Information:

Dr. Eric A. Surface
Principal
Surface, Ward & Associates
116 N. West Street
Suite 230
Raleigh, NC 27603
919.836.9970
919.454.4824
esurface@bellsouth.net