

Surface, E. A., Dierdorff, E. C., & Poncheri, R. M. (2006, May). *The “write” stuff: A preliminary assessment of a writing proficiency test*. Paper presented at the 21st annual conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

The “Write” Stuff: Preliminary Assessment of a Writing Proficiency Test

Eric A. Surface
Surface, Ward, & Associates

Erich C. Dierdorff
DePaul University

Reanna M. Poncheri
North Carolina State University
Surface, Ward, & Associates



MAY 2006

APPROVED FOR PUBLIC RELEASE;
DISTRIBUTION UNLIMITED

Copyright Notice

This document and its content is copyright ©1997-2010 of SWA Consulting Inc. All rights reserved.

Any redistribution or reproduction of part, or the entire document in any form is prohibited except for: (1) you may print or download to a local hard disk extracts for your personal and non-commercial use only, and (2) you may copy the content to individual third parties for their personal use, but only if you acknowledge the website and author(s) as the source of the material. You may not, except with our express written permission, distribute or commercially exploit the content, nor may you transmit it or store it on any other website or other form of electronic retrieval system.

The "Write" Stuff: Preliminary Assessment of a Writing Proficiency Test

Eric A. Surface
Surface, Ward, & Associates

Erich C. Dierdorff
DePaul University

Reanna M. Poncheri
North Carolina State University/Surface, Ward, & Associates

The importance of foreign language skills is clearly on the rise in the U.S. largely due to economic expansion, national security, and workforce demographic shifts. This study examines the psychometric properties of a newly created writing proficiency test designed by the American Council for the Teaching of Foreign Languages.

Recent historical and political events (e.g., September 11th) as well as the expanding global economy and the influx of native Spanish-speakers in the workforce have highlighted the importance of foreign language study in American society. However, while the demand for proficiency in foreign language is high and projected to increase, there are strong indications that this need is not being met. Weber (2005) recently suggested that the U.S. military, Central Intelligence Agency, National Security Agency, and Federal Bureau of Investigation are some organizations which have suffered the most from this deficiency and are much in need of individuals with proficiency in foreign languages. In response to this need, there have been several calls for increased attention and cooperation among interested parties to improve the current state of foreign language learning in American society (e.g., Rovira, 2003), and some private sector organizations have started to offer foreign language and English as a second language training to counter these deficiencies (e.g., Hammers, 2005).

At the National Language Conference in June 2004, representatives from several different areas of American society (e.g., government, industry, academia, and language associations) met to discuss the importance of foreign language. From this meeting, a white paper entitled "A Call to Action for National Foreign Language Capabilities" was produced (U.S. Department of Defense [DoD], 2005). In this paper, several recommendations were made to address the shortage of foreign language

capabilities in the United States. Two recommendations are particularly relevant for the present study: 1) "a system for the standardized assessment of achievement and proficiency in foreign languages, especially at high proficiency levels" and 2) "a system of assessments to test and certify the knowledge, skills, and/or abilities of language professionals and practitioners, such as instructors, trainers, translators, interpreters, and other language specialists" (U.S. DoD, 2005, p. 4). This call for action reinforces the need for research related to language skill assessment to ensure national security and future economic success.

Due to the growing need for foreign language capabilities, there is an increase in the demand for foreign language instructors. In order to become a foreign language instructor, certification and licensure are often required. Moreover, issues related to teacher certification have received additional attention due to the *No Child Left Behind* legislation. As part of this initiative, teachers are now required to demonstrate that they are "highly qualified," which means that they must have: 1) a bachelor's degree, 2) full state certification/licensure, and 3) prove that they know each subject they teach (U.S. Department of Education, 2004). Although specific certification requirements vary from state to state, some level of certification/licensure is required for foreign language instructors to be in compliance with this legislation.

In 2002, the American Council on the Teaching of Foreign Languages (ACTFL) published revised

guidelines for writing proficiency in languages (Breiner-Sanders, Swender & Terry, 2002). These guidelines were used to create the ACTFL Writing Proficiency Test (WPT). In accordance with the *Standards for Educational and Psychological Testing* (1999), published by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME), test publishers must document the psychometric properties of their instruments by providing empirical evidence of reliability and validity. As a new assessment, preliminary reliability and validity evidence must be provided for the WPT (AERA et al., 1999). Thus, the primary purpose of our study was to assess the psychometric properties of this instrument. This research represents the first assessment of reliability and validity evidence for the WPT.

Research Background

Writing Proficiency Research

Although the WPT is a new assessment with no previous research, there is evidence from two prior studies using writing proficiency measures developed from the original ACTFL proficiency guidelines (ACTFL, 1986). However, the findings from these two studies should be viewed with caution because very little information regarding the development and nature of the writing proficiency assessments were provided. The first study presented the results of a multitrait-multimethod validation study of the ACTFL Oral Proficiency Interview (OPI). In addition to the OPI (speaking), the study included measures of writing, listening and reading in French and English as a second language (Dandonoli & Henning, 1990). The interrater reliability (Pearson r) for the writing test was .87 for the English sample ($n = 59$) and .89 for the French sample ($n = 60$). Although, the writing and speaking tests were not rated by the same individuals, correlations between the speaking raters and writing raters were .85, .86, .92, and .88 for the English sample and .85, .80, .84, and .80 for the French sample.

A second study by Thompson (1996) presented results from an assessment of speaking, reading, listening, and writing proficiency of students studying Russian using tests based on the ACTFL proficiency guidelines (ACTFL, 1986). The writing, listening, and reading tests were specifically developed for the study. Speaking was measured using ACTFL OPI testers. The interrater reliability (Pearson r) of the writing test in the pilot study was reported as .88. Two additional university samples were examined with interrater reliabilities (Pearson r)

reported as .91 (University of Iowa) and .72 (Middlebury College). The relationship between the writing and speaking tests was reported as .64.

ACTFL Writing Proficiency Guidelines

The ACTFL proficiency guidelines are global characterizations of integrated performance across language skill modalities, including writing. As aforementioned, the *ACTFL Proficiency Guidelines—Writing (Revised 2001)* are the basis for the ACTFL WPT. These guidelines were first published in 1986 and revised in 2001 following a similar revision of the speaking guidelines in 1999 (ACTFL, 1986; Breiner-Sanders et al., 2002; Breiner-Sanders, Lowe, Miles & Swender, 1999). As with the speaking guidelines, the writing guidelines specify four major levels of proficiency (i.e., *Superior*, *Advanced*, *Intermediate*, and *Novice*) that are divided into ten hierarchical sublevels: *Superior*, *Advanced High*, *Advanced Mid*, *Advanced Low*, *Intermediate High*, *Intermediate Mid*, *Intermediate Low*, *Novice High*, *Novice Mid*, and *Novice Low* (ACTFL, 2002). Descriptions of the four major levels of writing proficiency are presented in Table 1. Highly trained, certified raters follow a standardized protocol that incorporates these proficiency guidelines when assigning one of the aforementioned proficiency levels to each test taker’s writing sample.

Research Objectives

Interrater Reliability and 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). In the specific case of WPT ratings, 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 raters involved in a particular judgment process. When interrater reliability estimates are high, the interpretation is a large degree of consistency across 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., test takers). 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 are ordinal or nominal.

The *Standards* (AERA et al., 1999) provide a number of guidelines designed to help users evaluate the reliability data. These *Standards* also provide guidance to test publishers as to what evidence should be made available to users. Importantly, empirical evidence of reliability that is relevant to the particular assessment should be provided. In the case of the WPT, the most relevant reliability data are interrater reliability and consistency coefficients, since the WPT is based on raters’ judgments. To investigate the interrater reliability and consistency of the WPT, the following research questions were proposed:

Q1: What is the reliability of the WPT across different languages?

Q2: What is the reliability of the WPT for the most commonly tested language (Spanish)?

Q3: Has the reliability of the WPT changed over time since its inception in 2002?

Validity

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” (AERA et al., 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*, one category—“evidence based on relations to other variables”—is reviewed in this study because it provides the basis for the validity examinations of the limited WPT data currently available.

The statistical relationship of a test to established measures of the same construct, related constructs, or different constructs can provide validity-related evidence (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. In this study, a strong correlation between the ACTFL WPT and OPI would provide some validity evidence because one would expect two language skill measures in the same language 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. Therefore, although a strong relationship between the WPT and OPI will provide only limited validity evidence, the lack of a

statistically significant relationship with an established assessment of speaking proficiency (OPI) might call the WPT into question. The following research question was proposed to explore convergent validity evidence for the WPT:

Q4: What is the relationship between WPT and OPI scores?

Method

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. These tests were completed between January 2002 and April 2004 and were for the purpose of teacher certification in two states (Connecticut and Pennsylvania). This study used data from tests in five 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. Four hundred and sixty cases had OPI scores as well as WPT scores, allowing us to assess the correlation between the two skill modalities.

WPT Rating Procedure

The WPT is a standardized global assessment of functional writing ability in a language and measures a test taker’s ability to write spontaneously in a language without the opportunity to revise responses and/or use reference or editing tools. The writing performance of the test taker on the WPT is compared to standardized criteria (ACTFL, 2002) in order 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 (e.g., *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” (ACTFL, 2002, 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 writing task covers multiple aspects (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, raters consider the following criteria from the wider perspective of how the test taker contributes 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 raters must go through a rigorous training program to become certified.

OPI Rating Procedure

The ACTFL OPI assessment procedure consists of four phases (Warm Up, Level Checks, Probes, and Wind Down) that are designed to efficiently elicit a ratable sample (Swender, 1999). As stipulated by the procedure, a pair of judges rates each case. In all cases, the first rater conducts and audiotapes the interviews. Subsequently, this rater judges each interviewee's speaking proficiency from the tape at some later time. Next, the taped interview is independently rated by the second rater. If the independent ratings provided by the rating pair disagree, a third rater is assigned as an arbitrator to rate the interview tape and provide a rating. This rater does not know the previously assigned scores, nor does the rater know he or she is the third rater. All raters are trained and certified and use the ACTFL rating scale to describe the proficiency levels of the interviewees. A number of studies have demonstrated the reliability (e.g., Magnan, 1987; Thompson, 1995) and the validity (e.g., Dandonoli and Henning, 1990) of the ACTFL OPI. In the most recent reliability study, Surface and Dierdorff (2003) found the interrater reliability ranged from .94 to .99 (Pearson r) across 19 different languages with an overall coefficient of .98, suggesting more than adequate reliability.

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. 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.

The interrater consistency measures used in this study were as follows: the Pearson correlation (r), the Spearman rank-order correlation (R), Kendall's tau,

Goodman and Kruskal's gamma, and Cohen's kappa (weighted). In addition, raw percentages of agreement were calculated to assess the extent to which raters display perfect agreement. This serves as an absolute agreement estimate of interrater consistency and was 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).

In order to assess the relationship between writing and speaking proficiency and provide evidence of convergent validity, a Pearson correlation (r) and a Spearman rank-order correlation (R) were computed.

Results

Table 2 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 3 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%). Similarly, raters judging the Spanish version of the WPT were 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., Intermediate-Low to Intermediate-Mid).

Tables 4 and 5 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 6 and 7 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 convergent validity evidence, 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 2 all fall within "acceptable" ranges as described by the relevant literature. For example, the interrater reliability estimates were above the .90 levels, which have been recommended for applied research (Kaplan & Saccuzzo, 2001; Nunnally & 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, in which the majority 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. This finding provides some validity evidence because one would expect two language skill measures (writing and speaking) in the same language to be at least moderately related, especially since writing and speaking are both productive skills. 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 unavailable. Therefore, our finding of a robust relationship between writing and speaking provides only limited evidence suggesting that both the OPI and WPT are assessing related, overlapping constructs. In other words, we found what was expected. More data and a well-designed experiment are needed to assess the validity of the WPT more thoroughly.

Overall, the results of our study are positive for the WPT, especially considering its newness. This preliminary reliability and validity evidence suggests that the WPT is consistently measuring writing proficiency and that organizations can use the WPT with confidence for assessing writing proficiency. As more data become available, future research should continue to examine the reliability and validity of the WPT. Studies employing construct- and criterion-related validity strategies would be especially useful. The development of the WPT is an important step in addressing the identified need for standardized assessments of foreign language proficiency.

References

- American Council on the Teaching of Foreign Language (1986). *ACTFL proficiency guidelines*. Yonkers, NY: Author.
- American Council on the Teaching of Foreign Language (2002). *ACTFL writing proficiency test familiarization guide*. 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. Nesselrode (eds.), *Handbook of multivariate experimental psychology: Perspectives on individual differences*, 2nd ed. (pp. 131–203). New York: Plenum Press.
- 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.
- 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.
- Hammers, M. (2005). Wyndham looks to leap language gap. *Workforce Management*, p. 17.
- Kaplan, R. W., & Saccuzzo, D. P. (2001). *Psychological testing: Principles, applications, and issues* (5th ed.). Belmont, CA: Brooks and Cole.
- Landis, J. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159-174.
- Magnan, S. S. (1987). Rater reliability of the ACTFL Oral Proficiency Interview. *The Canadian Modern Language Review*, 43, 267–76.
- Nunnally, J. C., Bernstein, I. H. (1994). *Psychometric Theory*, 3rd ed. New York, NY: McGraw Hill Book Company.
- Rovira, J. D. (2003). *Importance of foreign language capabilities*. (ALMAR 072/03). Washington, DC: Author. Retrieved September 9, 2005, from <http://www.marines.mil/almar/almar2000.nsf/1babcf316f87f38c852569b8008017e7/952d4b8516b446a585256df8006a75f9?OpenDocument>
- Stanley, J. C. (1971). Reliability. In R. L. Thorndike (ed.), *Educational measurement*, 2nd ed. (pp. 356–442). Washington, DC: American Council on Education.
- Swender, E. (ed.) (1999). *ACTFL oral proficiency interview tester training manual*. Yonkers, NY: ACTFL.
- Surface, E. A., & Dierdorff, E. C. (2003). Reliability and the ACTFL Oral Proficiency Interview: Reporting indices of interrater consistency and agreement for 19 languages. *Foreign Language Annals*, 36 (4), 507-519.
- 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, 407–22.
- 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.

United States Department of Defense. (2005, February 1). A call to action for national foreign language capabilities. *Proceedings of the National Language Conference*, Retrieved April 1, 2005, from http://www.nlconference.org/docs/White_Paper.pdf

United States Department of Education. (2004, March 14). New "No Child Left Behind" flexibility: Highly qualified teachers. Retrieved September 10, 2005 from, <http://www.ed.gov/nclb/methods/teachers/hqtflexibility.pdf>

Weber, G. (2004). English RULES. *Workforce Management*, 83(5), p. 47-51.

Table 1

Description of the Four Major Levels of Writing Proficiency According to ACTFL Guidelines

Level of Proficiency	Description
Superior	<ul style="list-style-type: none"> • 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. • 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.
Advanced	<ul style="list-style-type: none"> • 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 • Their writing is comprehensible to all native speakers due to breadth of generic vocabulary and good control of the most frequently used structures.
Intermediate	<ul style="list-style-type: none"> • Can meet a range of simple and practical writing needs. • Can communicate simple facts and ideas. • Their writing is comprehensible to those accustomed to the writing of non-natives.
Novice	<ul style="list-style-type: none"> • Can produce lists and notes and limited formulaic information on simple forms and documents. • Their writing is typically limited to words, phrases and memorized material.

Note. From *ACTFL Writing Proficiency Test Familiarization Guide* (ACTFL, 2002).

Table 2

Interrater Consistency for the WPT

Data Type	<i>N</i>	<i>r</i>	<i>R</i>	Γ	τ	K_{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; Γ = Goodman-Kruskal gamma; τ = Kendall's tau; K_{wt} = Cohen's weighted kappa coefficient; all correlations are significant ($p < .05$).

Table 3

Percentages of Interrater Agreement

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

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

Table 4

Full Sample WPT Interrater Agreement by Proficiency Category

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

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

Table 5

Interrater Agreement by Proficiency Category for Spanish WPT

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

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

Table 6

Full Sample WPT Interrater Agreement by Proficiency Level

Proficiency Level	Agreement		Disagreement Distance	
	Absolute	1 Step	2 Steps	3 Steps
<i>Novice</i>				
Low
Mid
High
<i>Intermediate</i>				
Low	0.25 (1)	.	.	.
Mid	6.86 (28)	9.30 (8)	14.29 (2)	.
High	11.27 (46)	29.07 (25)	21.43 (3)	.
<i>Advanced</i>				
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)	.	.
<i>Superior</i>				
	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 7

Interrater Agreement by Proficiency Level for Spanish WPT

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

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

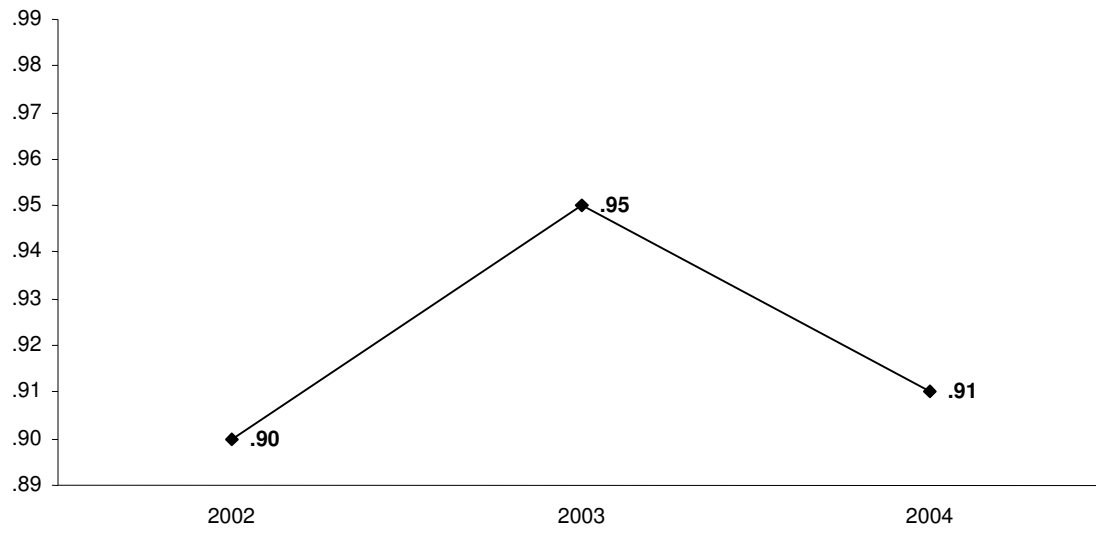
Figure Captions

Figure 1. Uncorrected interrater reliabilities.

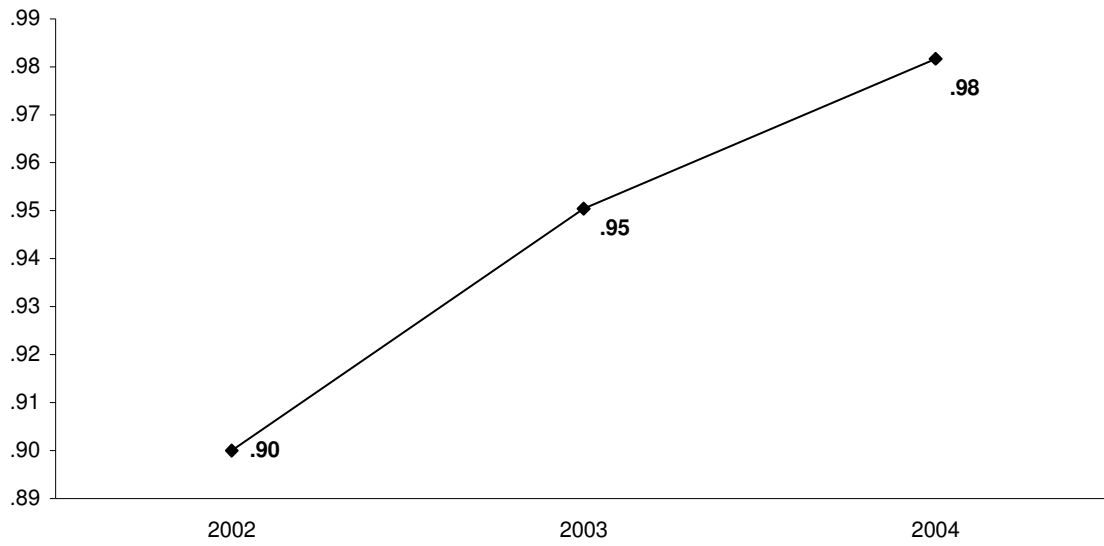
Figure 2. Corrected interrater reliabilities.

Figure 3. Uncorrected interrater reliabilities using bi-annual categories.

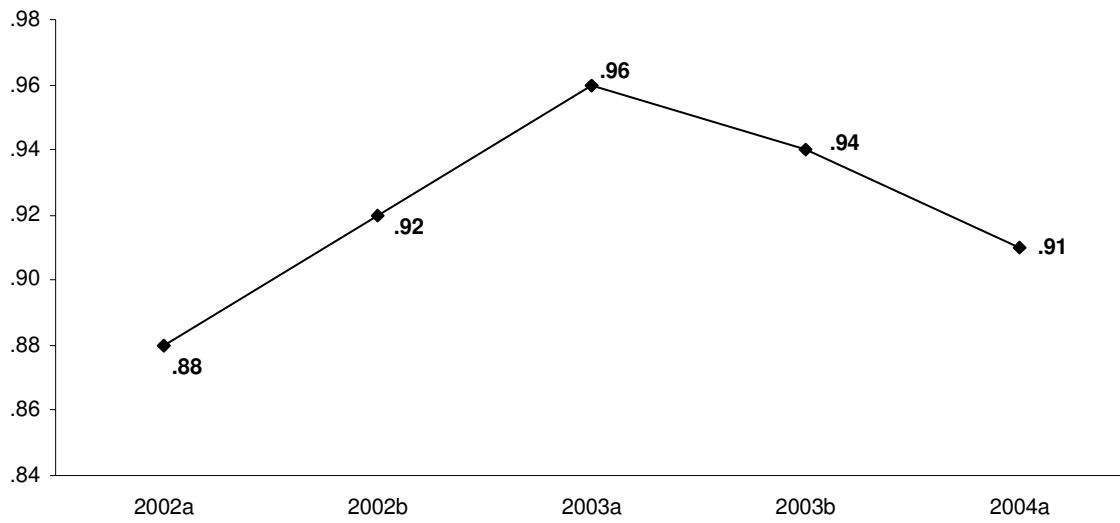
Figure 4. Corrected interrater reliabilities using bi-annual categories.



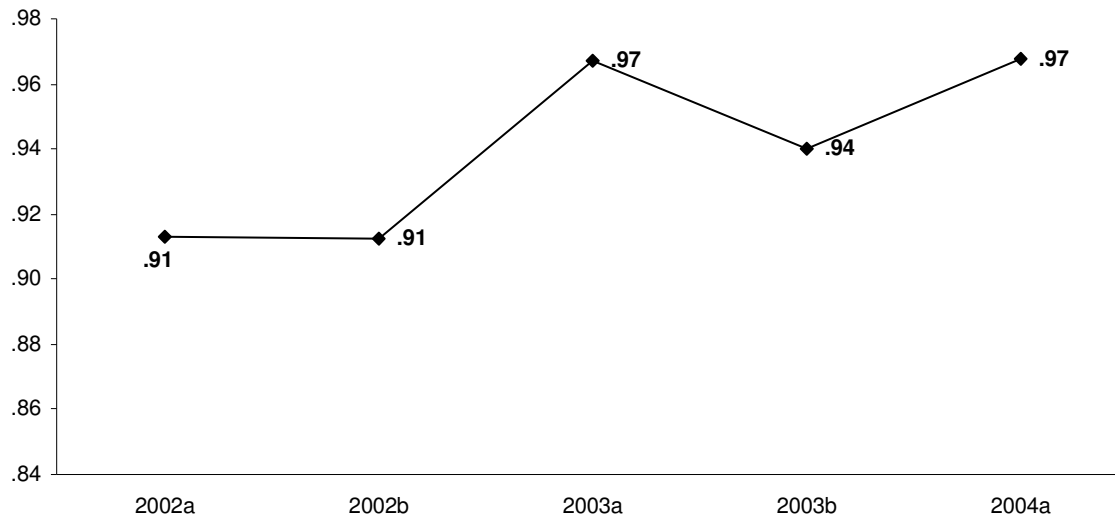
Note. N = 180 in 2002; N = 181 in 2003; N = 34 in 2004.



Note. All estimates were adjusted to 180 rater pairs using the Spearman-Brown formula.



Note. N = 70 in 2002a; N = 110 in 2002b; N = 81 in 2003a; N = 100 in 2003b; N = 34 in 2004a.



Note. All estimates were adjusted to 100 rater pairs using the Spearman-Brown formula.

ABOUT SWA CONSULTING INC.

SWA Consulting Inc. (formerly Surface, Ward, and Associates) provides analytics and evidence-based solutions for clients using the principles and methods of industrial/organizational (I/O) psychology. Since 1997, SWA has advised and assisted corporate, non-profit and governmental clients on:

- Training and development
- Performance measurement and management
- Organizational effectiveness
- Test development and validation
- Program/training evaluation
- Work/job analysis
- Needs assessment
- Selection system design
- Study and analysis related to human capital issues
- Metric development and data collection
- Advanced data analysis

One specific practice area is analytics, research, and consulting on foreign language and culture in work contexts. In this area, SWA has conducted numerous projects, including language assessment validation and psychometric research; evaluations of language training, training tools, and job aids; language and culture focused needs assessments and job analysis; and advanced analysis of language research data.

Based in Raleigh, NC, and led by Drs. Eric A. Surface and Stephen J. Ward, SWA now employs close to twenty I/O professionals at the masters and PhD levels. SWA professionals are committed to providing clients the best data and analysis with which to make solid data-driven decisions. Taking a scientist-practitioner perspective, SWA professionals conduct model-based, evidence-driven research and consulting to provide the best answers and solutions to enhance our clients' mission and business objectives. SWA has competencies in measurement, data collection, analytics, data modeling, systematic reviews, validation, and evaluation.

For more information about SWA, our projects, and our capabilities, please visit our website (www.swa-consulting.com) or contact Dr. Eric A. Surface (esurface@swa-consulting.com) or Dr. Stephen J. Ward (sward@swa-consulting.com).