ACTFL
AMERICAN COUNCIL ON THE TEACHING OF FOREIGN LANGUAGES

AMERICAN COUNCIL ON THE TEACHING OF FOREIGN LANGUAGES
1001 North Fairfax Street, Suite 200 | Alexandria, VA 22314 | P 703-894-2900 | F 703-894-2905
445 Hamilton Avenue, Suite 1104 | White Plains, NY 10601-1832 | P 914-963-8830 | F 914-963-1275

www.actfl.org | www.leadwithlanguages.org | facebook.com/actfl | @actfl

Examination Evaluation of the ACTFL WPT® in English, Russian, and Spanish for the ACE Review

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

Prepared by
Stephen Cubbellotti, Ph.D.
Independent Psychometric Consultant

EXECUTIVE SUMMARY

This report documents the American Council on the Teaching of Foreign Languages (ACTFL) Writing Proficiency Test (WPT®) from 2012 to 2014 to satisfy a review requirement of the American Council of Education College Credit Recommendation Service (CREDIT) program. The ACTFL WPT® is an assessment of functional writing proficiency in a foreign language which is evaluated by trained and certified experts in a writing format across numerous languages.

The structure of this document is outlined to address several areas including: general test information, item/test content development, reliability information, validity information, scaling and item response theory procedures, validity of computer administration, cut-score information, and other recommended items.

METHOD

ACTFL and LTI have an extensive collection of resources available publically that documents the rigor of defining language competency as well as the precision in their assessments. All documentation cited is publically available and citations for these resources are given in the bibliography at the end of this document. The reliability information section is the only section which contains uniquely generated statistics for the purposes of this study. An outline of the results can be found below.

Given the ordinal nature of the ACTFL proficiency scale and ACTFL WPT® scores, inter-rater reliability was measured by the Spearman's *R* correlation, which is a coefficient of reliability appropriate for ordinal data. Inter-rater agreement was measured by the extent to which ratings exhibited absolute (i.e., exact) and/or adjacent (i.e., +/- one level) agreement. The combination of Spearman's *R* and absolute/adjacent agreement results provides sufficient information about reliability.

Comparisons of ACTFL WPT® inter-rater reliability and agreement were made across three languages: English, Russian, and Spanish. Comparisons were also made across language categories (i.e., language difficulty) and interview years (i.e., 2012 to 2014 in this sample). For inter-rater agreement, rater concordance was further investigated by major proficiency level and sub-level.

FINDINGS

The ACTFL WPT® exceeded the minimum inter-rater reliability and agreement standards. Further, the findings are fairly consistent with results from SWA Consulting (2012) on the Arabic, Russian, and Spanish versions of the exam; thus indicating the ACTFL WPT® process yields relatively stable reliability results over time.

Overall, the findings support the reliability of the ACTFL WPT® as an assessment of writing proficiency. Areas for continued improvement include increasing rater agreement at the Advanced Mid sublevel and the Novice High-Intermediate Low border. Findings are presented in more detail in the report.

Table of Contents

General Test Information

Rationale and Purpose of the test

The ACTFL WPT® is an assessment of functional writing ability that measures how well a person spontaneously writes in the target language in response to four carefully constructed prompts dealing with practical, social, and professional writing tasks that are encountered in true-to-life informal and formal contexts. The individual whose writing proficiency is being evaluated is presented with tasks and contexts that represent the range of proficiency levels from Novice to Superior according to the *ACTFL Proficiency Guidelines 2012 – Writing.*

All instructions and prompts are written in English; responses are written in the target language. The ACTFL WPT® can be administered in booklet form or via the Internet. The ACTFL WPT® typically lasts for 80 minutes (with an additional 10 minutes allotted for instructions). However, the test can take anywhere between 40-90 minutes depending on the proficiency range being assessed and the writing competence of the test-taker. The beginning of the WPT® presents the candidate with a Self-Assessment consisting of six different descriptions of how well a person can write in a language; test takers select the description they feel most accurately describes their writing ability. They are also presented with a Background Survey. The Self-Assessment determines which one of three WPT® test forms is generated; the Background Survey determines some of the content areas for the individualized assessment.

In all forms of the WPT®, there are four separate requests, each of which encompasses multiple writing tasks, (i.e. descriptive, informative, narrative, and persuasive writing). Each request describes the audience, context, and purpose of the writing task, as well as a recommended length for the response. Depending on the form of the test, the prompts that are presented to the writer are designed to elicit writing at the Intermediate, Advanced and/or Superior levels, across a variety of contexts and content areas.

Further details of this test can be found in the *ACTFL – Writing Proficiency Test – Familiarization Manual*.

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

Principle Item writers for the ACTFL WPT®:

- Ray Clifford, Ph. D. Brigham Young University
- Pardee Lowe, Jr., Ph D (Ret.)
- John Lett Ph.D (Ret.) Defense Language Institute Foreign Language Center
- Lucia Caycedo Garner, Ph. D. (Emerita) University of Wisconsin – Madison
- Maria Teresa Garreton, Ph. D. Chicago State University
- Karen Breiner Sanders Ph.D. (Emerita) Georgetown University

Types of scores reported for examinees

Examinee scores are reported as the major level and sublevel according to the *ACTFL Proficiency Guidelines 2012 - Writing*. The ACTFL Guidelines describe the tasks that a writer can handle at each level, as well as the content, context, accuracy, and discourse types associated with tasks at each level.

The description of each major level is representative of a specific range of abilities. They also present the limits that writers encounter when attempting writing tasks at the next higher major level.

While the *ACTFL Proficiency Guidelines* are comprised of five major levels of proficiency – Novice, Intermediate, Advanced, Superior, and Distinguished – the current exam only tests through Superior. Together these levels form a hierarchy in which each level subsumes all lower levels. The major levels of Advanced, Intermediate, and Novice are divided into High, Mid, and Low sublevels. ACTFL publically shares their guidelines for defining the levels of proficiency describing what examinees have displayed during their examination.

Directions for scoring and procedures and keys

Certified WPT® Raters evaluate the entire writing sample holistically, determining whether the level of the response meets, does not meet, or exceeds the expectations for the targeted level according to the Assessment Criteria for the major level. A rating at any major level is arrived at by the sustained performance across ALL the criteria of the level. The sublevel is determined by the quality of the performance at that level and the proximity to the next higher major level. Once a preliminary rating is reached, the rater compares the sample to the descriptions in the *ACTFL Proficiency Guidelines 2012 – Writing* and selects the best match between the sample and proficiency descriptors.

The assessment criteria used to evaluate the ACTFL WPT® is provided in the chart below:

Proficiency Level	Tasks and Functions	Context / *Content*	Text Type	Accuracy
Superior	Can write most correspondence (memos, letters, summaries, reports). Can write in detail and explain complex matters, state opinions, present supporting arguments, and compose hypotheses and conjectures.	Most formal and informal settings. *Practical, professional, and social topics treated both concretely and abstractly*	Writes a clearly organized and articulated text that can extend from several paragraphs to pages.	Demonstrates no patterned errors in basic structures, vocabulary, punctuation, or spelling. Some occasional errors may occur, particularly in low-frequency structures, which rarely disturb the native reader.
Advanced	Can write informal and some routine formal correspondence and reports that require simple narratives, descriptions, and summaries of a factual nature. Can narrate and describe in major time frames, at times uses paraphrase and elaboration to provide clarity.	Informal settings and some routine formal settings on familiar topics. *Topics of personal and general interest.*	Writes a connected, cohesive text of at least a paragraph in length. Can extend to two or more paragraphs in length on familiar topics.	Expresses meaning that is comprehensible to those unaccustomed to the writing of non-natives, primarily through generic vocabulary, with good control of the most frequently used basic structures and punctuation.
Intermediate	Can meet practical writing needs, i.e., notes, simple messages, and requests for information. Can ask and respond to straightforward questions.	Routine informal settings and limited tasks involving the exchange of simple information. *Predictable, familiar topics related to self and daily routines and activities.*	Writes a loosely connected text made up of a collection of primarily discrete sentences that may or may not be presented in the semblance of a paragraph.	Expresses meaning through vocabulary and basic structures that is comprehensible to those accustomed to the writing of non-natives.
Novice	Can write words, lists, and notes and limited formulaic information to communicate the most basic information.	The most common informal settings. *Most common aspects of self and daily life.*	Words, lists, phrases, and some limited formulaic information.	May be difficult to comprehend, even for readers accustomed to dealing with non-native writers.

ACTFL Certified WPT® Raters are highly specialized language professionals who have completed a rigorous training process that concludes with a rater's demonstrated ability to consistently rate samples with a high degree of reliability. A pre-requisite for becoming a Certified WPT® Rater is that one is already a Certified ACTFL OPI Tester.

Certified WPT® Raters are always expected to respect and follow WPT® rating protocol. Confidentiality and exclusivity are important practices for all Certified WPT® Raters. Every Rater agrees to respect the rules and regulations regarding WPT® rating, and the exclusivity of the WPT® as ACTFL property. Work with the WPT® rating process must be done exclusively through Language Testing International, the ACTFL Testing Office. Raters are required to follow all WPT® procedures and guidelines, as well as any other information received on behalf of LTI and ACTFL.

Item/Test Content Development

Specifications that define the domain(s) of content, skills, and abilities that the test samples

The ACTFL WPT® utilizes a Background Survey and a Self-Assessment to determine appropriate topics and linguistic levels for the test taker. The Background Survey is a questionnaire which elicits information about the test taker's work, school, home, personal activities, and interests. The survey answers determine the pool of prompts from which the computer will randomly select topics for writing tasks. The response to the Self-Assessment ensures that the test taker is provided with tasks that are appropriate to his/her linguistic ability. The variety of topics, the types of questions, and the range of possible computer-generated combinations allow for individually designed assessments. Even if two test takers select the same combination of Background Survey and Self-Assessment responses, the resulting tests will be different.

The ACTFL Guidelines describe the tasks that writers can handle at each level, as well as the content, context, accuracy, and discourse types associated with tasks at each level. They also present the limits that writers encounter when attempting to function at the next higher major level. Further descriptions of each level are available online.

Statement of test's emphasis on each of the content, skills, and ability areas

The tested content, skills and ability areas are based on the Assessment Criteria for Writing and the descriptions contained in the *ACTFL Proficiency Guidelines - Writing*. The ACTFL WPT® measures how well a person spontaneously writes in the target language in response to carefully constructed prompts dealing with practical, social, and professional topics that are encountered in true-to-life informal and formal contexts. These tasks range from writing short messages and invitations (Intermediate level), to writing paragraph-length narrations and descriptions in major time frames (Advanced), to dealing abstractly with current issues of general interest, supporting one's opinion and hypothesizing in multi-paragraph, essay-like discourse (Superior level).

Rationale for the kinds of tasks (items) that make up the test

The tasks of the ACTFL WPT® reflect the linguistic writing functions of each of the major levels of proficiency as described in the *ACTFL Proficiency Guidelines 2012 – Writing*. Test takers are presented with writing tasks that span two or more major levels across a variety of content areas. In this way, the sample that is produced provides sufficient evidence of a writer's patterns of linguistic strengths (their "floor performance") and weaknesses (their "ceiling").

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

The *ACTFL Proficiency Guidelines – 2012 - Writing* describe the range of contents and contexts a writer at each major level should be able to handle. This was the main driver behind the topics generated for each level. Additionally, candidates fill out a Background Survey which elicits information about the test taker's work, school, home, personal activities, and interests. The survey answers determine the pool of prompts from which the computer will randomly select topics for writing tasks. The variety of topics, the types of questions, and the range of possible computer-generated combinations allows for individually designed assessments. Even if two test takers select the same combination of Background Survey responses, the resulting tests will be different. Based on the Background Survey, questions are pulled that reflect the background and interests of the candidate.

Information on the currency and representativeness of the test's items

The representativeness of the items in a test is guaranteed by providing a diversity of topics, subtopics, genres, domains and rhetorical organization so that the test can provide ample evidence of the proficiency of the test-taker across a broad spectrum of target language use domains.

Some of the topics from which the test-taker may choose include: education, business, history, languages, the environment, sports, entertainment, popular culture, current events. New topics are always being developed and old ones revised as they become less current.

Description of the item sensitivity panel review

The use of a Background Survey allows the test taker to avoid the selection of items which may be insensitive or irrelevant for the test taker. In an effort to ensure that test-takers are not offended or made uneasy while taking a WPT®, item writers are instructed to avoid sensitive topics (e.g., immigration, national origin, sexual preference, religion, marital status, racism, etc.) when developing WPT® writing prompts.

Whether and/or how the items pre-tested (field tested) before inclusion in the final form

As each WPT® is generated based on the test taker's responses to the Background Survey and Self-Assessment, there is no standard "final form." However, items are pre-tested before they are added to the item pool; items that do not elicit the expected level of response are modified or removed.

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

All WPT® items target the linguistic tasks, contexts and content areas as described in the *ACTFL Proficiency Guidelines 2012 – Writing.*

Reliability Information

Previous studies provided psychometric support for the use of writing proficiency measures developed according to the *ACTFL Proficiency Guidelines*.

In 2004, Dandonoli and Henning presented the results of a multitrait-multimethod validation study, which included tests of speaking, writing, listening and reading in French and English as a Second Language (ESL). The inter-rater reliabilities for the writing test for the English and French samples were strong (reported Pearson *r*s of .87 and .89, respectively).

Surface and Dierdorff (2004) presented results from a reliability and validity study on the WPT® traditional "non-adaptive" version. 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 agreement indicated that for the full sample, the majority of judges provided identical scores (80% perfect agreement). Similar results were found for the Spanish-only sample as well (78% perfect agreement). The longitudinal reliability trends indicate that the inter-rater reliability has generally increased during the time the revised procedures have been in place (as of the date of the report).

Bärenfänger and Tschirner (2011) examined the ratings of 166 internet English WPT®s that were administered in Korea in November and December 2010 to adult second-language learners of English. As opposed to the test takers in Surface and Dierdorff (2004) study, these test-takers took a new adaptive version of the internet ACTFL WPT® that requires test-takers to self-identify their range of proficiency through Self-Assessment Statements. The researchers found a high level of interrater consistency with a Spearman's rho of 0.917 and a summed absolute and adjacent interrater agreement of greater than 95% for all levels of ability. Bärenfänger and Tschirner summarized their results by indicating that changing the format of the test has not changed its rating reliability either in a more positive or negative way.

Most recently, SWA consulting (2012) analyzed the Arabic, Russian, and Spanish ACTFL WPT®s and found Spearman Rs exceeded the standard for use, ranging from 0.92 to 0.98 across languages and years analyzed. In addition, overall inter-rater agreement was higher than 70% for all languages and lowest for Novice High. These results were consistent across languages and highest for Novice-Mid and Superior.

To start, a concordance analysis is seen below. It cannot be used to judge the correctness of measuring or rating techniques; rather, it shows the degree to which different measuring or rating techniques agree with each other.

Note that category names were shortened to fit into the tables below. They follow the following abbreviations:

NL="Novice Low", NM="Novice Mid", NH="Novice High, IL="Intermediate Low", IM="Intermediate Mid", IH="Intermediate High", AL="Advanced Low", AM="Advanced Mid", AH="Advanced High", S="Superior"

Table 1 Concordance Table for English WPT® from 2012 to 2014

		Rater 1									
		NL	NM	NH	IL	IM	IH	AL	AM	AH	S
Rater 2	NL	2	0	0	0	0	0	0	0	0	0
	NM	0	5	1	0	0	0	0	0	0	0
	NH	0	0	8	1	0	0	0	0	0	0
	IL	0	0	1	14	0	0	0	0	0	0
	IM	0	0	0	5	88	6	1	0	0	0
	IH	0	0	0	0	8	256	51	9	2	0
	AL	0	0	0	0	3	97	255	88	12	1
	AM	0	0	0	0	0	10	79	348	82	4
	AH	0	0	0	0	0	0	13	102	506	59
	S	0	0	0	0	0	0	1	8	80	931

Table 2 Concordance Table for Russian WPT® from 2012 to 2014

		Rater 1									
		NL	NM	NH	IL	IM	IH	AL	AM	AH	S
Rater 2	NL	6	1	1	0	0	0	0	0	0	0
	NM	0	21	2	1	0	0	0	0	0	0
	NH	0	5	30	5	2	0	0	0	0	0
	IL	0	0	11	35	14	0	0	0	0	0
	IM	0	0	0	11	57	10	1	0	0	0
	IH	0	0	0	1	6	80	15	0	0	0
	AL	0	0	0	0	0	17	45	9	0	0
	AM	0	0	0	0	0	4	7	50	6	0
	AH	0	0	0	0	0	0	0	6	23	4
	S	0	0	0	0	0	0	0	0	0	97

Table 3 Concordance Table for Spanish WPT® from 2012 to 2014

		Rater 1									
		NL	NM	NH	IL	IM	IH	AL	AM	AH	S
Rater 2	NL	6	1	0	0	0	0	0	0	0	0
	NM	0	15	3	0	0	0	0	0	0	0
	NH	0	0	53	9	1	0	0	0	0	0
	IL	0	0	10	166	41	1	0	0	0	0
	IM	0	0	5	40	772	106	4	0	0	0
	IH	0	0	0	7	106	1426	271	18	0	0
	AL	0	0	0	0	6	272	1125	234	10	0
	AM	0	0	0	0	0	19	173	839	76	7
	AH	0	0	0	0	0	1	6	83	257	48
	S	0	0	0	0	0	0	0	6	28	80

The concordance tables illustrate generally good agreement between the raters as there are no ratings that are strikingly different than one another as seen by the large quantity of 0s seen in the upper right and bottom left of the rater matrix.

Internal consistency reliability

There are two types of inter-rater reliability evidence for rater-based assessments: inter-rater reliability coefficients and inter-rater agreement (concordance of ratings). Although there are many types of reliability analyses, the choice of a specific technique should be governed by the nature and purpose of the assessment and its data.

Spearman's rank-order correlation (R) is a commonly used correlation for assessing inter-rater reliabilities, and correlations should be at or above .70 to be considered sufficient for test development and .80 for operational use (e.g., LeBreton et al., 2003). Spearman's R is the most appropriate statistic for evaluation of the ACTFL WPT® data because the proficiency categories used for ACTFL WPT® ratings are ordinal in nature.

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

Table 4 Spearman's Correlations by Language from 2012-2014

Language	N	ρ	95% CI LL	95% CI UL	p
English	3137	0.936	0.929	0.942	<0.001
Russian	585	0.973	0.966	0.980	<0.001
Spanish	6337	0.913	0.907	0.919	<0.001

Table 5 Spearman's Correlations by Year

Language	Year	N	ρ
English	2012	1018	0.91
	2013	1008	0.94
	2014	1111	0.96
Russian	2012	121	0.99
	2013	191	0.98
	2014	273	0.96
Spanish	2012	1740	0.91
	2013	2232	0.93
	2014	2365	0.90

Overall, the ACTFL WPT® exceeded inter-rater reliability minimum standards and was quite high. The Spearman's R correlation was .936 for English, .973 for Russian, and .913 for Spanish. Inter-rater reliability was high across language categories and interview year. These results are consistent with previous years' results (Surface and Dierdorff, 2004; Bärenfänger and Tschirner, 2011; SWA Consulting, 2012) providing evidence of acceptable inter-rater agreement for operational use over time.

Evidence for equivalence of forms of the test

Before beginning the WPT®, test takers receive clear instructions for taking the test. These instructions are delivered in English. They then complete a Background Survey which elicits information about the test taker's work, school, home, personal activities, and interests. The survey answers determine the pool of prompts from which the computer will randomly select topics for writing tasks. The variety of topics, the types of questions, and the range of possible computer-generated combinations allows for individually designed assessments.

The Self-Assessment provides six different descriptions of how well a person can write in a language. Test takers select the description that they feel most accurately describes their writing ability in the target language. The Self-Assessment choice determines which one of three WPT® test forms is generated for the specific individual. The choices made by the test taker in response to the Background Survey and the Self-Assessment ensure that each test taker receives a customized and unique test.

The WPT® directions at the beginning of the assessment, provide instructions on how to navigate the test. To ensure that the WPT® test taker can make the necessary diacritical marks in the target language which are not represented on a standard U.S. keyboard, several keyboard options are available within the test software. Institutions can determine in advance which keyboard options should be made available to their test takers. At the time of the test, the test taker will make a choice based on the options set forth by the client/institution. To ensure that the test taker understands these options, a warm-up task is provided before the start of the test to allow the candidates to become familiar with the key-board options available. Once the warm-up is completed and the actual test is started, the test taker cannot change the selected keyboard. The WPT® is also available in traditional paper and pencil format and with the same customization and adaptive features as the online version.

Scorer reliability for essay items

Another common approach to examining reliability is to use measures of inter-rater agreement. Whereas inter-rater reliability assesses how consistently the raters rank-order test-takers, inter-rater agreement assesses the extent to which raters give the same score for a particular test-taker. Since the rating protocol assigns final test scores based on agreement (concordance) between raters rather than rank-order consistency, it is important to assess the degree of interchangeability in ratings for the same test taker. Inter-rater reliability can be high when inter-rater agreement is low, so it is important to take both into account when assessing a test.

Inter-rater agreement can be assessed by computing absolute agreement between rater pairs (i.e., whether both raters provide exactly the same rating). Standards for absolute agreement vary depending on the number of raters involved in the rating process. When two raters are utilized, there should be absolute agreement between raters more than 80% of the time, with a minimum of 70% for operational use (Feldt & Brennan, 1989). Absolute agreement closer to 100% is desired, but difficult to attain. Each additional rater employed in the process decreases the minimum acceptable agreement percentage. This accounts for the fact that agreement between more than two raters is increasingly difficult. Adjacent agreement is also assessed in this reliability study. Adjacent agreement occurs when raters are within one rating level in terms of their agreement (e.g., rater one gives a test taker a rating of Intermediate Mid and rater two gives a rating of Intermediate Low). In the ACTFL process, when there is not absolute agreement, an arbitrating third rater will provide a rating that resolves the discrepancy.

Table 6 Absolute/Adjacent Agreement by Language from 2012-2014

Language	N	Absolute Agreement (exact)	Adjacent Agreement (+/- 1)	None (+/- 2)
English	3137	77%	21%	2%
Russian	585	75%	22%	3%
Spanish	6337	75%	24%	1%

Table 7 Absolute/Adjacent Agreement by Language and Year

Language	Year	N	Absolute Agreement (exact)	Adjacent Agreement (+/- 1)	None (+/- 2)
English	2012	1018	71%	26%	3%
	2013	1008	74%	24%	2%
	2014	1111	85%	13%	2%
Russian	2012	121	86%	12%	2%
	2013	191	84%	15%	1%
	2014	273	66%	32%	2%
Spanish	2012	1740	76%	22%	2%
	2013	2232	77%	22%	1%
	2014	2365	72%	26%	2%

Table 8 Absolute/Adjacent Agreement by Language and Sublevel Proficiency from 2012-2014

Language	Rating	N	Absolute Agreement (exact)	Adjacent Agreement (+/- 1)	None (+/- 2)
English	Novice Low	2	100%	0%	0%
	Novice Mid	5	100%	0%	0%
	Novice High	10	80%	20%	0%
	Intermediate Low	20	70%	30%	0%
	Intermediate Mid	99	89%	8%	3%
	Intermediate High	369	69%	28%	3%
	Advanced Low	400	64%	33%	4%
	Advanced Mid	555	63%	34%	3%
	Advanced High	682	74%	24%	2%
	Superior	995	94%	6%	1%
Russian	Novice Low	6	100%	0%	0%
	Novice Mid	27	78%	22%	0%
	Novice High	44	68%	30%	2%
	Intermediate Low	53	66%	30%	4%
	Intermediate Mid	79	72%	25%	3%
	Intermediate High	111	72%	24%	4%
	Advanced Low	68	66%	32%	1%
	Advanced Mid	65	77%	23%	0%
	Advanced High	29	79%	21%	0%
	Superior	101	96%	4%	0%
Spanish	Novice Low	6	100%	0%	0%
	Novice Mid	16	94%	6%	0%
	Novice High	71	75%	18%	7%
	Intermediate Low	222	75%	22%	3%
	Intermediate Mid	926	83%	16%	1%
	Intermediate High	1825	78%	21%	1%
	Advanced Low	1579	71%	28%	1%
	Advanced Mid	1180	71%	27%	2%
	Advanced High	371	69%	28%	3%
	Superior	135	59%	36%	5%

Absolute agreement was higher than 70% for all high level comparisons within a major level. Absolute agreement and adjacent agreement all summed to at least 95% excluding the Novice High category in the Spanish WPT® exam. Absolute agreement was similar across language and language category. There was a slight improvement in inter-rater agreement from 2012 to 2014 excluding Russian which declined slightly in 2014. Comparisons made at the Language by Sublevel Proficiency should be viewed with caution as sample sizes can be limited and thus they should be used as a tool to identify possible areas for improvement in rater training.

Overall, the findings support the reliability of the ACTFL WPT® as an assessment of writing proficiency. Although the research is based on a very limited sample, there are indications that the NH/IL border is an area for continued improvement in interrater reliability. This is especially true for Russian (68% and 66%

absolute agreement at Novice High and Intermediate Low, respectively). This however has less of an impact on ACE Credit recommendations as the number of credits recommended by ACE for the ratings of Novice High and Intermediate Low is the same. Current ACE credit recommendations for ACTFL WPT ratings are listed in the chart below:

Official ACTFL WPT Rating	ACE Credit Recommendation
AH/S	6 (LD) + 8 UD)
AM	6 (LD) + 3 (UD)
IH/AL	6 (LD) + 1(UD)
IM	6 (LD)
NH/IL	3 (LD)

Errors of classification percentage for the minimum score for granting college credit (cut score)

The minimum score for granting college credit for an ACTFL OPI rating is Novice High. ACE determines the number of credits to be conferred based on the recommendations of expert reviewers, foreign language faculty who are familiar with language proficiency and the skills that students are expected to attain after various sequences of college language study.

Validity Information

Content-related validity

Content validity addresses the alignment between the test prompts and the content area they are intended to assess. There are two types of content-related validity: face validity and curricular validity. Face validity refers to the extent to which a test or the questions on a test *appear* to measure a particular construct. While curricular validity is the extent to which the content of the test matches the objectives of a specific curriculum. Both types of validity are evaluated by groups of content experts. The content validity evidence for the WPT® is represented by the degree to which the content of the test relates to the construct of writing proficiency as defined by the *ACTFL Proficiency Guidelines 2012 – Writing*.

Criterion-related validity

Similar to content-related validity, criterion-related validity also has two types. One type of criterion-related validity is predictive validity which refers to the power or usefulness of test scores to predict future performance. Concurrent validity, the other type of criterion-validity, focuses on the power of the test to *predict* outcomes on another test with similar content-related validity.

The ACTFL WPT® is a standardized procedure for the global assessment of functional language ability. Interactive and adaptive to the experiences, interests, and linguistic competence of the candidate, the WPT® measures written language production holistically by determining patterns of strengths and

weaknesses. Furthermore, it identifies a candidate's level and range of functional ability. The WPT® is a criterion-referenced testing method that measures how well a person functions in a language by comparing the individual's performance on specific language tasks with the criteria for each of the 10 levels described in the *ACTFL Proficiency Guidelines 2012 -Writing.*

Construct validity (if appropriate)

Construct validity refers to the degree to which a test or other measure assesses the underlying theoretical construct it is supposed to measure. Within construct validity there are two types: convergent validity and discriminant validity. Convergent validity consists of providing evidence that two tests are believed to measure closely related skills and addresses the reciprocity/correlation between measures that share the same content-related validity. Conversely, discriminant validity consists of evidence that two tests do not measure closely related skills.

Surface and Dierdorff (2004) studied the validity and reliability of the WPT® and found that the relationship between the OPI and WPT® scores was robust suggesting that both OPI and WPT® are assessing related and overlapping constructs. While this is a positive finding, it is an expected one as both are measures of language skill in the same language using the same assessment method.

Possible test bias of the total test score

Bias exists when a test makes systematic errors in measure or prediction (Murphy & Davidshofer, 2005, p.317). An example of this would occur when a test yields higher or lower scores on average when it is administered to specific criterion groups such as people of a particular race or sex than when administered to an average population sample. Negative bias is said to occur when the criterion group scores lower than average and positive bias when they score higher.

Bias is typically identified at the item level. Since this test's content is routed based on the ability and interests of the test taker, no two interviews are the same and thus a test of item bias would not be appropriate A bias analysis of total test score may be appropriate; however, demographic information is not tracked, therefore, this is not possible.

Evidence that time limits are appropriate and that the exam is not unduly speeded

The Writing Proficiency Test is proctored and begins with an Introduction, Background survey, Self-Assessment, Key-board selection and Warm-up, for which the candidate is given 10 minutes. Then the candidate begins the actual assessment, consisting of four requests for a variety of writing tasks. The candidate is given 80 minutes to complete the four writing tasks. Based on the Self-Assessment, the assessment will focus on only two levels of proficiency. For each of the four tasks, the candidate is given instructions on the recommended length and organization of the response (i.e., 2-3 paragraphs) as well as a recommendation for how long they should spend writing their response to assist them in finishing the test with enough time to re-read responses. Test-takers typically complete the test in 40-70 minutes depending on their level of writing proficiency.

Provisions for standardizing administration of the examination

The WPT® format guides the candidates through the test in the same standardized fashion.

I. Introduction and Directions

Introduction

This section contains an overview of the assessment directions, key-board selection and a warm-up activity to test the keyboard. Directions and demo tests are also made available in advance of the scheduled testing time. All directions are written in English. Special accommodations may be requested when directions and prompts need to be provided in a language other than English. Approximately ten (10) minutes are allotted for this introductory section of the test.

Background Survey: Selecting Topics for Writing

The Background Survey is a questionnaire which elicits information about the test taker's work, school, home, personal activities, and interests. The survey answers determine the pool of prompts from which the computer will randomly select topics for writing tasks. The variety of topics, the types of questions, and the range of possible computer-generated combinations allows for individually designed assessments. Even if two test takers select the same combination of Background Survey responses, the resulting tests will be different.

Self-Assessment: Defining the Level of the WPT®

The Self-Assessment provides six different descriptions of how well a person can write in a language. Test takers select the description that they feel most accurately describes their writing ability in the target language. The Self-Assessment choice determines which one of three WPT® test forms is generated for the specific individual (Novice/Intermediate, Intermediate/Advanced or Advanced/Superior). The choices made by the test taker in response to the Background Survey and the Self-Assessment ensure that each test taker receives a customized and unique test.

WPT® Test Administration & Keyboard Options

The WPT® provides directions on how to navigate within a page and from page to page. To ensure that the WPT® test taker can make the necessary diacritical marks in the target language which are not represented on a standard U.S. keyboard, several keyboard options are available within the test software. Institutions can determine in advance which keyboard options should be made available to their test takers. At the time of the test, the test taker will make a choice based on the options set forth by the client/institution. Once the test-taker selects a keyboard option, they get to try it out on the Warm-Up task to practice using it and can change their keyboard option at the end of the Warm-up and try an alternative key-board. Once the actual test has started, the candidate can no longer change the keyboard options.

II. Test

Writing Prompts

There are four separate prompts, each of which encompasses multiple writing tasks, (i.e., descriptive, informative, narrative, and persuasive writing) presented in English. Each request describes the audience, context, and purpose of the prompt. The four prompts that are presented to the writer are designed to elicit writing at the Intermediate, Advanced, and Superior levels, across a variety of contexts and content areas. Most prompts will target more than one task associated with one or more levels within the same context.

Each request also describes the suggested length of the response (i.e., several sentences, multiple paragraphs, etc.) and suggests a time allotment (i.e., 10 minutes, 25 minutes, etc.) for completing the response to that specific request. The total time allotted for all four requests is 80 minutes.

III. Rating the Texts Produced

Certified WPT® Raters rate the texts produced holistically, meaning that linguistic components are viewed from the wider perspective of how successfully they contribute to the overall texts produced. The rating criteria considered are: the tasks or functions the test taker produced, the range of social contexts and specific content areas they could handle, the accuracy of the written language, and the length and organization of the texts the test taker is capable of producing.

Provisions for exam security

Official WPT®s are administered in proctored environments. All proctors must read and review proctor instructions and sign an official proctor agreement before given access to any logins for assessments.

When the WPT® is administered to an academic institution, educational organization, or corporate clients, the following personnel qualify as potential proctor candidates:

K-12 Schools and School Districts

A proctor at a K-12 school or school district may only be a Principal, Assistant Principal, Dean, Administrative Assistant to the Principal or Dean, School District HR personnel, or Academic Chair. No other administrators or staff are permitted to act as proctors. All must submit a signed proctor agreement.

University or College

A proctor at a college may be a Professor, Department Chair, Department Administrative Assistant or Department Coordinator. No other administrators or staff are permitted to act as proctors. All must submit a signed proctor agreement.

Corporate clients

A proctor at a corporate site must be a managerial-level Human Resource staff member, or executive staff member, or, for branch offices without an on-site human resource representative, a senior level manager may act as proctor. All must submit a signed proctor agreement.

Security Measures

Each test candidate is required to fill out a Background Survey before the start of the WPT. Responses to the survey trigger the random selection of four requests for writing (from a test request pool of over 1800 requests). Additionally, tasks are retired in correlation to the frequency with which the task has been administered.

All official WPT®'s are proctored to ensure that candidates do not copy the prompts they receive or use pre-written responses. Logins for assessments are only valid for use for two weeks and once a candidate has logged into an assessment, they must complete that assessment in one sitting within two hours. If a

test candidate tries to access another website while logged into the assessment, the WPT® will close and only a proctor can log the candidate back in.

Raters also read for suspicious behavior: a significant change in writing ability from one task to another, patterned errors suddenly disappearing, change in hand writing. Raters are instructed to assign the score of UR for unratable and notify LTI test administration of "suspicious behavior" which is then investigated by the Director of Test Administration.

Scaling and Item Response Theory Procedures

Types of IRT scaling model(s) used

Item Response Theory (IRT) models are not used in the calibration or scoring model for this exam. Test takers are scored based on meeting criteria fitting the description of a major level which is representative of a specific range of abilities. Written descriptions of language abilities that a test taker must perform can be found in *ACTFL Proficiency Guidelines 2012 – Writing*.

Evidence of the fit of the model(s) used

The primary goal of the WPT® is to produce a ratable sample of writing. The Self-Assessment provides six different descriptions of how well a person can write in a language. Test takers select the description that they feel most accurately describes their writing ability in the target language. The Self-Assessment choice determines which one of three WPT® test forms is generated for the specific individual: Novice/Intermediate, Intermediate/Advanced or Advanced/Superior. The choices made by the test taker in response to the Background Survey and the Self-Assessment ensure that each test taker receives a customized and unique test. Writing requests will target more than one task associated with one or more contiguous levels within the same context/content.

Evidence that new items/tests fit the current scale used

The *ACTFL Proficiency Guidelines* and the Assessment Criteria for Writing describe the range of content and contexts a speaker at each major level should be able to handle. For example, at the Intermediate level, topics of personal interest and related to one's immediate environment are selected; at the Advanced level, topics move beyond the autobiographical to topics of general community, national, and international interest; at the Superior level, topics are presented as issues to be discussed from abstract and/or hypothetical perspectives.

Validity of Computer Administration

Size of the operational test item pool for test

Each test candidate is required to fill out a Background Survey before the start of the WPT®. Responses to the survey trigger the random selection of prompts from a test prompt pool of over 1829 prompts. Prompts are rotated on a regular basis; new prompts are created and implemented while existing prompts are disabled.

Exposure rate of items when examinees can retake the test

The somewhat adaptive nature of the WPT® allows for some level of exposure control. There are1829 prompts available per language and records of retests are maintained to ensure that candidates receive alternative tasks. Additionally, ACTFL controls for testing effects by limiting future retests to be 90 days from the most recent testing event.

Cut-score information

Rationale for the particular cut-score recommended

Once a ratable sample of writing has been provided by the test taker, that sample is compared to the assessment criteria of the rating scale. A rating at any major level is determined by identifying the writer's floor and ceiling. The floor represents the writer's highest sustained performance across ALL of the criteria of the level all of the time for that particular level; the ceiling is evidenced by linguistic breakdown when the writer is attempting to address the tasks presented that exceed the writer's ability to control. An appropriate sublevel can then be determined, and one of ten possible ratings is assigned by comparing the sample to the descriptions in the *ACTFL Proficiency Guidelines 2012 – Writing* and assigning the rating that best matches the sample.

Evidence for the reasonableness and appropriateness of the cut-score recommended

The *ACTFL Proficiency Guidelines* are descriptions of what individuals can do with language in terms of speaking, writing, listening, and reading in real-world situations in a spontaneous and non-rehearsed context. For each skill, these guidelines identify five major levels of proficiency: Distinguished, Superior, Advanced, Intermediate, and Novice. The major levels of Advanced, Intermediate, and Novice are subdivided into High, Mid, and Low sublevels. The levels of the *ACTFL Proficiency Guidelines* describe the continuum of proficiency from that of the highly articulate, well-educated language user to a level of little or no functional ability.

These Guidelines present the levels of proficiency as ranges, and describe what an individual can and cannot do with language at each level, regardless of where, when, or how the language was acquired. Together these levels form a hierarchy in which each level subsumes all lower levels. The Guidelines are not based on any particular theory, pedagogical method, or educational curriculum. They neither describe how an individual learns a language nor prescribe how an individual should learn a language, and they should not be used for such purposes. They are an instrument for the evaluation of functional language ability.

The *ACTFL Proficiency Guidelines* were first published in 1986 as an adaptation for the academic community of the U.S. Government's Interagency Language Roundtable (ILR) Skill Level Descriptions.

The third edition of the *ACTFL Proficiency Guidelines* includes the first revisions of Listening and Reading since their original publication in 1986, and a second revision of the ACTFL Speaking and Writing Guidelines, which were revised to reflect real-world assessment needs in 1999 and 2001 respectively. New for the 2012 edition are: the addition of the major level of Distinguished to the Speaking and Writing Guidelines; the division of the Advanced level into the three sublevels of High, Mid, and Low for the Listening and Reading Guidelines, and; the addition of a general level description at the Advanced, Intermediate, and Novice levels for all skills.

Another new feature of the 2012 Guidelines is their publication online, supported with glossed terminology and annotated, multimedia samples of performance at each level for Speaking and Writing, and examples of oral and written texts and tasks associated with each level for Reading and Listening.

The direct application of the *ACTFL Proficiency Guidelines* is for the evaluation of functional language ability. The Guidelines are intended to be used for global assessment in academic and workplace settings. However, the Guidelines do have instructional implications. The *ACTFL Proficiency Guidelines* underlie the development of the *ACTFL Performance Guidelines for K-12 Learners* (1998) and the *ACTFL Performance Descriptors for Language Learners (2012)* and are used in conjunction with the National Standards for Foreign Language Learning (1996, 1998, 2006, 2014) to describe how well students meet content standards. For the past 25 years, the *ACTFL Proficiency Guidelines* have had an increasingly profound impact on language teaching and learning in the United States.

Procedures recommended to users for establishing their own cut scores (e.g. granting college credit)

The summary of the Official ACTFL credit recommendations can be found on the Language Testing International (LTI) website, the ACTFL testing office. Depending on the rating level achieved, ACE recommends anywhere from three lower division baccalaureate/ associate degree category credits for the achievement of Novice High/Intermediate Low, up to six lower division baccalaureate /associate degree category credits and eight upper division baccalaureate / associate degree category credits for the achievement of Advanced high/Superior language proficiency.

Bibliography

ACTFL (2012). ACTFL Proficiency Guidelines 2012. Retrieved October 1, 2015 (http://www.actfl.org/publications/guidelines-andmanuals/actfl-proficiency-guidelines-2012)

American Council on the Teaching of Foreign Languages. (2015). Performance descriptors for language learners. Available online at: http://www.actfl.org/sites/default/files/pdfs/ACTFLPerformance_Descriptors.pdf

Breiner-Sanders, K.E., Lowe, Jr., P., Miles, J., Swender, E. (2000). ACTFL proficiency guidelines – Writing revised 1999. *Foreign Language Annals*. 33(1). 13-18.

Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). Council of Europe, Language Policy Unit, Strasbourg (2001) http://www.coe.int/t/dg4/linguistic/source/framework_en.pdf

Dandonoli, P., Henning, G. (1990). An investigation of the construct validity of the ACTFL proficiency guidelines and oral interview procedure. *Foreign Language Annals. 23(1). 11-19.*

LeBreton, J.M., Burgess, J.R.D., Kaiser, R.B., Atchley, E.K., James, L.R. (2003). The restricution of variance hypothesis and inter-rater reliability and agreement: Are ratings from multiple sources really dissimilar? *Organizational Research Methods*, 6(1), 80-128.

Murphy, K.R., Davidshofer, C.O. (2005). *Psychological testing: Principles and Applications.* New Jersey, USA: Pearson Prentice Hall.

Surface, E.A., Dierdorff, E.C. (2004). Preliminary reliability and validity findings for the ACTFL writing proficiency test. *SWA Technical Report 2004-C04-R01*. Available online at: http://www.languagetesting.com/wp-content/uploads/2013/08/ACTFL-WPT®-Technical-Report-2004.pdf

SWA Consulting. (2012). Reliability study of the ACTFL WPT® ® in Arabic, Russian, and Spanish for the ACE review. *Technical Report.* Available online at: http://www.languagetesting.com/wp-content/uploads/2013/08/ACTFL-WPT®-Reliability-2012.pdf

Swender, E., Conrad, D.J, Vicars, R. (2012). ACTFL proficiency guidelines 2012. Available online at: http://www.actfl.org/sites/default/files/pdfs/public/ACTFLProficiencyGuidelines2012_FINAL.pdf

Tschirner, E., Bärenfänger, O. (2011). An extension of inquiry into reliability issues of the ACTFL writing proficiency test (WPT®). *Technical Report 2011-US-PUB-1.* Availabile online at http://www.languagetesting.com/wp-content/uploads/2013/08/ACTFL-WPT®-Technical-Report-2011.pdf