



ACTFL WRITING PROFICIENCY TEST

Part A: General Test Information

Part B: Statistical Analysis &
Evidence of Validity

Publication No. AAR/WPT/ACE/I-2020-001



ACTFL WRITING PROFICIENCY TEST

Part A: General Test
Information

Nina Lin, Ph.D.
Stanford Language Center,
Stanford University

Table of Contents

Rationale and purpose of the WPT	1
Name(s) and institutional affiliations of the principle author(s) or consultant(s).....	2
Types of scores reported for examinees	3
Scoring (rating) procedures	3
Cut scores.....	5
Procedures recommended to users for establishing their own cut scores.....	5
Equivalence of forms	6
Information on norms and normative groups (if appropriate)	6
Item/Test content development	6
<i>Specifications that define the domain(s) of content, skills, and abilities that the test elicits</i>	<i>6</i>
<i>Statement of test's emphasis on each of the content, skills, and ability areas</i>	<i>7</i>
<i>Rationale for the kinds of tasks (items) that make up the test.....</i>	<i>7</i>
<i>Information about the adequacy of the items on the test as a sample from the domain(s)</i>	<i>7</i>
<i>Information on the currency and representativeness of the test's items</i>	<i>8</i>
<i>Description of the item sensitivity panel review</i>	<i>8</i>
<i>Whether and/or how the items pre-tested (field tested) before inclusion in the final form</i>	<i>8</i>
Item analysis results (e.g. item difficulty, discrimination, item fit statistics, correlation with external criteria).....	8
References	9

Table of Figures

Figure 1: WPT assessment criteria chart.....	4
Figure 2: Time as a critical component for developing language performance.....	6

Rationale and purpose of the WPT

The American Council on the Teaching of Foreign Languages (ACTFL) Writing Proficiency Test (WPT) is a standardized test which assesses a test taker's ability to use written language effectively and appropriately in real-life contexts. The WPT measures how well a person spontaneously writes in a target language by comparing their performance in four to five specific writing samples to the criteria stated in the *ACTFL Proficiency Guidelines- 2012 - Writing*. WPT writing prompts deal with practical, social, and professional topics encountered in informal and formal contexts that represent the range of proficiency levels from Novice to Superior.

The test length for an ACTFL WPT is approximately 90 minutes with the first 10 minutes allotted for instructions. The actual writing test can take anywhere between 30 to 80 minutes depending on the proficiency range being assessed and the writing competence of the test taker. All instructions and prompts are given in English with the expectation that the test taker's responses be written or typed in the target language. The test taker is not allowed to use any ancillary materials (notes, print, or online resources) during the test.

The ACTFL WPT is primarily administered via the Internet, though a fixed form paper/pencil booklet is available in cases where Internet access is not available. Before taking the online version of the WPT, a test taker completes a Background Survey and a Self-Assessment. The Background Survey elicits information related to the test taker's work, school, home, and personal activities in order to identify appropriate content areas for an individualized assessment. In addition, the Self-Assessment invites test takers to select one of six descriptions they feel most accurately describes their writing ability. By utilizing information from the Background Survey and Self-Assessment, the computer generates an individually designed WPT tailored to both the test taker's experience and self-assessed proficiency level. Three possible forms may be generated:

- Form 1 targets Novice and Intermediate tasks and may be rated Novice Low to Intermediate Mid.
- Form 2 targets Intermediate and Advanced tasks and may be rated Novice Low to Advanced Mid.
- Form 3 targets Advanced and Superior tasks and may be rated Novice Low to Superior.

In all three forms of the online WPT, there are four to five separate multi-part prompts. Each prompt provides a clear explanation of the purpose of the writing task, the audience, and the context.

The paper/pencil booklet form of the WPT does not include a Background Survey or Self-Assessment. In this fixed form, test takers are first prompted to provide basic information at the Novice word-level, followed by four three-part prompts that require the writer to address tasks at the Intermediate, Advanced and Superior levels. The tasks increase in complexity throughout the test, ranging from simple informative writing to descriptive, narrative, and persuasive writing. This form may be rated from Novice Low to Superior. Because of the higher opportunity for exposure of test items, this form is replaced annually.

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

Principle Item writers for the ACTFL WPT included:

- 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

Subsequent item refreshes have taken place and include item writers and target language experts such as:

- Reuben Vyn, Ph.D. (French) University of Iowa
- Jane Shuffelton, Brighton High School (emerita), Rochester, NY
- Danila Nika, (Albanian)
- Addisu Hodes (Amharic)
- Razima Chowdhury, MA (Bengali)
- Larisa Zlatic, Ph.D. (Bosnian)
- Donna Kovacheva, MA (Bulgarian)
- Pui Shan Fiona Hui (Chinese-Cantonese)
- Sayad Farid Saydee, Ph.D. (Dari)
- Siddhi Talati (Gujarati)
- Marky Jean-Pierre (Haitian-Creole)
- Doron Friedman, Ed.D. (Hebrew)
- Bhabha Padmanabhan, MA (Malayalam)
- Mohammad Yunus Bazel (Pashto)
- Gerardina Malgorzata Szudelski, MA (Polish)
- Aileen Cole, MA (Swahili)
- Rose Bybee, MA (Tagalog)
- Emine Fougner (Turkish)
- Mara Sukholutskaya (Ukrainian)

- Tauseef Baig (Urdu)
- Kimloan Hill, Ph.D. (Vietnamese)
- Akinsola Ogundeji, MA (Yoruba)

Types of scores reported for examinees

Examinees' scores are reported as a 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 major level, as well as the content, context, discourse types, and accuracy associated with that 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. 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. There are no sublevels for Superior. The description of each major level is representative of a specific range of abilities. They also present the limitations that the writers encounter when attempting writing tasks at the next higher major level. An ACTFL WPT is assigned one of the following ratings: Superior, Advanced High, Advanced Mid, Advanced Low, Intermediate High, Intermediate Mid, Intermediate Low, Novice High, Novice Mid, or Novice Low.

Scoring (rating) procedures

Once the online WPT is completed, the sample is saved automatically on a secure Internet site. For WPTs completed using pen and paper, the booklet is scanned by the proctor and the PDF file is emailed to LTI. Once LTI confirms receipt of the file, the hard copies and digital files destroyed on site. An ACTFL Certified WPT Rater evaluates the entire sample holistically according to the Assessment Criteria described in the *Guidelines*.

The rater first determines the major level by evaluating whether the sample demonstrates sustained performance across ALL the criteria of a major level and whether there is evidence of breakdown from the criteria for the next higher level. Once the major level is decided, the sublevel is determined by the quality and quantity of the performance at that level and the proximity of performance to the next higher major level. 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 WPT is then blindly second rated by another certified WPT rater following the same protocol. Any rating

discrepancy is blindly arbitrated by a third rater, and an official ACTFL rating is assigned when two ratings agree exactly.

The assessment criteria for the major levels used to evaluate the ACTFL WPT is provided in the chart below:

Proficiency Level	Global Tasks and Functions	Context/Content	Text Type	Accuracy
Superior	Can write most kinds of correspondence (in-depth summaries, reports, and research papers). Can write in detail and explain complex matters, present and support opinions by developing cogent arguments, and compose hypotheses and conjectures.	Most informal and formal settings. <i>Practical, professional, and social topics treated both concretely and abstractly.</i>	Writes a clearly organized and articulated text that can extend from a series of paragraph to pages.	Demonstrates no patterned errors in basic structures, vocabulary, punctuation, and spelling. Some occasional errors may occur, which rarely disturb the reader.
Advanced	Can write routine, informal, and some formal correspondence, as well as narratives, descriptions in detail, and summaries of a factual nature. Can narrate and describe in all major time frames, at times uses paraphrasing and elaboration to provide clarity.	Informal settings and some routine formal settings on familiar topics. <i>Topics of personal and general interest.</i>	Writes a connected, cohesive text of at least a paragraph in length. Can extend to two or more paragraphs in length on familiar topics.	Demonstrates good control of the most frequently used structures and generic vocabulary, comprehensible to readers unaccustomed to the writing of language learners.
Intermediate	Can Create with language. Can meet practical writing needs, such as simple messages and letters, requests for information, and notes. Can ask and respond to simple questions.	Routine informal settings and limited tasks involving the exchange of simple information. <i>Familiar, predictable topics related to self and daily routine and activities.</i>	Writes a loosely connected text made up of a collection of primarily discrete sentences.	Expresses meaning through vocabulary and basic structures that is comprehensible to readers accustomed to the writing of language learners.
Novice	Can write words, lists and notes, and formulaic information to communicate the most basic information.	The most common informal settings. <i>Most common aspects of self and daily life.</i>	Writes words, lists, phrases, and limited formulaic information.	May be difficult to comprehend even for readers accustomed to dealing with the writing of language learners.

Figure 1: WPT assessment criteria chart

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. Prerequisites for becoming a Certified WPT Rater require Superior-level writing proficiency in the language of certification, minimum Advanced Mid level oral proficiency in English, and possession of an undergraduate degree in a related field.

Certified WPT Raters are expected to respect and follow all WPT rating protocols. WPT Raters also uphold the expectations related to confidentiality. As per their rater agreement, every rater agrees to abide by the rules and regulations regarding WPT rating, remaining calibrated to ACTFL proficiency standards and following all WPT procedures and guidelines. Under the supervision of the ACTFL Quality Assurance Program, WPT raters are authorized to rate WPTs and assign official ACTFL ratings exclusively through Language Testing International (LTI), the exclusive licensee of ACTFL assessments.

Cut scores

The WPT® does not have numeric cut scores. The WPT is an assessment of language proficiency that is rated holistically according to the *ACTFL Proficiency Guidelines* (2012).

Procedures recommended to users for establishing their own cut scores

As previously referenced, the ACTFL WPT is a proficiency-oriented assessment with no recommended cut scores. That is, the OPIc should result in a description of the test taker's spontaneous, unrehearsed language abilities. As such, the 2015 – 2019 ACE credit recommendations relate proficiency levels to credit recommendations.

ACTFL RATING	WPT
Novice High/Intermediate Low	3LD
Intermediate Mid	6LD
Intermediate High/Advanced Low	6LD + 1UD
Advanced Mid	6LD + 3UD
Advanced High/Superior	6LD + 6UD

For any language program, the proficiency levels can be mapped to course and program goals by analyzing the descriptors and comparing them to course and/or program objectives in addition to factors such as time.

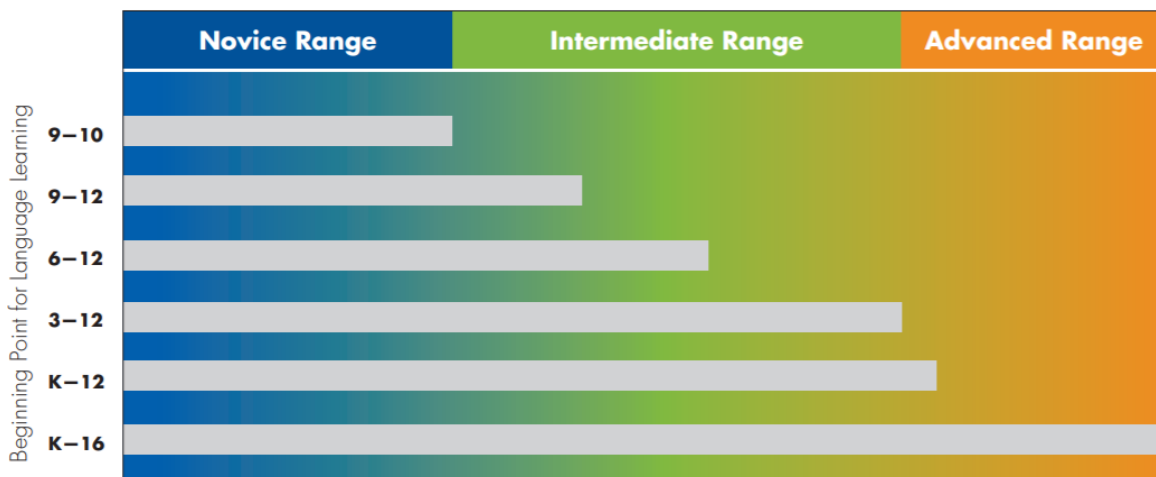


Figure 2: Time as a critical component for developing language performance

ACTFL suggests that the credit recommendations and proficiency targets above are in line with the number of courses and years of study that an undergraduate student of typical aptitude might achieve (see Figure 2).

Equivalence of forms

The WPT is made up of a pool of over 1,800 prompts; as referenced above, prompts are selected from the pool based on an algorithm which builds a test form based on the test taker's responses to the Background Survey and the Self- Assessment. Each examinee should receive a unique set of items in many instances.

ACTFL WPT prompts are function-based which are outlined in the *ACTFL Proficiency Guidelines*. This allows for a standardized approach to test development such that the content of a prompt along with tasks used to convey the functions differ from prompt to prompt and thus examinee to examinee; however, the functions for which test takers must demonstrate a sustained ability to communicate remain consistent. Prompt writer's adherence to the function-informed and rating scale-normed item writing protocol along with adherence to the process of awarding ratings according to the ACTFL Proficiency Descriptors allow for equivalence between forms.

Information on norms and normative groups (if appropriate)

The WPT® is a criterion-referenced test. No norm-referenced information is reported.

Item/Test content development

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

With regard to the content of the test, as referenced above, the ACTFL online WPT utilizes a Background Survey. The Background Survey is a questionnaire that elicits information about

the test taker's work, school, home, personal activities, and interests. The test taker's answers determine the pool of prompts from which the computer randomly selects topics for writing tasks. In order to determine the proficiency levels targeted in an online WPT, the test taker completes a Self-Assessment by comparing their own writing abilities in the target language to six descriptions of how well a person can write in a language. Based on the variety of topics and the self-assessed range of proficiency, a computer algorithm generates appropriate topics and linguistic levels for the test taker. The variety of topics, the types or questions, and the range of possible computer-generated combinations ensure that each test taker receives a customized and unique test

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 *ACTFL Proficiency Guidelines- 2012 – 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 requests for information (Intermediate level) to writing paragraph-length narrations and descriptions in all major time frames (Advanced) to dealing abstractly with current issues of general interest, supporting opinions and hypothesizing in extended 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*. Each form of the WPT consists of four to five separate prompts. Each prompt is presented in the form of a testlet that encompasses multiple (2-3) interrelated writing tasks that either focus on one major level or span two major levels. Furthermore, each WPT presents the opportunity for the test taker to write about a variety of content areas. As such, the test taker is given ample opportunities to demonstrate their patterns of linguistic strengths ("floor") and their limitations ("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. Topics generated at each major level follow the Guidelines. Additionally, as noted above, a Background Survey elicits information about the test taker's work, school, home, personal activities, and interest to determine the pool of prompts from which the computer randomly selects topics and generates testlets. The diversity of topics, variety of question types, and the range of

possible computer-generated combinations allows for individually customized assessments that present the test taker with the opportunity to demonstrate writing proficiency across a range of content and context areas.

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 for the items include home, education, free-time activities, sports, work, business, history, travel, language, the environment, entertainment, popular culture, technology, education, current events, etc. New topics and new items are always being developed and old ones revised as they become less current.

Description of the item sensitivity panel review

Since prompt selections of each online WPT is based on the Background Survey of the individual test taker, selection of items which may be sensitive or irrelevant for the test taker can be avoided. In order to ensure that test takers are not offended or made uncomfortable while taking a WPT, item writers are instructed to avoid sensitive topics when developing WPT writing prompts. They avoid topics such as immigration, national origin, sexual preference, religion, marital status, racism, etc.

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

Since each online 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 eliminated. Item performance is continually monitored over time for cases where topics later become sensitive (e.g. pandemics). In such cases, content is also 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 area as described in the *ACTFL Proficiency Guidelines 2012 – Writing*. Please refer to Alpine Testing Solutions (2020c) for a statistical analysis of the ACTFL Writing Proficiency Test® (WPT).

References

- ACTFL (2012). *ACTFL proficiency guidelines – writing*.
<https://www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012/english/writing>
- ACTFL (2020). *ACTFL writing proficiency familiarization guide*.
<https://www.actfl.org/sites/default/files/assessments/2020%20WPT%20Familiarization%20Guide.pdf>
- Alpine Testing Solutions. (2020c). *Examination of the ACTFL Writing Proficiency Test® (WPT) in English, Russian, and Spanish for the ACE Review - Part B: Statistical analysis & evidence of validity*. Orem, UT: Alpine Testing Solutions.
- ACTFL (no date). *Demo version of the ACTFL Writing Proficiency Test*.
<https://wptdemo.actfltesting.org/>
- Cubbellotti, S. (2015a). *Examination evaluation of the ACTFL WPT® in English, Russian, and Spanish for the ACE Review* (ACTFL Publication No. AAR/WPT/ACE/R–2015-001). Alexandria, VA: ACTFL.
- Cubbellotti, S. (2015b). *Examination evaluation of the ACTFL OPIc® in Arabic, English, and Spanish for the ACE Review* (ACTFL Publication No. AAR/OPIc/ACE/R–2015-001). Alexandria, VA: ACTFL.
- Cubbellotti, S. & Cox, T. (2015). *Examination evaluation of the ACTFL OPI® in French, Korean, Mandarin for the ACE Review* (ACTFL Publication No. AAR/OPI/ACE/R–2015-001) Alexandria, VA: ACTFL.
- Language Testing International (2019). *ACTFL WPT examinee handbook*.
<https://www.languagetesting.com/pub/media/wysiwyg/manuals/wpt-examinee-handbook.pdf>
- Language Testing International (2017). *ACTFL assessments catalog*.
<https://www.languagetesting.com/pub/media/wysiwyg/manuals/actfl-assessments-brochure.pdf>

Examination Evaluation of the ACTFL Writing Proficiency Test® (WPT) in Arabic, Russian, and Spanish for the ACE Review

Part B: Statistical Analysis & Evidence of Validity

Submitted to the American Council on the Teaching of Foreign Languages
(ACTFL)

Submitted May 29, 2020



Table of Contents

Executive Summary.....	3
Statistical Performance	4
Item Analysis Results (e.g., Item Difficulty, Discrimination, Correlation with External Criteria)	4
Reliability Information, Scorer Reliability for Essay Items, Errors of Classification When Single or Multiple Cut-Scores are Used	4
Score Stability Over Time.....	10
Evidence of Validity.....	12
Content Related	12
Criterion Related	12
Construct Related	13
Possible Test Bias	13
Evidence that Time Limits are Appropriate and that the Exam is not Unduly Speeded ..	13
Provisions for Standardizing Administration of the Examination.....	14
Irrelevant Sources of Difficulty Affecting Test Scores.....	14
Provisions for Exam Security.....	14
Interpretations and Conclusions.....	16
References	17



Executive Summary

This document is structured to parallel the ACE Examination Checklist, which addresses the following topics: statistical performance and validity evidence.

This report documents the American Council on the Teaching of Foreign Languages (ACTFL) Writing Proficiency Test (WPT®) from 2016 to 2020 to satisfy a review requirement of the American Council of Education (ACE) College Credit Recommendation Service (CREDIT) program. The ACTFL WPT® is an assessment of functional writing proficiency in a foreign language that is evaluated by trained and certified experts in a writing format across numerous languages.

Inter-rater reliability and rater agreement were analyzed for three languages of the ACTFL WPT: Arabic, Russian, and Spanish. Additionally, comparisons were analyzed across language proficiency levels, as well as for testing years (i.e. 2016-2020, in this sample).

Results show that the ACTFL WPT met the minimum inter-rater reliability and agreement requirements. Agreement between raters occurred in all examined languages within one sublevel of each other over 85% of the time, and within two sublevels 99% of the time. Additionally, the findings of the Spearman's R Correlation analyses demonstrate that the correlations of the ratings are almost always positive and strong, ranging from 0.75-0.90 across languages. Areas for improvement include a focus on the absolute agreement between raters within the Intermediate High – Advanced Low and Advanced High – Superior borders across languages. These findings are expanded upon and discussed in detail below.

Please refer to Part A for general test information.

Statistical Performance

Item Analysis Results (e.g., Item Difficulty, Discrimination, Correlation with External Criteria)

Examinees are scored at the “highest level of sustained functional ability,” which means a single holistic proficiency rating is assigned for the whole exam (see Examinee Handbook, page 20). Individual item (prompt) data is not collected.

Reliability Information, Scorer Reliability for Essay Items, Errors of Classification When Single or Multiple Cut-Scores are Used

An inter-rater agreement analysis was conducted for each language from 2016 to 2020. In this analysis, the number of times Rating 1 and Rating 2 agreed exactly, within one category (proficiency level), within two categories, or beyond two categories was counted. When two ratings did not agree, a third rating contributed to the score. If there was still disagreement, a fourth rating contributed to the decision. It is noteworthy that Ratings 1, 2, 3, and 4 does not mean a specific *Rater 1, 2, 3, and 4*. Instead, Rating 1 refers to the rating assigned by “Rater 1”, where Rater 1 was selected from a pool of trained raters. An individual assigned as “Rater 1” for one candidate may be Rater 2, 3, or 4 for another candidate. In other words, the rating number is not consistently connected to a specific individual.

The exam is initially scored by two raters (i.e., Rating 1 and Rating 2). If these two raters do not agree, a third rater is brought in for adjudication. If the third rater agrees with either of the first two raters, then the rating is finalized. However, if the third rater disagrees with both of the first two raters, a fourth rater is used. This process is followed for nearly all scores; however, there are cases in which scores are finalized after conversations with the involved raters.

Table 1 lists the number of examinees analyzed by year. Table 2 lists the percent of examinees that had exactly two, exactly three, or four ratings for their exam. Overall, the percentage of the number of ratings was fairly consistent when comparing Spanish and Russian. The Arabic exam had more exams with 3 or 4 raters and fewer with 2 raters compared to Spanish and Russian. However, the percentages of the number of ratings are all within 10% of each other.

Table 1. Number of Examinees by Year

	2016	2017	2018	2019	2020*	Total
Arabic	119	210	192	228	96	845
Spanish	2356	2480	3052	2720	739	11,347
Russian	191	119	131	129	21	591

*Arabic data collected through March 20, 2020; Spanish data collected through March 14, 2020; Russian data collected through March 11, 2020.

Table 2. Percent of Examinees with 2, 3, or 4 Ratings from 2016 to 2020

	N	2 Ratings	3 Ratings	4 Ratings
Arabic	845	42%	54%	4%
Spanish	11,347	52%	47%	<1%
Russian	591	53%	46%	<1%

Tables 3-5 list the agreement of Rating 1 and Rating 2 by category. Table 6 summarizes the percent of exact agreement, adjacent agreement (within one category), and agreement within two categories.

Table 3. Arabic WPT: Rating 1 and Rating 2 Agreement from 2016-2020 (N = 845)

		Rating 1*									
		NL	NM	NH	IL	IM	IH	AL	AM	AH	S
Rating 2*	NL	4	0	0	0	0	0	0	0	0	0
	NM	1	0	1	1	1	0	0	0	0	0
	NH	0	0	4	2	0	0	0	0	0	0
	IL	0	0	1	15	14	5	1	0	0	0
	IM	0	0	0	12	23	31	10	0	0	0
	IH	0	0	0	3	19	63	32	10	3	1
	AL	0	0	0	1	3	40	66	26	3	4
	AM	0	0	0	0	0	7	46	66	19	15
	AH	0	0	0	0	0	3	14	43	21	24
	S	0	0	0	0	1	3	7	27	51	98

*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 4. Spanish WPT: Rating 1 and Rating 2 Agreement from 2016-2020 (N = 11,347)

		Rating 1*									
		NL	NM	NH	IL	IM	IH	AL	AM	AH	S
Rating 2*	NL	15	9	1	0	0	0	0	0	0	0
	NM	6	29	40	1	1	0	0	0	0	0
	NH	2	12	113	62	21	3	0	0	0	0
	IL	0	2	65	209	184	23	1	0	0	0
	IM	0	0	16	162	886	403	38	0	0	0
	IH	0	0	0	7	549	1690	715	79	0	0
	AL	0	0	0	0	62	828	1649	493	27	1
	AM	1	0	0	0	3	90	810	1085	235	23
	AH	0	0	0	0	0	2	37	232	217	60
	S	0	0	0	0	0	0	1	34	72	41

*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 5. Russian WPT: Rating 1 and Rating 2 Agreement from 2016-2020 (N = 591)

		Rating 1*									
		NL	NM	NH	IL	IM	IH	AL	AM	AH	S
Rating 2*	NL	0	0	0	0	0	0	0	0	1	0
	NM	0	0	4	0	0	0	0	0	0	0
	NH	0	0	13	8	1	0	0	0	0	0
	IL	0	0	5	4	13	0	0	0	0	0
	IM	0	0	2	21	48	30	5	0	0	0
	IH	0	0	0	2	16	58	23	6	0	0
	AL	0	0	0	0	5	25	54	18	2	0
	AM	0	0	0	0	0	8	31	48	12	2
	AH	0	0	0	0	0	0	3	10	28	8
	S	0	0	0	0	0	0	0	2	12	63

*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

As shown in Table 6, Rating 1 and Rating 2 had exact agreement 43% of the time for the Arabic exam, 52% for the Spanish exam, and 54% for the Russian exam. All three were within one category of each other over 85% of the time. Tables 7-9 expand on these values by listing the percentage (and number) of exact agreements, adjacent agreements (within one category), and agreements within two categories, respectively.

Table 6. Agreement between Rating 1 and Rating 2

	N	Exact Agreement	Adjacent Agreement (within 1 category)	Agreement within 2 Categories
Arabic	845	42.6%	85.4%	97.0%
Spanish	11,347	52.3%	95.8%	99.9%
Russian	591	53.5%	93.4%	99.8%

Table 7. Percent (N) of Exact Agreement

Language	Rating	Rating		
		2	3	4
Arabic	1	42.6% (360)	32.1% (158)	29.4% (10)
	2	---	35.0% (172)	26.5% (9)
	3	---	---	11.8% (4)
Spanish	1	52.3% (5934)	38.3% (2071)	14.3% (3)
	2	---	48.8% (2639)	28.6% (6)
	3	---	---	42.1% (8)
Russian	1	53.5% (316)	40.0% (110)	0.0% (0)
	2	---	37.8% (104)	0.0% (0)
	3	---	---	50.0% (1)

Table 8. Percent (N) of Adjacent Agreement (within 1 Category)

Language	Rating	Rating		
		2	3	4
Arabic	1	85.4% (722)	80.9% (398)	73.5% (25)
	2	---	79.9% (393)	52.9% (18)
	3	---	---	73.5% (25)
Spanish	1	95.8% (10,871)	92.1% (4978)	66.7% (14)
	2	---	95.6% (5164)	66.7% (14)
	3	---	---	78.9% (15)
Russian	1	93.4% (552)	90.9% (250)	50.0% (1)
	2	---	87.3% (240)	50.0% (1)
	3	---	---	100.0% (2)

Table 9. Percent (N) of Agreement within 2 Categories

Language	Rating	Rating		
		2	3	4
Arabic	1	97.0% (820)	97.4% (479)	91.2% (31)
	2	---	96.3% (474)	82.4% (28)
	3	---	---	94.1% (32)
Spanish	1	99.9% (11,334)	99.8% (5391)	95.2% (20)
	2	---	99.9% (5398)	95.2% (20)
	3	---	---	94.7% (18)
Russian	1	99.8% (590)	99.6% (274)	100.0% (2)
	2	---	99.6% (274)	50.0% (1)
	3	---	---	100.0% (2)

The Spearman rank-order correlation (ρ) was computed between each pair of Ratings. This correlation is a non-parametric measure of the strength and direction associated with the two variables of interest, in this case, two independent Ratings. The range of possible values is -1.00 to +1.00. This correlation is computed by first ranking the items for one variable (in this case, one of the Ratings) and then correlating it to the ranking of the items for the other variable (in this case, another Rating). A statistical significance test of the correlation determines whether the correlation is statistically significant, reported as a p -value.

The Spearman rank-order correlation is similar to a Pearson correlation, except the Pearson correlation involves interval level data while the Spearman rank-order correlation involves ordinal level data. Similar to the Pearson correlation, positive values would indicate a positive correlation between the two Ratings and negative values would indicate an inverse relationship

between the two Ratings. For this dataset, a positive correlation is expected, i.e., as the rating increases for one Rating, it is expected that the rating would also increase for the other Rating. The strength of the correlation is determined by the magnitude of the correlation. Correlations with absolute values of at least 0.70 generally indicate a strong correlation.

Table 10 displays the Spearman rank-order correlation results for each pair of Ratings. Ratings involving Rating 4 are not shown due to the small sample sizes. All correlations were positive, strong, and statistically significant.

Table 11 breaks down the correlations by year. Nearly all correlations were strong, positive, and statistically significant. The ratings that were moderately correlated with two values below 0.70 were for the Arabic exam. The correlation was 0.617 in 2020 between Ratings 2 and 3 and 0.693 between Ratings 1 and 3 in 2017. Ratings involving Rating 4 are not shown due to the small sample sizes.

Table 10. Spearman Rank-Order Correlations by Language from 2016-2020

Ratings Compared	Language	N	ρ	p -value
1 and 2	Arabic	845	0.814	< 0.001
1 and 2	Spanish	11,347	0.826	< 0.001
1 and 2	Russian	591	0.894	< 0.001
1 and 3	Arabic	492	0.750	< 0.001
1 and 3	Spanish	5403	0.755	< 0.001
1 and 3	Russian	275	0.828	< 0.001
2 and 3	Arabic	492	0.762	< 0.001
2 and 3	Spanish	5403	0.824	< 0.001
2 and 3	Russian	275	0.812	< 0.001

Table 11. Spearman's Correlations by Year

Language	Ratings	Year	N	ρ	p -value
Arabic	1 and 2	2016	119	0.784	< 0.001
	1 and 2	2017	210	0.810	< 0.001
	1 and 2	2018	192	0.810	< 0.001
	1 and 2	2019	228	0.804	< 0.001
	1 and 2	2020	96	0.741	< 0.001
Spanish	1 and 2	2016	2356	0.809	< 0.001
	1 and 2	2017	2480	0.817	< 0.001
	1 and 2	2018	3052	0.833	< 0.001
	1 and 2	2019	2720	0.828	< 0.001
	1 and 2	2020	739	0.829	< 0.001

Table 11. Spearman's Correlations by Year

Language	Ratings	Year	N	ρ	p -value
Russian	1 and 2	2016	191	0.925	< 0.001
	1 and 2	2017	119	0.891	< 0.001
	1 and 2	2018	131	0.868	< 0.001
	1 and 2	2019	129	0.853	< 0.001
	1 and 2	2020	21	0.947	< 0.001
Arabic	1 and 3	2016	72	0.721	< 0.001
	1 and 3	2017	108	0.693	< 0.001
	1 and 3	2018	110	0.765	< 0.001
	1 and 3	2019	147	0.718	< 0.001
	1 and 3	2020	55	0.721	< 0.001
Spanish	1 and 3	2016	1213	0.763	< 0.001
	1 and 3	2017	1220	0.752	< 0.001
	1 and 3	2018	1424	0.762	< 0.001
	1 and 3	2019	1244	0.739	< 0.001
	1 and 3	2020	302	0.761	< 0.001
Russian	1 and 3	2016	95	0.848	< 0.001
	1 and 3	2017	47	0.895	< 0.001
	1 and 3	2018	60	0.729	< 0.001
	1 and 3	2019	63	0.779	< 0.001
	1 and 3	2020	10	0.871	0.001
Arabic	2 and 3	2016	72	0.827	< 0.001
	2 and 3	2017	108	0.763	< 0.001
	2 and 3	2018	110	0.742	< 0.001
	2 and 3	2019	147	0.736	< 0.001
	2 and 3	2020	55	0.617	< 0.001
Spanish	2 and 3	2016	1213	0.803	< 0.001
	2 and 3	2017	1220	0.832	< 0.001
	2 and 3	2018	1424	0.838	< 0.001
	2 and 3	2019	1244	0.827	< 0.001
	2 and 3	2020	302	0.793	< 0.001
Russian	2 and 3	2016	95	0.790	< 0.001
	2 and 3	2017	47	0.888	< 0.001
	2 and 3	2018	60	0.797	< 0.001
	2 and 3	2019	63	0.747	< 0.001
	2 and 3	2020	10	0.987	< 0.001

Overall, the results of this analysis suggest that the Ratings are reasonably in agreement with each other and the correlations of the ratings are positive and strong. In the summary of the Rating 1 and Rating 2 correlations over time, Figure 1 shows that the correlations of the first two Ratings of the exams have a correlation above 0.853 for Russian, between 0.809 and 0.833

for Spanish and between 0.741 and 0.810 for Arabic. The correlations for the Russian exam were consistently the highest correlations over time among the three exams, and the correlations for the Arabic exam were consistently the lowest. The greatest disparity occurred in 2020 when the correlation was below 0.741 for Arabic but nearing 0.947 for the Russian exam. However, it is important to recognize that only 21 examinees completed the Russian exam in 2020 at the time of this report. It is possible that examinees in the early part of the year are not representative of the full year.

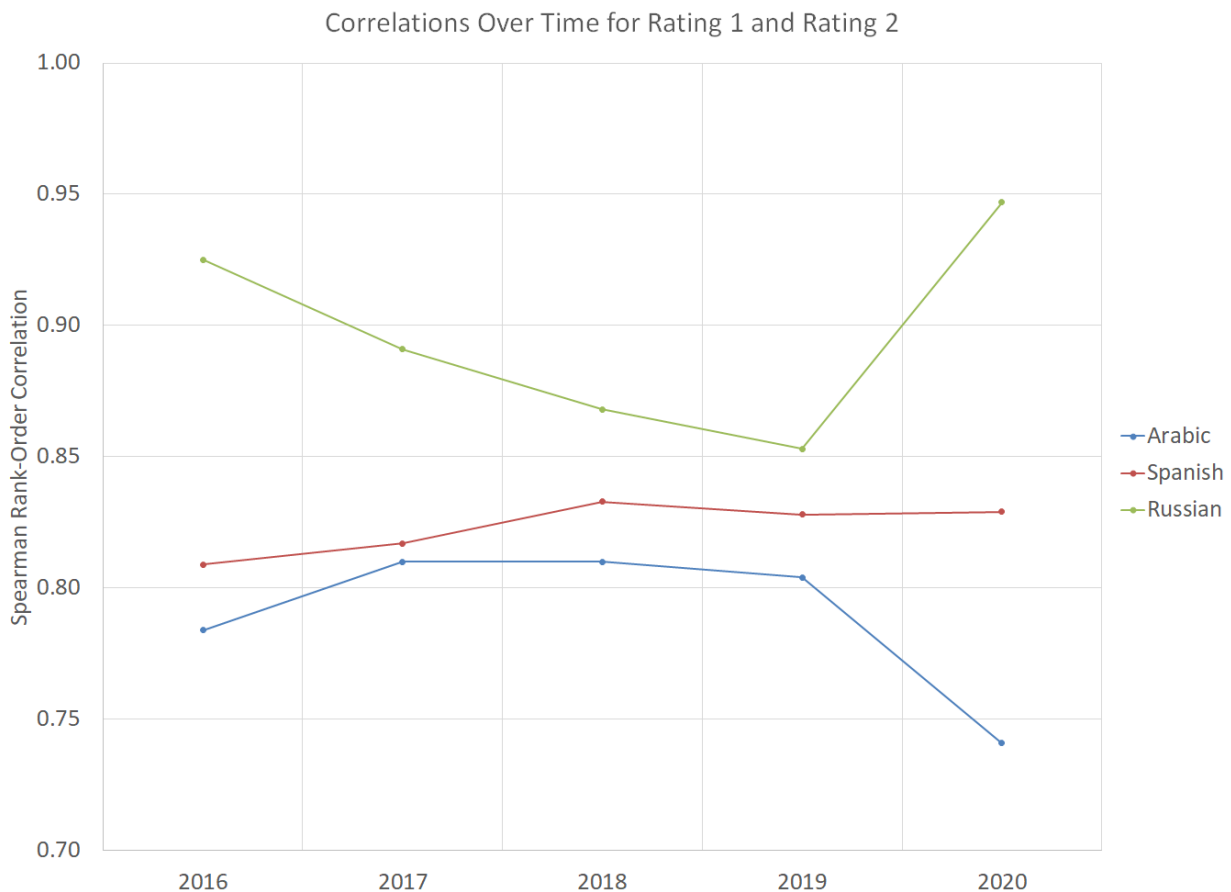


Figure 1. Spearman-rank correlations of Rating 1 and Rating 2 from 2016 to 2020

Score Stability Over Time

An analysis was conducted to analyze the percent of each final rating over time. Figures 2-4 show the results graphically. For the Arabic exam, the distribution of the final exam ratings was similar (within approximately 10%) over time for ratings of “IM” and lower. However, there was more fluctuation in the distribution of scores at the higher end of the scale. The greatest difference occurred in the “S” rating in which nearly 40% of the exams had an “S” rating in 2016, but less than 5% had ratings of “S” in the available 2020 data. In addition, 2020 had a higher percentage of “IH” and “AL” ratings than any previous year. The distribution of the

Spanish ratings was fairly consistent across the years for each rating. The Russian exam saw some variability but were also fairly consistent across the years for each rating. As with the Spearman correlations shown in Figure 1, it is possible the shortened data collection for 2020 had some impact on these results.

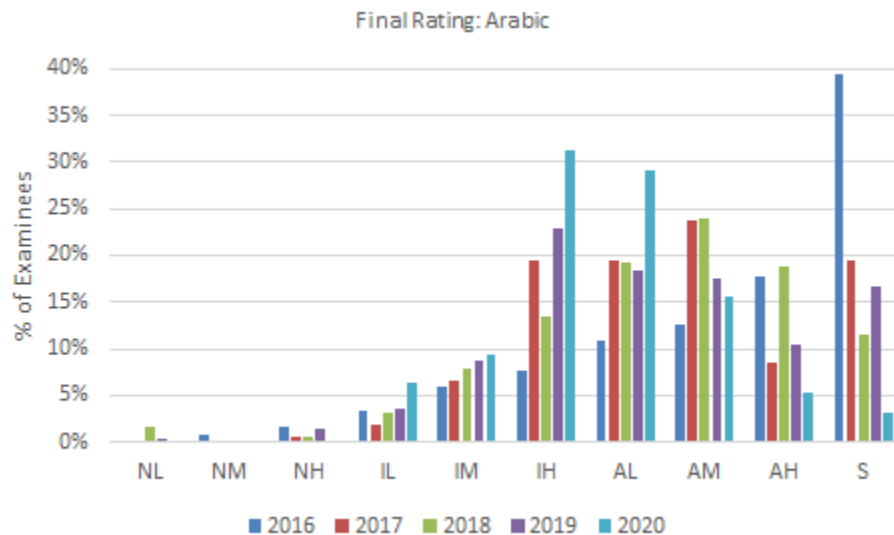


Figure 2. Final ratings from 2016 to 2020 for the Arabic WPT

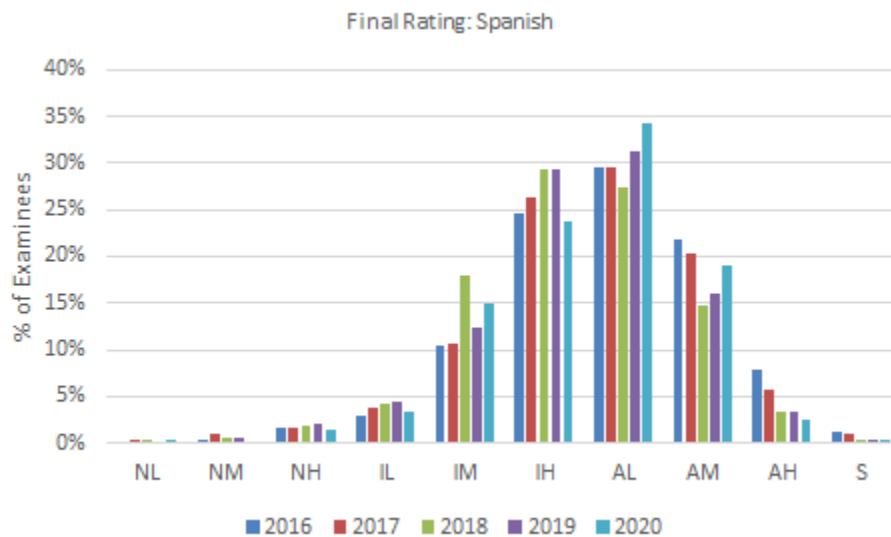


Figure 3. Final ratings from 2016 to 2020 for the Spanish WPT

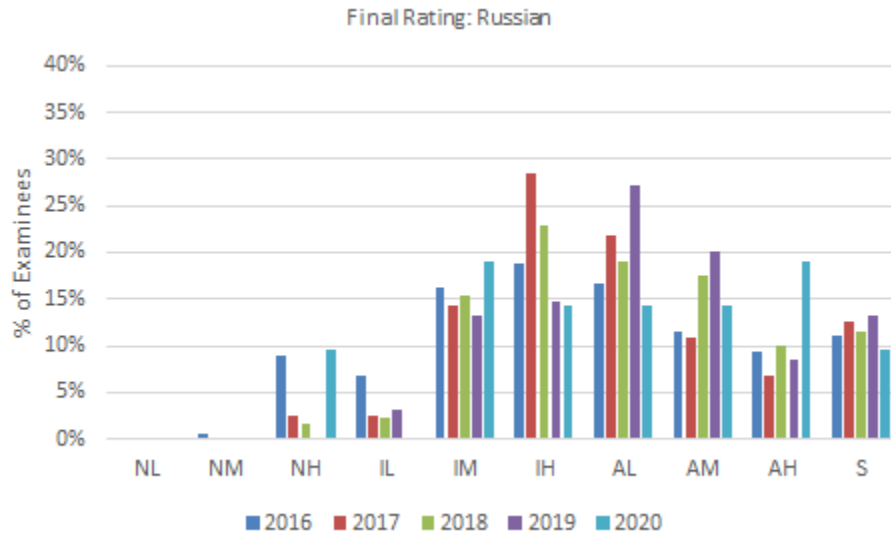


Figure 4. Final ratings from 2016 to 2020 for the Russian WPT

Evidence of Validity

Content Related

The *ACTFL Proficiency Guidelines – 2012 - Writing* describe the range of content and contexts a writer at each major level should be able to handle. These guidelines serve as the core underpinning of the WPT prompts. The writing guidelines are based on a hierarchy of global tasks and specify four major levels of proficiency (Superior, Advanced, Intermediate, and Novice). The WPT prompts are written to elicit a writing sample that aligns with each of the major proficiency levels. To do this, the prompts focus on global functions, contexts and content areas, discourse type and length, and accuracy, as outlined for the major levels in the *ACTFL Proficiency Guidelines – 2012 – Writing*.

Criterion Related

Scores from the current WPT® have not been compared with any related measures of language performance that would allow for criterion-related validity evidence. The exam scores are used for a variety of purposes including language fluency certification, employment selection, placement, and college credit; therefore, standardized measures of later performance would be difficult to obtain. In addition, the WPT® is not meant for use as a predictor of performance, but rather as a global assessment of functional writing ability in a language that can indicate readiness for a given purpose. Since the intended use of the exam is not to predict levels of performance, traditional criterion-related validity evidence is not directly applicable.

Construct Related

Traditional construct-related evidence typically involves correlation of one measure of a trait with other measures of the same or similar traits. It is not unusual for researchers to gather such data with, for example, psychological measures where the trait is tested indirectly (e.g., depression inventories). Scores from the current WPT® have not been compared with any related tests of language ability largely because the WPT® is a direct measure of language ability, and high correlations with similar direct measures of language ability would add little to the validity argument.

Possible Test Bias

The use of a Background Survey allows the test taker to avoid selecting items that might 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, political viewpoint) when developing WPT® writing prompts.


However, no demographic data is collected on the examinees that would allow for measurement of bias or adverse impact.

Evidence that Time Limits are Appropriate and that the Exam is not Unduly Speeded

The ACTFL WPT® has a fixed amount of time of 90 minutes allowed for the completion of the exam. Ten minutes are allotted at the beginning for test takers to complete the Background Survey, Self-Assessment, keyboard selection, and warm-up. The candidate has 80 minutes to complete the actual assessment. Based on the Self-Assessment, the test will focus on two levels of proficiency, thereby, avoiding time spent on unnecessary writing prompts.

Test takers can choose to allocate whatever time they deem necessary to each prompt, within the total time frame of the assessment. Additionally, each prompt includes a suggested amount of time to respond to the prompt and a suggested length for each response. The suggested times and lengths are specific to the proficiency level of the prompt – how long it would take for an individual to provide a sufficient response that meets the criteria for the level and the task. More time is suggested to respond to a higher-level task than for a lower level task. The higher a test taker's proficiency, the more language they can produce. As such, those with higher proficiency levels will take more time with the assessment than those with lower proficiency levels.

Once the allotted time has elapsed, the test taker is automatically exited from the online assessment. When taking a paper and pencil booklet form of the WPT®, candidates are required to stop writing at the conclusion of 90 minutes. The proctor is responsible for adhering to the allotted time limit.



Test takers typically complete the assessment in 40-70 minutes, depending on their level of language proficiency.

During the development of WPT prompts, item writers pay close attention to the amount of time required for successful completion of each level task when attempted by test takers of varying levels of proficiency. Prior to their inclusion into the item pool, the writing prompts are pre-tested. Those items that do not elicit the expected level of response within the expected allotted time are modified or removed.

Provisions for Standardizing Administration of the Examination


The WPT test structure is governed by a detailed format and by the global language functions delineated by the criteria referenced by the Assessment i.e. the *ACTFL Proficiency Guidelines*. As per the *ACTFL Test Development Overview* and the *WPT Examinee Handbook*, test administration begins with an introduction to the assessment directions, online navigation, keyboard selection, and warm-up activity. The Background Survey and Self-Assessment determine the pool of prompts from which a variety of topics will be randomly selected for the writing tasks. The choices the test taker makes in response to the Background Survey and Self-Assessment ensures the uniqueness of the test for each test taker. Four separate prompts are presented to the test taker are designed to elicit written language that demonstrates a test taker's ability to consistently complete linguistic tasks which provide evidence of their proficiency. These tasks are designated by the protocol and are in-line with the functions that are identified within the *ACTFL Proficiency Guidelines*.

Irrelevant Sources of Difficulty Affecting Test Scores

A formal study of construct irrelevant variance for the WPT® has not been undertaken. However, some likely sources of construct irrelevant variance are addressed through ACTFL's exam policies and procedures. Rater training is extensive, and scoring is done against a standardized rubric (see the *ACTFL Examinee Handbook*, page 22 and the *ACTFL Writing Proficiency Test Familiarization Manual*, pages 4-6). The use of the background survey to select prompts most likely to be familiar to the examinee may help to minimize context effects (see the *ACTFL Writing Proficiency Test Familiarization Manual*, page 7). As described above, administration procedures are standardized to ensure the examinee testing experience varies as little as possible.

Provisions for Exam Security

Per *ACTFL's Assessment Integrity Policy*, "A test taker's language must be representative of their own language abilities (speaking, writing, listening, or reading) at the time of the test." Measures have been put into place in order to protect both test content but also the proficiency-based framework for this assessment.



Official WPT®s are administered in proctored environments. As per ACTFL's standard operating procedure document, proctors must apply and be accepted by the test administration office. They must sign an agreement verifying that they understand and can apply ACTFL proctoring protocols.

Administration of the WPT Paper/Pencil Booklet requires the proctor to maintain security of the test materials prior to, during, and post administration to ensure the integrity of the assessment and test takers' responses. Prompt action to proctoring responsibilities post-WPT paper/pencil administration are required to maintain the integrity of the test takers' responses.

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 must 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 members are permitted to act as proctors.

University or College

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

Corporate Clients


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

In addition, educational or business proctors must have a work e-mail address; the e-mail address must contain the proctor's name and the organization's name. Personal e-mail addresses (e.g., AOL, Hotmail, Comcast, Verizon) are not accepted for proctors.

In addition to face to face proctoring, ACTFL also offers remote (virtual) proctoring which make use of a test taker's webcam to identify the test taker and monitor the computer screen and testing environment.

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). In addition, *ACTFL's Re-test Policy* prohibits re-tests within ninety days of a test date in support of the proficiency-based framework. WPT tasks per language are retired based on their ability to elicit the targeted linguistic features (i.e. performance) and/or due to overexposure.

The written samples are digitally stored within the Language Testing International (LTI) secure database. The record is stored under a test identification number which may be looked up on the certificate verification site. 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 2 weeks; once a candidate has logged into an assessment, they must complete that assessment in one sitting within 2 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 ready for suspicious behavior: a significant change in writing ability from one task to another, patterned errors that suddenly disappear, change in handwriting. 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.

Interpretations and Conclusions

To conclude, the ACTFL WPT® met the minimum inter-rater reliability and agreement requirements. Agreement between raters occurred in all examined languages within one sublevel of each other over 85% of the time, and within two sublevels 99% of the time. The highest absolute agreement was Novice Low in both Spanish and Arabic, and at Superior for Russian. The lowest absolute agreement was found at Advance High for Spanish and Arabic, and Intermediate Low for Russian.

The findings of the Spearman's *R* Correlation analyses demonstrate that the correlations of the ratings are almost always positive and strong, ranging from 0.75-0.90. The correlations for the Russian exam were consistently the highest among the three exams, while the Arabic exam were the lowest. Suggested areas of improvement based on the analyses include raising absolute rater agreement within the Intermediate High – Advanced Low and Advanced High – Superior borders across languages. The results of this analysis confirm the reliability of the ACTFL WPT® as an assessment of written proficiency.

References

ACTFL (2020, May). *Quality Control of ACTFL Assessments: Assessment Integrity Policy*.

<https://www.actfl.org/center-assessment-research-and-development/actfl-assessments/quality-control-actfl-assessments>

ACTFL (2020). *Methods of Safeguarding ACTFL Assessments within Test Administration Processes and Procedures*. (ACTFL Publication No. AAR/SOP/P--2020-001). Alexandria, VA: ACTFL Assessment Program.

ACTFL (2020, May). *Quality Control of ACTFL Assessments: Re Test Policy*.

<https://www.actfl.org/center-assessment-research-and-development/actfl-assessments/quality-control-actfl-assessments>

ACTFL (2020). *ACTFL Test Development Overview*. (ACTFL Publication No. AAR/SOP/P—2020-002). Alexandria, VA: ACTFL Assessment Program.

ACTFL (2020). *ACTFL Writing Proficiency Test Familiarization Guide*.

<https://www.actfl.org/sites/default/files/assessments/2020%20WPT%20Familiarization%20Guide.pdf>

ACTFL and Language Testing International (2019). *ACTFL WPT Examinee Handbook*.

<https://www.languagetesting.com/pub/media/wysiwyg/manuals/wpt-examinee-handbook.pdf>

ACTFL (2012). *ACTFL Proficiency Guidelines*. <https://www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012>.

Swender, E. & Vicars, R. (eds.) (2012). *ACTFL Oral Proficiency Interview Tester Training Manual*. Alexandria, VA: ACTFL.