UNIVERSITÄT LEIPZIG

Centre for Test Development and Research
Professor Dr. Erwin Tschirner
Dr. Olaf Bärenfänger
☎ 0341 9737 570
🖷 0341 9737 547

Universität Leipzig, Herder-Institut, Beethovenstraße 15, 04107 Leipzig

Leipzig, April 23, 2011

An Extension of Inquiry into Reliability Issues of the

ACTFL Writing Proficiency Test (WPT)

Technical Report 2011-US-PUB-1

Prepared for:

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

Language Testing International
White Plains, NY

Prepared by:
Centre of Test Development and Research

Dr. Erwin Tschirner
Gerhard-Helbig-Professor of German as a Foreign Language

Dr. Olaf Bärenfänger
Director, Language Learning Centre

Universität Leipzig
Herder-Institut
Beethovenstr. 15
04107 Leipzig

Telefon 0341 97-37570
Fax 0341 97-37547

tschirner@uni-leipzig.de
www.uni-leipzig.de/herder

An Extension of Inquiry into Reliability Issues of the ACTFL Writing Proficiency Test (WPT)

Olaf Bärenfänger, Erwin Tschirner

Introduction

In 2004, Surface and Dierdorff investigated reliability and validity aspects of the ACTFL Writing Proficiency Test (WPT) in a preliminary study. Their main focus of interest was on reliability and interrater consistency as well as on concurrent validity of the WPT and the ACTFL Oral Proficiency Interview (OPI). The study involving a total of 509 WPTs provided evidence for a very high interrater reliability (r = .94). Measures of interrater agreement corroborated these results with 80% of the raters showing perfect agreement. Under a longitudinal perspective, interrater reliability measures indicated a moderate increase in reliability over three years. Additional evidence of validity was provided by a correlation of a data subset (n = 460) between WPT and OPI ratings of r = .81. Taken collectively, the results of the preliminary study offered strong support for the reliability of the WPT rating procedure as well as for the validity of the construct.

In the original 2004 WPT study, a traditional "fixed form" version of the ACTFL Writing Proficiency Test (WPT) was administered which encompassed writing tasks from the Novice through the Superior levels of Writing Proficiency according to the ACTFL Proficiency Guidelines - Writing (revised 2001). The language examined in that WPT study was Spanish.

This current study examines the ratings of 166 internet English WPTs that were administered in Korea in November and December 2010 to adult second language learners of English. In this study, the 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. Based on the statement selected, each test-taker receives a version of the WPT that addresses his or her potential range of writing proficiency: Novice to Intermediate, Intermediate to Advanced or Advanced to Superior. This change was initiated by ACTFL in order to meet the marketplace demand to shorten the test time of the WPT for lower level test-takers, while still maintaining its validity and rating reliability. The new format gives the test-taker more opportunity to demonstrate the writing tasks they can perform and those at the next level which they cannot perform in order to elicit a more focused and ratable sample. It is also the intention of the new version of the test to be more efficient and avoid presenting lower level writers with writing tasks far beyond their ability and upper level writers with writing task far below their capability.

This current study intends to replicate the findings of the 2004 WPT study as far as interrater consistency is concerned, both in terms of a correlation between the ratings of the two WPT ratings as well as in terms of their agreement. As in the previous study, these measures include:

Raw percentages of agreement. The ratings from judges 1 and 2 are cross-tabulated. This kind of analysis reveals the degree of match as well as deviant ratings.

Pearson's correlation. At the level of interval data, Pearson's correlation r_s assesses the degree to which ratings covary. The closer r_s is to 1.00, the higher is the positive linear dependence between two entities; an $r_s = 0.00$ indicates that two entities are independent of one another.

Spearman's rho. This measure assesses the extent to which a relationship between two variables may be described as a monotonic function. It is commonly used in the case of ordinal data. Since the intervals between the levels at the ACTFL scale increase as one moves up the scale, the computing of *rho* is appropriate in this case. *Rho* is calculated by ordering the data by rank and by subsequently correlating the two rank orders. *Rho* may be interpreted in a similar way as Pearson's correlation.

Kendall's tau. The relationship between the two ratings is also analyzed using Kendall's *tau*, a measure of agreement. Like Spearman's *rho*, *tau* is calculated on the basis of rank orders. Whereas *rho* is based on the proportion of variability accounted for, *tau* is a measure of agreement. Thus, *tau* expresses the difference between the probability that participants are rated in the same order and the probability that participants are rated in different orders. Again, a *tau* value of 1.00 stands for a perfect correspondence between the two tests and 0.00 for a non-existing correspondence.

Goodman and Kruskal's gamma. This is another measure of agreement. Unlike Kendall's *tau*, *gamma* ignores bindings, i.e. cases when two participants were assigned the same rank (e.g. because they both had received an *Intermediate Mid* rating). A computation method that is insensitive to bindings provides helpful insights especially when there are few categories on a scale as is the case here. As a consequence of the calculation method, *gamma* tends to be higher than Kendall's *tau*.

Cohen's kappa. Cohen's *kappa* is one of the most common measures of interrater reliability, comparing the observed agreement of ratings with agreement that would be expected by chance. A *kappa* of 1.00 stands for perfect interrater reliability and a *kappa* of 0.00 for an agreement that is completely random. As in the present study ordinal data are involved, Cohen's weighted *kappa* was calculated. This measure assigns different weights to categories of agreement and of disagreement. This way, different levels of disagreement contribute differently to the overall reliability.

Procedures

Subjects. In this study, 166 tests of the English WPT were rated by 7 ACTFL certified raters.

Design. Each rater assessed between 1 and 71 tests, the average being 47.43 tests. Each test received at least two ratings. In the case of a discrepancy between the two raters, arbitration by

a third rater was needed. In the present study, 34 tests were rated by a third rater. Note that the measures of agreement and correlation reported always refer to the non-arbitrated ratings.

Data analysis

Table 1
Percentages of Interrater Agreement

Agreement	Disagreement Distance	
Absolute	1 Step	2 Steps
79.52	19.28	1.20
(132)	(32)	(2)

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

Table 1 presents the results from the interrater agreement analysis. In the previous study, absolute agreement was 80.1%. The results of this study are almost exactly the same, i.e., 79,52%. As in the 2004 study, almost all disagreements were by only one step (19,28%), with only 2 (1,2%) involving two steps.

Table 2
Percentages of Interrater Agreement by Proficiency Category

	Agreement per Category	Disagreement Distance per Category	
	Absolute	1 Step	2 Steps
Novice	72.97	24.32	2.70
	(27)	(9)	(1)
Intermediate	82.11	17.07	0.81
	(101)	(21)	(1)
Advanced	66.67	33.33	.
	(4)	(2)	

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

Table 2 shows the degree of agreement for the major ACTFL proficiency levels. Since judges did not award any Superior ratings, this category is not included in the table. As can be seen, the most mismatched ratings occurred at the Advanced level. This finding needs, however, to be considered with caution, because only a very small number of cases form part of this category. A future analysis focusing on the Advanced level with a larger number of cases might shed more light on the nature of the relatively low absolute agreement in this study. Since the overall agreement is 79.52% and thus very high, there is no reason for concern regarding the reliability of the rating procedure as such.

Table 3
Percentages of Interrater Agreement by Proficiency Subcategory

	Agreement per Subcategory	Disagreement Distance per Subcategory	
	Absolute	1 Step	2 Steps
Novice			
Low	.	.	.
Mid	58.33 (7)	41.67 (5)	.
High	80.00 (20)	16.00 (4)	4.00 (1)
Intermediate			
Low	79.63 (43)	18.52 (10)	1.86 (1)
Mid	84.31 (43)	15.69 (8)	.
High	83.33 (15)	16.67 (3)	.
Advanced			
Low	66.67 (4)	33.33 (2)	.

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

Table 3 allows for an even more fine-grained analysis of interrater agreement. In addition to the relatively low proportion of absolute agreement at the Advanced level mentioned above, this table shows a similar relatively low proportion of agreement at the Novice Mid level. Again, this is most likely due to the limited number of cases in that category and appears to be an artifact of the present study. Future analyses, however, might want to increase the number of cases at this level as well as at the Advanced Low level.

Table 4
Interrater Consistency for the Writing Proficiency Test

n	**Pearson r_s**	**Spearman's *rho***	**Kendall's *tau***	**Goodman-Kruskal's *gamma***	**Cohen's weighted *kappa***
166	.914	.917	.875	.970	.830

Note: All correlations and measures of agreement are significant ($p < .001$).

As the degree of rater agreement in the previous sections indicates, there is a close relationship between the two ratings. In fact, all consistency measures in table 4 replicate the high interrater consistency reported in the 2004 WPT study. All measures of correlation (Person's r_s, Spearman's *rho*) and agreement (Kendall's *tau*, Goodman-Kruskal's *gamma*, Cohen's weighted *kappa*) indicate a very high interrater consistency that is needed to satisfy the requirements of a high stakes test.

Conclusion

As far as interrater consistency is concerned, the present study replicated all of the relevant findings of the 2004 WPT study. All measures of correlation and agreement clearly indicate that the rating procedure of the WPT is highly reliable. Changing the format of the test has not changed its rating reliability either in a more positive or negative way.

References

Surface, E.A. & Dierdorff, E.C. (2004). Preliminary reliability and validity findings for the ACTFL Writing Proficiency Test. Yonkers, NY: ACTFL.