UNIVERSITÄT LEIPZIG
c/o

Institute for Test Research and Development

Professor Dr. Erwin Tschirner
Herder-Institut, Universität Leipzig
Beethovenstraße 15, 04107 Leipzig
sprachtests@uni-leipzig.de

Leipzig, June 22, 2013

Assessing Evidence of Validity of the

ACTFL Listening Proficiency Test (LPT)

Technical Report 2013-US-PUB-2

Prepared for:

American Council on the Teaching of Foreign Languages
Alexandria, VA

Prepared by:
Institute for Test Research and Development

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

Dr. Olaf Bärenfänger
Director, Language Learning Centre

Assessing Evidence of Validity of the ACTFL Listening Proficiency Test (LPT)

Olaf Bärenfänger, Erwin Tschirner

Introduction

The ACTFL Listening Proficiency Test (LPT) is a criterion referenced test developed by the Institute for Test Research and Development at the University of Leipzig in 2012. It is based on the ACTFL Proficiency Guidelines – Listening 2012. It may also be rated according to the ILR Skill Level Descriptions – Listening and the Common European Framework of Reference for Languages (CEFR). Currently, there are ACTFL LPT parallel tests in seven languages: English, French, German, Italian, Russian, Spanish, and Portuguese.

This report describes the ACTFL LPT and it summarizes the original study conducted with test-takers at the University of Leipzig, Germany, to assess evidence of validity of the ACTFL LPT as a measure of listening proficiency in English and to determine the original cut points. This was accomplished by completing a side-by-side study of the ACTFL LPT with NATO's Benchmark Advisory Test – Listening (BAT-L). In addition, this study summarizes and explains the internal validity studies completed for every single ACTFL LPT.

The ACTFL Listening Proficiency Test (LPT)

The ACTFL Listening Proficiency Test (LPT) developed by the Institute for Test Research and Development at the University of Leipzig is based on a crosswalk between the ACTFL, ILR, and CEFR skill level descriptions for listening proficiency. The crosswalk was completed by ACTFL certified tester trainers familiar with all three scales and with ACTFL, ILR and CEFR training. At each level, common features were identified and weighted. Common features across all three scales were retained. Where there were any discrepancies, the features of the ACTFL 2012 level descriptions were selected.

On the basis of the crosswalk, test rubrics were developed for the levels Intermediate Low (IL) to Superior (S) that include genre, content areas, rhetorical organization, vocabulary, what is understood, clarity of speech, and text length. Genre, content areas, what is understood, clarity of speech, vocabulary and rhetorical organization were derived from the ACTFL, ILR and CEFR listening descriptors. Vocabulary ranges are defined according to frequency, idiomaticity, precision and other features, while rhetorical organization is defined on a continuum of very simple, loosely connected texts via description and narration to argumentation and hypothesis. The variation in text length depends on genre and type of rhetorical organization as well as on increasing difficulty and varies from 40 to 300 words per text.

Each proficiency level, from IL to S, is tested separately. There are five levels. They include Intermediate Low (IL), Intermediate Mid (IM), Advanced Low (AL), Advanced Mid (AM) and Superior (S). The differences between IL and IM on one hand and AL and AM on the other are differences in quantity and quality. With respect to the Low levels, the Mid levels exhibit a greater variety of topics, longer texts with longer sentences, and greater breadth of vocabulary. The Mid levels also exhibit a few features of the next higher level, e.g., an emerging ability to make inferences at the AM level. The text type or rhetorical organization of Low and Mid, however, remains the same, i.e., loosely connected texts imparting basic information at the IL and IM levels and description and narration at the AL and AM levels.

Each level tested consists of five listening passages accompanied by three tasks with four multiple choice responses only one of which is correct. Tasks vary from level to level. At the lower levels these tasks include global, detailed, and selective questions and at the higher levels global, detailed, and inference questions. Passages and tasks align at each level. A detailed or global question at the IL and IM levels may be answered by understanding single utterances while at the AL and AM levels these questions require an understanding of information that is spread out across several utterances or across the whole passage.

The ACTFL LPT is a timed test with a total test time of 25 minutes per level. Test takers usually take two or three levels at the same time. Two levels are rated together, either the two levels taken or, if more than two levels were taken, the two highest levels that can be rated according to the specific algorithm of the test. The algorithm and cut points for each level were determined empirically (see below). The test is internet-administered and computer-scored.

Linking the LPT to the ACTFL Proficiency Guidelines – Listening 2012

The ACTFL LPT is linked to the ACTFL Proficiency Guidelines – Listening 2012. A formal, documented linking procedure was followed: The linking process used consists of two phases.

- Systematic documentation of the test information:
    - Description and analysis of the test quality including the general examination content; the process of test development; the process of marking, grading, and establishing results; a test analysis and post-examination review;
    - Description of the test in relation to its framework including the overall estimation of the examination level as well as the communicative activities and aspects of communicative language competence tested.

- Empirical validation of the test by analyzing test data; documentation of both, the psychometric characteristics of the test and its relation to other tests:
    - Analyzing the psychometric quality of the test and collecting evidence for its internal validity;
    - Comparing the test with an external criterion – e.g., another previously validated test – collecting evidence for its external validity.

The ACTFL LPT was linked to the ACTFL Proficiency Guidelines according to the procedures discussed above. The design statement of the ACTFL LPT is based on the ACTFL Proficiency Guidelines – Listening 2012. Test items were authored in complete accordance with the test specifications as outlined in the design statement, the test blueprint, and the construct matrix. Test items were developed by trained item writers with a university degree and extensive experience in foreign language teaching.

Each test is piloted and a number of psychometric analyses are carried out. A test is released only when both, measures of classical and probabilistic test theory point to a high degree of internal validity (see below). Released tests meet all requirements of a standardized high stakes test. To determine the degree of external validity, a correlation study between the LPT and NATO's Benchmark Advisory Test – Listening (BAT-L) was undertaken. The BAT-L is based on NATO's STANAG 6001 (version 3) descriptors which in turn are based on the US Government's ILR Skill Level Descriptions. There are clear and well-established correspondences between the ILR Skill Level Descriptions and the ACTFL Proficiency Guidelines. The STANAG descriptors were developed by NATO's Board of International Language Coordination

(BILC). Because of the correspondence between the ACTFL Proficiency Guidelines and the STANAG 6001.3 and because of the high validity and reliability coefficients of the BAT-L, a correspondence between the ACTFL LPT and BAT-L provides a large measure of external validity and supports the validity argument for the ACTFL LPT.

Table 1 shows the correspondence between the ACTFL, STANAG 6001.3, and ILR skill level descriptions. As can be seen, STANAG 6001.3 stops at ILR level 3 (= Superior). The correlation study below, therefore establishes correspondences between the ACTFL LPT and BAT-L up to the Superior level.

Table 1
Correspondence between ACTFL, STANAG 6001.3, and ILR skill level descriptions

ACTFL	STANAG	ILR
		5
Distinguished		4
Superior	3	3
Advanced	2	2
Intermediate	1	1
Novice	0	0

Procedure

The following section describes in detail the analyses that were carried out to determine the internal validity of the ACTFL LPT as well as how insights about its external validity were gained.

SUBJECTS: The subjects were students of English at the University of Leipzig ranging from beginning to very advanced levels. A total of 88 students took both the ACTFL LPT and the BAT-L. To assure a relatively even distribution of proficiency levels, an almost equal number of participants were selected from Beginning, Intermediate 1, Intermediate 2 and Advanced English courses. Also included in the sample were advanced students of English teacher education, American Studies, and Translation Studies to gain insights into the ACTFL Superior level. Since beginners in university language classes in Germany are rare, the proportion of participants with beginning proficiency in English is smaller than that of participants with more advanced proficiency.

DESIGN: Both, the ACTFL LPT and the BAT-L were administered to the same group of students in a split test design. Half the participants took the ACTFL LPT first; the other half took the BAT-L first. Participants took both tests internet-delivered under controlled proctored conditions in University of Leipzig computer labs. The tests were taken at different days to prevent participant fatigue. Lower proficiency students took ACTFL LPT levels IL, IM, and AL and BAT-L levels 1 and 2. Mid-level proficiency students took ACTFL LPT levels AL and AM and BAT-L levels 1 and 2. High-level proficiency students took ACTFL LPT levels AL, AM, and S and BAT-L levels 2 and 3. Participants were given 75 minutes for the three-level ACTFL LPT and the BAT-L and 50 minutes for the two-level ACTFL LPT. Tests were computer-scored according to their internal scoring algorithms. For the three-level ACTFL LPT, the two highest levels that had at least sixty per cent of the items correct were scored to arrive at the final rating.

STATISTICAL ANALYSES: To determine the *internal validity* of the ACTFL LPT, two types of analyses were carried out. Within the framework of classical test theory, Cronbach's alpha was computed for each level of the test as a measure of overall reliability. In addition, information about the reliability of each individual item was collected by calculating item difficulty parameters and item discrimination parameters. Because classical test theory has been criticized for a number of shortcomings (Bond & Fox, 2007), probabilistic test theory (Rasch dichotomous model) was used to provide a further perspective and to gain more fine-grained insights into the validity of the ACTFL LPT.

The following measures were calculated:

Cronbach's alpha: This is a measure of the internal consistency of a test and provides an overall reliability estimate (Cronbach, 1951/1980). The higher the correlation between test items, the higher also is Cronbach's alpha. Although this parameter may take negative values, only positive values may be interpreted meaningfully. An alpha close to zero indicates low reliability; an alpha close to one indicates very high reliability.

Difficulty index: Traditionally, the difficulty of an item is calculated by determining the proportion of participants who answered the item correctly. A fraction of .5 means that 50% of all participants answered the item correctly. An item revision is usually recommended when the item is too easy, i.e., when its difficulty index is larger than .9, or when it is too difficult, i.e., when the difficulty index is smaller than .1 (Kline 2005).

Separation index: The separation index is a measure of the degree of how well an individual item discriminates between participants of different levels of ability. A participant with low ability, for example, is more likely to answer an item wrong than a participant with higher ability. One way of calculating the separation index is to correlate all answers for an individual item with the answers of all the other items (item-to-scale correlation, cf. Kline 2005). The separation index may take values between -1.0 and +1.0 with a value of .0 indicating no discrimination at all. The closer the discrimination index is to 1.0, the better the item discriminates between participants with different degrees of ability. As a general rule, a separation index value of .5 and better is considered to be sufficient.

Rasch dichotomous model: As part of probabilistic test theory, the Rasch dichotomous model assumes a mathematical relationship between three variables: the difficulty of a test item, the test candidate's ability, and the probability that the candidate will answer the item correctly (Wright/Mok 2004). The probability, e.g., that a person with low ability will answer a difficult item correctly is low. Formally, this relationship may be described as follows:

$$\log_e\left(\frac{P_{ni}}{1-P_{ni}}\right) = B_n - D_i$$

where p_{ni} denotes the probability that a person answers item i correctly; B_n denotes the person's ability; and D_i the difficulty of the item. All parameters are located on the same scale and are expressed in logits (because of the logarithmic transformation in the formula above). Winsteps software version 3.70.1 (Lincacre, 2010) was used to compute the three parameters of the dichotomous Rasch model.

The dichotomous Rasch model is the original model on which a growing number of different Rasch models are based. Because they all compensate for one of the major shortcomings of classical test theory –

that all parameters may only be interpreted with respect to the subject sample – they are increasingly used to analyze data in a broad range of subject areas (Bond & Fox 2007). Because all parameters are scaled on the same metric, one can easily interpret the meaning of the parameters calculated. These parameters are objective in the sense that they are sample independent and they provide more detailed information. Therefore, they have become the accepted tool to assess the quality of test items and to link specific tests to other tests as well as to calibrate item difficulty by means of test equating.

The dichotomous Rasch model not only allows for estimating item difficulty and person ability, but also for calculating the degree of fit between the model and the data (expressed in infit and outfit values) as well as for judging the ability of the test to discriminate between subjects (Bond & Fox, 2007). Infit statistics are sensitive to irregular answer patterns for test takers who respond to items appropriate for their ability. Outfit statistics provide information about irregular answer patterns for test takers whose competence does not correspond to the difficulty of an item. Both fit values as well as the person separation value may be interpreted in terms of reliability. Ideally, infit and outfit values are close to 1.0. The separation value is interpreted in a similar way as Cronbach's alpha, i.e., the closer to 1.0 it is, the better.

To gain insights about the *external validity* of the ACTFL LPT, raw percentages of agreement between the ACTFL LPT and BAT-L were cross-tabulated and several correlation values were computed.

Raw percentage of agreement: After determining the final ratings for both the ACTFL LPT and BAT-L, their absolute and relative frequencies were cross-tabulated. This kind of analysis reveals the degree of match as well as the number of deviant ratings.

Pearson's correlation: For the final ratings, Pearson's r_s was computed. At the level of interval data, r_s assesses the degree to which ratings covary. The closer r_s is to 1, the higher is the positive linear correspondence between two entities; an $r_s = 0$ 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 computed for analyses involving ordinal data. Because the final ratings of the two tests cannot be assumed to be interval data, the calculation of *rho* is more appropriate. *Rho* is calculated by ordering the data by rank and subsequently correlating the two rank orders. *Rho* is therefore also called a rank-order correlation. Its results are interpreted in a similar way as Pearson's correlation.

Kendall's *tau*: The relationship between final ratings was 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. *Tau* expresses the difference between the probability that participants are rated in the same order and the probability that participants are rated in a different order. A *tau* value of 1.0 stands for a perfect correspondence between the two tests and 0 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 where two participants are assigned the same rank (e.g., because they both received an IM rating). Due to the fact that the final rating consists only of 5 values in the case of the LPT (IL, IM, AL, AM, S) and 6 values in the case of the STANAG 6001.3 (0, 0+, 1, 1+, 2, 2+, 3), a computation method that is insensitive to bindings provides helpful insights. As a consequence of the calculation method, *gamma* tends to be higher than Kendall's *tau*.

Data analysis

According to classical test theory, the ACTFL LPT displays high overall reliability at each level with Cronbach's *alpha* at or above .80 (see Table 2 below).

Table 2
Classical reliability estimates for the ACTFL LPT

	LPT Level IL	LPT Level IM	LPT Level AL	LPT Level AM	LPT Level S
Number of items	15	15	15	15	15
Cronbach's *alpha*	.90	.89	.81	.80	.80

The Rasch analysis using the Rasch model for dichotomous data corroborates this finding. Table 3 below shows the fit parameters and person separation reliability for all levels of the ACTFL LPT. Both infit and outfit parameters are close to 1.0 and indicate an excellent model fit. The person separation reliability is close to 1.0 and confirms the excellent classical reliability parameters in the previous table.

Table 3
Rasch model fit and separation parameters for the ACTFL LPT

	Mean Model Infit [MNSQ]	Mean Model Outfit [MNSQ]	Person Separation Reliability
LPT	.97	.88	.81

Note: Infit and outfit parameters are expressed in mean square roots [MNSQ] of the residuals.

Classical item difficulties were calculated for the 75 items of the ACTFL LPT individually. 10 items were considered to be too easy because their values were .9 or higher. Five of them pertained to the IM level. This result is most likely due to a ceiling effect. Because most of the participants had a higher level of proficiency than IM, the proportion of correct responses was very high. At the AL level, there were four easy items and at the S level, there was one. Only four of the 75 items displayed separation indices that were lower than .25. Taking into account the total number of items in the ACTFL LPT, the amount of items with low discrimination is very small and thus only marginally affects the accuracy of the measurement.

Considering all parameters of classical test theory investigated in this study, the individual reliabilities of items are very high. This fact is also reflected at the level of overall reliability as expressed in Cronbach's *alpha* values (see Table 2 above). In addition, all Rasch parameters point to a very high degree of reliability in terms of individual items and of the overall test. To conclude, the internal validity of the ACTFL LPT may be considered high and appropriate for a high stakes test.

Table 4 below lists all measures that were computed to establish the ACTFL LPT's *external validity*. It contains four parameters which indicate the relationship between the ACTFL LPT and BAT-L. Two correlation and two agreement measures were computed. Both correlation parameters, Pearson's r_s and

Spearman's *rho*, show a high interdependence between the two tests. As for the agreement measures, Kendall's *tau* is obviously affected by bindings in the data and hence slightly lower than Goodman-Kruskall's *gamma*. Both indicators support, however, the conclusion that there is high agreement between the ratings of both tests, and by implication, that the ACTFL LPT and BAT-L measure the same construct.

Table 4
Correlation and agreement measures between final ratings of the ACTFL LPT and BAT-L

N	Pearson's r_s	Spearman's *rho*	Kendall's *tau*	Goodman-Kruskall's *gamma*
88	.842	.833	.753	.898

Note: All correlations are significant ($p < 0.01$).

The frequency distribution in Table 5 below also points to a strong relationship between the two tests and corroborates the correlation parameters and agreement measures reported in Table 4.

Table 5
Frequency of agreement in final ratings of the ACTFL LPT and BAT-L

		BAT-L Final Rating						
		0	0+	1	1+	2	2+	3
ACTFL LPT Final Rating	0	1 (1.0)						
	IL		2 (.40)	3 (.60)				
	IM			8 (.57)	3 (.21)	3 (.21)		
	AL			3 (.09)	8 (.23)	23 (.66)	1 (.03)	
	AM			1 (.14)		1 (.14)	2 (.29)	3 (.43)
	S					4 (.15)	6 (.23)	16 (.62)

Note: The proportion of agreement is indicated in parentheses.

As to the nature of the correspondence between LTP and BAT-L, Table 5 above shows the following: IL corresponds to STANAG/ILR 0+ 40% and to STANAG/ILR 1 60% of the time. IM corresponds to STANAG/ILR 1 57% and to STANAG/ILR 1+ or higher 42% per cent of the time. AL corresponds to STANAG/ILR 1+ or lower 32% and to STANAG/ILR 2 66% of the time. AM corresponds to STANAG/ ILR 2 or lower 28% and to STANAG/ILR 2+ or higher 72% of the time. S corresponds to STANAG/ ILR 3 62% of the time.

In order to externally validate the ACTFL level of the ACTFL LPT, the relationship between ILR and ACTFL levels needs to be taken into account. ILR level 1 corresponds to both IL and IM; level 1+ often corresponds to IH but may also correspond to IM; level 2 corresponds to AL and AM; level 2+ often corresponds to AH but may also correspond to AM; and level 3 corresponds to baseline Superior.

The finding that IL corresponds to 0+ (40%) and 1 (60%), i.e. the lower level 1 ranges, is consistent with the relationship between ACTFL and ILR established above. IM corresponds to 1 (57%) and 1+ or higher (42%), i.e., the higher level 1 ranges. This is also consistent with the relationship between ACTFL and ILR. The finding that AL corresponds to 1+ (23%) and 2 (66%), i.e., the lower level 2 ranges is equally consistent. AM corresponds to 2 and 2+ (43%) and even 3 (43%). This points to a correspondence between AM and the higher level 2 ranges. S, finally, clearly corresponds to 3 (62%). Because participant numbers for levels IL and AM are somewhat low, it is suggested that future analyses pay special attention to these two levels.

In summary, both internal and external validity point to the conclusion that the ACTFL LPT may be assumed to be a valid measure of listening competence as defined by the ACTFL Proficiency Guidelines – Listening 2012.

Explaining the Data Reports

All ACTFL LPT go through a rigorous pilot study process. All tests are usually taken by at least 100 participants ideally with 20 participants at each of the five levels. Data reports are completed of all pilot tests. This section explains the format of the data reports.

The data report provides the date on which the report was completed, the name of the test, e.g., English LPT 01A, the name of the person completing the report, the date or dates of data collection and the number of participants.

The data report provides both classical item analyses as well as a Rasch analysis. The classical item analysis provides Cronbach's *alpha* for two adjoining levels, i.e., IL/IM, IM/AL, AL/AM, and AM/S as well as for all levels combined. Cronbach's *alpha* reflects the degree to which the items of two adjoining levels reliably discriminate between test participants of different degrees of ability. Its value for all five levels is an indicator of the overall reliability of the test. Cronbach's alpha should not be lower than .8.

In addition to Cronbach's alpha, the report also provides difficulty and separation indices for each item. Difficulty indices should be close to .5, not lower than .1 and not larger than .9. Separation indices should not be lower than .25.

The data report also provides a Rasch analysis indicating the overall separation reliability, the model fit, and misfitting items if any. The overall separation reliability is interpreted in a similar way as Cronbach's *alpha* and should not be lower than .8.

The model fit values are calculated by comparing empirical answer patterns with the patterns predicted by the Rasch model in the form of a residual analysis. Whereas infit statistics refer to the randomness of the data and thus to threats to the validity of the model with respect to the data, outfit statistics yield information on outliers (Eckes, 2009). Generally, infit statistics are considered more important than outfit statistics (Bond & Fox, 2007; Eckes, 2009). Both infit and outfit mean square values range from 0 to infinity. An infit value of 1.0 indicates that the amount of variance in the data is exactly the amount that is predicted by the model. Mean square values below 1.0 represent less variance in the data than predicted and mean square values larger than 1.0 represent more variance. While mean square values below 0.5 or between 1.5 and 2.0 are considered to be less productive but not degrading, mean square

values above 2.0 distort or degrade the measurement system (Linacre, 2012). For this reason, items with fit values above 2.0 are usually recommended for revision. The closer fit values are to 1.0, the better the model fits the data. Fit values may also be computed for individual items. Again, an item should ideally have infit and outfit values close to one and should not exceed 2.0.

When either classical test or Rasch analyses have identified items with problematic values, the report recommends either a revision of individual items or a further study. When revisions of items would only lead to a minor improvement of the overall test, e.g., when only a few items are slightly beyond a critical threshold, the report usually recommends not making any changes to the items.

Each report concludes with a general statement as to the quality of the psychometric properties of the test and if it may be used for high stakes testing.

Conclusion

Taken collectively, all measures investigated in this study indicate that the ACTFL LPT shows a very high degree of reliability, both in terms of classical and probabilistic test theory. An evaluation of the psychometric characteristics of the ACTFL LPT and its relation to another validated test, the BAT-L, support the argument that the ACTFL LTP is a valid test of listening proficiency as measured by the ACTFL Proficiency Guidelines – Listening 2012 based upon both internal and external validity studies.

Bibliography

Bond, T. G. & Fox, C. M. (2007). *Applying the Rasch Model: Fundamental measurement in the human sciences*, 2nd edn. Mahwah, NJ: Lawrence Erlbaum.

Cronbach, L. J. (1951/1980). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.

Eckes, T. (2009). Many-facet Rasch measurement. In S. Takala (ed.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment* (Section H). Strasbourg, France: Council of Europe.

Kline, J. B. (2005). *Psychological testing. A practical approach to design and evaluation*. Thousand Oaks, CA: Sage.

Linacre, J. M. (2010). *Winsteps*® (Version 3.70.1) [Computer Software]. Online: winsteps.com. Retrieved January 1, 2010.

Linacre, M. (2012). *Many-facet Rasch measurement: Facets tutorial*. Online: www.winsteps.com/a/ftutorial2.pdf. Retrieved August 27, 2012.