

Language Research and Development, Inc.

Erwin Tschirner, PhD
580 White Plains Rd., Ste. 660
Tarrytown, NY 10591
USA

May 2, 2023

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

Technical Report 2023/2-PUB-2

Prepared for:

American Council on the Teaching of Foreign Languages
Alexandria, VA

Prepared by:

Language Research and Development, Inc.

Dr. Erwin Tschirner
President

Table of Contents

1	<i>General Exam Information</i>	4
2	<i>Rationale and Purpose of the Examination</i>	6
3	<i>Names and Institutional Affiliations of the Principal Authors/Consultants</i>	7
4	<i>Score Types Reported for Examinees</i>	10
5	<i>Scoring Directions/Procedures</i>	10
6	<i>Domain Specifications of Content, Skills, and Abilities Tested</i>	11
7	<i>Rationale for the Kinds of Tasks/Items</i>	12
8	<i>Task/Item Inclusion Rationale</i>	13
9	<i>Item Adequacy as a Domain Sample</i>	15
10	<i>Item Currency and Representativeness</i>	22
11	<i>Item Sensitivity Panel Review Description</i>	22
12	<i>Field/Pre-Test Processes and Procedures</i>	23
13	<i>Item Analysis Results</i>	23
14	<i>Internal Consistency Reliability</i>	31
15	<i>Equivalence of Exam Forms Evidence</i>	33
16	<i>Scorer Reliability</i>	36
17	<i>Cut-Score Classification Errors</i>	36
18	<i>Content-Related Validity</i>	36
19	<i>Criterion-Related Validity</i>	37
20	<i>Construct-Related Validity</i>	39
21	<i>Item Bias and Differential Item Functioning</i>	42
22	<i>Time Limit Appropriateness</i>	43
23	<i>Exam Administration Standardization Provisions</i>	43
24	<i>Exam Security Provisions</i>	44
25	<i>IRT Scaling Models Used</i>	45

26	<i>IRT Model’s Evidence of Fit</i>	<i>45</i>
27	<i>Evidence that New Items/Tests Fit the Current Scale Used.....</i>	<i>45</i>
28	<i>Exposure Rate of Items and Operational Test Item</i>	<i>45</i>
29	<i>Equivalency Between Hardcopy and Computer Administration.....</i>	<i>46</i>
30	<i>Cut-Score Rationale, Reasonableness, and Appropriateness</i>	<i>46</i>
31	<i>Procedures Recommended to Users for Establishing their own Cut-Scores</i>	<i>50</i>
32	<i>Information on Norms and Normative Groups (If Appropriate).....</i>	<i>52</i>
	<i>References.....</i>	<i>52</i>
	<i>Appendices</i>	<i>52</i>

ACTFL Listening Proficiency Test (LPT)

This evaluation of the ACTFL Listening Proficiency Test (LPT) follows the Examination Evaluation Checklist as provided by ACE. Where appropriate, the evaluation references documents provided as appendices. Item analysis results, reliability information, and evidence of validity are based on the following three languages: Arabic, German, and Spanish. This report covers the time period January 1, 2020 to January 1, 2023. Where appropriate, additional data for these three languages are used. Table 1 shows the number of tests administered from 1/1/20 to 1/1/23 by language and test form.

Table 1
Assessments Administered from 1/1/20 to 1/1/23 by Language and Form

	Arabic	German	Spanish
Form 1	68	85	
Form 2	277	74	
Form 3	428	69	423
Form 4			631
Form 5			388
Form 6			226
Form 7			5
Form 8			1
Total	773	228	1674

Table 1 shows that three forms of the Arabic and German LPT and six forms of the Spanish LPT were administered in the time period under review for a total of 773 Arabic, 228 German, and 1,674 Spanish tests.

1 General Exam Information

This section provides general information about the examination (see Appendix 1 – Familiarization Manual). The LPT is a standardized test for the global assessment of listening ability in a language. It is a carefully constructed assessment based on the *ACTFL Proficiency Guidelines 2012 – Listening* that evaluates Novice to Superior levels of listening ability. It is an online assessment. The test assesses specific ranges of proficiency. The available ranges are shown in Table 2 below. These options ensure that the test administered targets the range of the examinee’s listening ability economically in terms of time and effort.

Item (Task) Types

There are four item types: Global, Detail, Selective, and Inference (see Table 3 for number of item types per sublevel):

- IL passages have one global, one selective, and one detail item.
- IM passages have one global and two detail items.
- AL passages have one global and two detail items.
- AM passages have one global, one detail, and one inference item.
- S passages have one global, one detail, and one inference item.

Table 3
Number of Item Types per Sublevel

Level	IL	IM	AL	AM	S
Number of items	Global: 5 Selective: 5 Detail: 5	Global: 5 Detail: 10	Global: 5 Detail: 10	Global: 5 Detail: 5 Inference: 5	Global: 5 Detail: 5 Inference: 5

Time Allotment

The time limit for a two-sublevel test is 50 minutes; for a three-sublevel test, it is 75 minutes; for the non-adaptive full-range test (H), it is 125 minutes, and for the semi-adaptive full-range test (G), it is 75 minutes. This amounts to approx. five minutes per task. However, there is only an overall time limit for the complete test. There is a time gauge to let examinees know how much time is still remaining.

2 Rationale and Purpose of the Examination

This section summarizes the rationale and purpose of the examination (see Appendix 3 – Design Statement and Appendix 10 – Examinee Handbook). The ACTFL LPT is the official listening proficiency test of the American Council on the Teaching of Foreign Languages (ACTFL). It assesses how well a person understands listening passages in a world language when presented with passages and tasks as described in the *ACTFL Proficiency Guidelines 2012 – Listening* without access to dictionaries or grammar references. The *ACTFL Proficiency Guidelines 2012 – Listening* describe five major levels of proficiency: Distinguished, Superior, Advanced, Intermediate, and Novice. The description of each major level is representative of a specific range of abilities. Together these levels form a hierarchy in which each level subsumes all lower levels. The major levels Advanced, Intermediate, and Novice are divided into High, Mid, and Low sublevels. The ACTFL LPT assesses listening proficiency at all levels except Distinguished, i.e., from Novice Low to Superior.

3 Names and Institutional Affiliations of the Principal Authors/Consultants

The ACTFL LPT was developed by Dr. Erwin Tschirner (Gerhard Helbig Professor of German as a Foreign Language, University of Leipzig, and President of the Institute for Test Research and Test Development, Leipzig, Germany) and Dr. Olaf Bärenfänger (Director of the Language Learning Center, University of Leipzig, and Vice-President of the Institute for Test Research and Test Development, Leipzig, Germany). Dr. Tschirner and Dr. Bärenfänger also designed the item development process, and both are in charge of overall test validity and quality assurance.

Item development was managed by staff members of the Institute of Test Research and Test Development (ITT). They included Jupp Möhring (M.A. in German as a Foreign Language, University of Leipzig), Elisabeth Muntschick (M.A. in German as a Foreign Language, University of Leipzig), Robin Ide (M.A. in German as a Foreign Language, University of Leipzig), and Sabine Kutschera (M.A. in German as a Foreign Language, University of Leipzig) for the time period under review.

Item Development Process

All items undergo a rigorous, standardized quality assured development process. Text and item writers are native speakers of the language in question with a college degree in foreign language teaching or applied linguistics and with a considerable amount of language teaching and test writing experience. Test reviewers and senior test development officers are native or near-native speakers of the language in question and trained for language proficiency testing. Authors, reviewers, and final quality control specialists undergo a rigorous selection and training process as well as ongoing quality assurance measures as appropriate for high stakes testing.

Items are developed in multiple stages in a controlled process. Authors who are native speakers of the target language develop passages and items according to the Item Writing Manual and the Construct Matrix and submit a first draft. The first draft is reviewed for style and correctness by another native speaker of the target language. The main focus of this review is to ensure that the passages are culturally and idiomatically authentic, well written, and able to hold the listener's interest. Tests are revised by the original author and submitted to an assessment specialist, who checks if the passages and items are at the appropriate levels, if the author has followed the instructions in the Item Writing Manual precisely, and if all items, keys, and distractors follow the norms established. This includes a first round of item sensitivity review to ensure that passages and items are not offensive or biased towards certain groups of examinees. The main focus is on the level appropriateness of the texts and the quality of the items. The assessment specialist is a native or near-native speaker of the target language. Tests are revised again by the original author or by a different native speaker author with similar qualifications.

Before the listening passages are recorded, they undergo a second round of sensitivity review. They are spoken by professional speakers who are trained to speak in a level-specific and criterion-based, authentic manner for the various proficiency levels and text types. The speakers are experienced television or radio speakers, actors, speech scientists and/or world languages teachers with a substantial amount of teaching experience with a special talent for acting. They receive additional training for speaking LPT passages.

The sound recordings for the LPT are completed in sound studios, which comply with the guidelines and the specifications of public broadcasting (developed by the German television networks ARD and ZDF). Recordings and postproduction are undertaken by trained sound engineers. The recordings are professionally edited with background noise and other acoustic features that make them appear more authentic. After postproduction, several rounds of proof listening are carried out until the audio files are entered into the test system at LTI and checked again during User Assurance Testing (UAT).

The items are checked for spelling and punctuation and undergo a second round of item sensitivity review before the test is uploaded to the LTI Assessment System to begin UAT, which typically consists of at least two rounds and often results in additional revisions made to passages and items. The test then enters the operational testing phase with at least 300 examinees at all proficiency levels taking the test. Detailed data reports are developed using IRT analysis (Rasch modeling) (item difficulty logits, SEM, infit and outfit values, separation indices). Any misfitting items or any items that are too difficult or too easy for a particular level are revised or removed.

Table 4 lists the names, qualifications, and languages of the staff, item writers and reviewers who were developing and reviewing items in the time period covered in this report. The column *Other Languages* lists second languages with a proficiency of at least *Advanced Mid* but in many cases *Superior*. The columns *Assessment* and *Teaching* list the years of experience in both fields. People who show no experience in teaching are translators or interpreters, usually with a considerable amount of experience in their profession.

Table 4
Current Staff, Item Writers, and Reviewers

Name	Sex	Degree	Subject	Native Language	Other Languages	Assessment	Teaching
Erwin Tschirner	m	PhD (Berkeley)	Linguistics/SLA	German	Spanish	34	42
Olaf Bärenfänger	m	PhD (Bielefeld)	German	German	French	22	20
Jupp Möhring	m	MA (Leipzig)	German	German	English	12	15
Elisabeth Muntschick	f	MA (Leipzig)	German	German	English	8	9
Robin Ide	m	MA (Leipzig)	German	German	Spanish	10	10
Elisa Hartmann	f	MA (Leipzig)	German	German	English	5	6
Sabine Kutschera	f	MA (Leipzig)	German	German	English	11	15
Emely Barreto	f	BA (Leipzig)	Translation	Spanish	German, English	3	3
Writer/Reviewer 1	f	MA (Guadalajara)	German & Philosophy	Spanish	German, English	5	10
Writer/Reviewer 2	f	MA (Guadalajara)	German	Spanish	German	4	6
Writer/Reviewer 3	m	MA (Guadalajara)	Applied Linguistics	Spanish	German, English	13	17

Writer/Reviewer 4	f	MA (Guadalajara)	German	Spanish	German	3	6
Writer/Reviewer 5	f	MA (Guadalajara)	German	Spanish	German	9	12
Writer/Reviewer 6	m	MA (Leipzig)	Spanish & Translation	Spanish	German	3	7
Writer/Reviewer 7	m	BA (Leipzig)	Communication	Arabic	English, German	3	4
Writer/Reviewer 8	m	MA (Leipzig)	Medical Studies	Arabic	German	3	-
Writer/Reviewer 9	m	MA (Isra University)	Engineering	Arabic, Russian	English, German	3	-
Writer/Reviewer 10	m	MA (Heidelberg)	German	Mandarin	German, English	8	10
Writer/Reviewer 11	f	MA (Beijing)	Chinese	Mandarin	German	3	11
Writer/Reviewer 12	f	PhD (Leipzig)	Media studies	Mandarin	German	11	13
Writer/Reviewer 13	f	MA (Leipzig)	German	Russian	German, Ukrainian	7	7
Writer/Reviewer 14	f	BA (Leipzig)	Management	Russian	German	3	3
Writer/Reviewer 15	f	MA (Leipzig)	European Studies	Russian	German	3	6
Writer/Reviewer 16	f	MA (UFPR Brazil)	Communication	Portuguese	German	18	23
Writer/Reviewer 17	f	MA (UFPR Brazil)	German	Portuguese	German	4	4
Writer/Reviewer 18	f	MA (UFPR Brazil)	German	Portuguese	German, English	13	4
Writer/Reviewer 19	f	MA (Leipzig)	Digital Humanities	Italian	German	3	6
Writer/Reviewer 20	f	MA (Leipzig)	German	Italian	German, Spanish	3	5
Writer/Reviewer 21	m	MA (Leipzig)	English	Italian	English, German	5	-
Writer/Reviewer 22	f	MA (Leipzig)	Educational Studies	Arabic	German	3	5
Writer/Reviewer 23	f	MA (Beijing)	TCSOL	Mandarin	German	7	7
Writer/Reviewer 24	f	MA (Nagoya)	Japanese	Japanese	German	1	2
Writer/Reviewer 25	m	MSc (Leipzig)	Economics	Japanese	German	3	-
Writer/Reviewer 26	m	MA (Leipzig)	Japanese	German	Japanese	2	7
Writer/Reviewer 27	m	MA (Napoli)	Italian	Italian	German	10	13
Writer/Reviewer 28	f	MA (Leipzig)	German	Spanish	German	5	5
Writer/Reviewer 29	f	MA (Guadalajara)	German	Spanish	German	5	5
Writer/Reviewer 30	f	MA (Leipzig)	German	German	English, Spanish	5	-

4 Score Types Reported for Examinees

The ACTFL LPT is a proficiency test reporting proficiency levels as described in the *ACTFL Proficiency Guidelines 2012 – Listening*. Test scores are converted to ACTFL proficiency levels and reported as such (see Section 5).

In addition to the ACTFL listening proficiency level, the certificate also provides a brief description of what examinees who have reached a particular level can do. This helps examinees to place themselves within a continuum of proficiency levels (see Appendix 11 – Certificate).

5 Scoring Directions/Procedures

This section summarizes the scoring procedures (see Appendix 4 – Blueprint). The ACTFL LPT is machine-scored. At least two sublevels are administered and scored together, i.e., IL and IM; IM and AL; AL and AM; or AM and S. To assign a rating, the combined total of the two levels that are rated is used. When there are more than two levels administered, the highest two levels that have at least 18 points between them are used. When there are no two levels that have at least 18 points between them, the highest two levels that have at least 11 points between them are used. When there are no two levels that have at least 11 points between them, the two lowest levels are used. Table 5 shows how test scores are converted to ACTFL ratings. (See Section 30 for information on how the cut scores were determined.)

Table 5
Scoring Algorithm

Sublevels Rated	Total Score	ACTFL Rating
IL-IM	0-11	NL
IL-IM	12-14	NM
IL-IM	15-17	NH
IL-IM	18-23	IL
IL-IM	24-30	IM
IM-AL	0-11	BR*
IM-AL	12-14	NH
IM-AL	15-17	IL
IM-AL	18-21	IM
IM-AL	22-23	IH
IM-AL	24-30	AL
AL-AM	0-11	BR*
AL-AM	12-14	IM
AL-AM	15-17	IH
AL-AM	18-23	AL

AL-AM	24-30	AM
AM-S	0-11	BR*
AM-S	12-14	IH
AM-S	15-17	AL
AM-S	18-21	AM
AM-S	22-23	AH
AM-S	24-30	S

*BR (Below Range) is assigned when the test-taker's ability is lower than the lowest rating that may be assigned by a particular test version.

Table 5 shows what ratings are assigned to what scores given two particular sublevels. BR (Below Range) is assigned to scores of 0-11, because such scores could potentially be achieved by guessing only (see Section 30). For the sublevels IL and IM, the rating NL is assigned to scores of 0-11.

6 Domain Specifications of Content, Skills, and Abilities Tested

This section summarizes the specifications that define the domains of content, skills, and developed abilities that the exam samples (see Appendix 1 – Familiarization Manual, Appendix 2 – Assessment Use Argument, Appendix 3 – Design Statement, Appendix 4 – Blueprint, and Appendix 5 – Construct Matrix).

Based on the *ACTFL Proficiency Guidelines 2012 – Listening*, the construct matrix defines the domains of content, skills, and abilities that the exam measures. The target language use (TLU) task that was selected as the basis for developing assessment tasks (passages and items) is listening in general, i.e., retrieving information from a variety of spoken passages in daily life, at work, university or school etc., indicating different aspects of comprehension (global, selective, detail understanding, and making inferences), depending on the sublevel. Tasks are described in terms of function, content, context, text type, vocabulary, grammar, and culture at all major ACTFL levels (see Table 6 for a summary of the task descriptors).

Table 6
Summary of Task Descriptors at the Proficiency Levels Represented by Test Tasks

	Function	Content	Context	Text Type	Vocabulary	Grammar	Culture
Superior	Argumentation; Supported opinion; Hypothesis	Familiar and unfamiliar abstract topics	Professional; Academic; Literary	Complex, lengthy passages	Broad; Precise; Specialized	Complex structures	Cultural references; Aesthetic properties

Ad- vanced	Description; Narration; Exposition; Explanation;	Concrete current and gen- eral inter- est topics	Public; Education; Work; News	Paragraph- based con- nected passages with a clear pre- dictable structure	Broad general vocabu- lary	Sequencing; Time frames; Chronology	Most com- mon cul- tural pat- terns
Interme- diate	Convey basic infor- mation	Highly fa- miliar eve- ryday con- tent	Highly fa- miliar eve- ryday con- texts	Simple, predicta- ble, loosely connected passages	High fre- quency vocabu- lary	Simple sen- tence pat- terns and strings of sentences	Some of the most common cultural patterns

- The term *function* refers to the different purposes spoken passages may have such as instruction, description, narration, explanation, or argumentation.
- The term *content* refers to the general content areas that the listener can understand in the language.
- The term *context* refers to the different domains in which discourse occurs such as the public, educational or work domain.
- The term *text type* refers to the quantity, quality and organization of passages that the listener can understand in the language.
- The term *vocabulary* refers to the range of vocabulary the listener can understand in the language.
- The term *grammar* refers to the range of grammatical structures that the listener is able to use for comprehension purposes.
- The term *culture* refers to the range of idiomatic expressions and cultural references the listener can understand in the language.

7 Rationale for the Kinds of Tasks/Items

The content, skill, and ability areas are based on the *ACTFL Proficiency Guidelines 2012 – Listening*. Each exam contains items for at least two sublevels. Thus, at least ten passages and 30 items form the basis of a rating. This allows the test to assess a representative sample of real-life topics and to make a meaningful statement about the language proficiency of the examinee. Depending on the sublevels assessed, the listening passages have different functions such as description, narration, explanation, exposition, argumentation, and hypothesis and different contexts such as familiar everyday contexts, work, public, education, academic, professional, and art contexts. For example, the test that assesses the sublevels Advanced Mid and Superior contains ten passages, which represent the functions of both levels, i.e., description, narration, explanation, and exposition at the Advanced level and argumentation, supported opinion, and hypothesis at the Superior level. A similar distribution applies to content and genre. The test involves passages of concrete, current, and general interest topics as well as familiar and unfamiliar abstract topics such as

discussions between educated native speakers, radio broadcasts, news stories, oral reports, and lectures concerned with contemporary social problems, biographical accounts, stories, and opinion/editorial pieces, analyses and commentaries.

8 Task/Item Inclusion Rationale

This section presents the rationale for the kinds of tasks/items included in the exam (see Appendix 3 – Design Statement, Appendix 4 – Blueprint, and Appendix 5 – Construct Matrix). Please see Sections 6 and 7 for the rationale for the kinds of passages included in the exam. This is the rationale for the items:

There are four item types: Global (for the sublevels IM to S), Detail (for all sublevels), Selective (for IL only), and Inference (for the sublevels AM to S). These item types were derived from the *ACTFL Proficiency Guidelines 2012 – Listening* and from the cognitive processing approach to defining comprehension of Weir and Khalifa (2008)¹, in particular, their model of intent (goal setter) with its dimensions of local vs. global and careful vs. expeditious. *Expeditious* was redefined as *casual* for the model used by the LPT. The distribution of item types across sublevels is as follows:

- IL passages have one selective and two detail items.
- IM passages have one global and two detail items.
- AL passages have one global and two detail items.
- AM passages have one global, one detail, and one inference item.
- S passages have one global, one detail, and one inference item.

Passages and items align with each other with respect to function. Intermediate passages, e.g., may be understood sentence by sentence. Intermediate items consequently focus on information contained within the context of an individual sentence-length utterance. Advanced passages consist of descriptive and narrative texts that require paragraph-length comprehension and the understanding of cohesive devices to signal, e.g., sequencing, time frames, and chronology. Advanced items consequently focus on information that is distributed across several sentence-length utterances within a passage. Depending on the sublevel, item types, therefore, are defined differently as follows:

Global

- IL: Able to identify general subject matter, gets an idea of the content. The general subject matter is put in very broad terms. Distractors are viable passage-based options, i.e., there are words and phrases in the passage that refer plausibly to these options.

¹ The model by Weir & Khalifa was developed for reading comprehension. Because it works equally well for listening comprehension, it was used in the development of the LPT.

- IM: Able to identify general subject matter, understands the gist of the passage. The general subject matter is put in terms that require a global understanding of the passage at hand.
- AL: Ability to understand the main idea depends on comprehending supporting details. Examinee needs to understand some details to answer the question correctly. The correct answer needs to be synthesized from understanding different parts of the passage. The main idea is of a factual nature rather than focusing on author intent.
- AM: Ability to understand the main idea and/or argument depends on comprehending supporting details. The correct answer is spread out over different parts of the passage. It is based on what the author is intending to say. Author intent is clearly signaled.
- S: Fully able to understand the main argument and all supporting facts. It is the main argument the author is making. The correct answer is spread out over different parts of the passage. Distractors refer to other arguments the authors are making or to an argument they could be making based on statements contained in the passage.

Detail

- IL: Able to understand simple single facts. These facts are the easiest to understand and do not necessarily have to be important for the passage as a whole. Distractors must be viable passage-based options, which must be clearly false.
- IM: Able to understand single straightforward facts. These facts contribute to the gist of the passage. Still, their comprehension only requires understanding single simple sentence-length utterances. Distractors must be viable passage-based options. Key must use synonyms or paraphrases that consist of highly frequent or shared international vocabulary.
- AL: Able to understand explicitly mentioned facts and thoughts. They go beyond simple sentence-based facts. Their understanding is dependent on understanding the gist of the passage. They require understanding more than one sentence-length utterance. Distractors focus on other relevant facts mentioned. Key must use synonyms or paraphrases that contain general vocabulary.
- AM: Able to understand explicitly mentioned facts, thoughts, and argument. Their understanding is dependent on understanding the gist of the passage. They require understanding complete subsections of the passage rather than single sentence-length utterances. Keys and distractors focus on explicitly mentioned facts or argument. Key must use synonyms and paraphrases that contain a broad general vocabulary.
- S: Able to understand argument, finer points of detail and abstraction. They require understanding complete subsections of the passage rather than single sentence-length utterances. Keys and distractors focus on finer points of detail and abstraction that support the main argument of the passage. Key must use synonyms and paraphrases. Stem, key, and distractors often contain precise, specialized and low-frequency vocabulary.

Selective

- IL: Able to understand familiar words and very basic phrases. Both stem and options repeat words and phrases from the passage. The main task is to understand the question and to notice the answer in the passage. Both key and distractors need to contain language that is taken from the text.

Inference

- AM: Able to identify the main conclusions in clearly signaled explanatory or argumentative passages and to make straightforward inferences. Items refer to the complete passage and focus on something that is clearly understood but not explicitly mentioned.
- S: Able to infer attitude, mood, and intentions; able to infer implied as well as stated opinions; able to draw conclusions. Items refer to the complete passage, the main argument or subordinate arguments. They refer to something the speaker or speakers clearly had in mind, to their attitude towards the issue, or the reasons why they said what they said.

Item Difficulty

Items align with their level with respect to function, vocabulary, and grammar.

- IL: Most frequent common basic words and phrases, common names, cognates and shared international vocabulary; short, simple sentence-length utterances, predominantly in the present tense.
- IM: High-frequency words and phrases, cognates, and shared international vocabulary; short simple sentence-length utterances.
- AL: Variety of frequent words and phrases, cognates, and shared international vocabulary; longer and more complex turns containing some subordinate clauses, prepositional phrases and other features of connected discourse.
- AM: Broad active listening vocabulary and some low-frequency words and expressions; complex paragraph-length turns containing subordinate clauses, prepositional phrases and other features of connected discourse.
- S: Precise, often specialized and low-frequency vocabulary and expressions, including idioms and colloquialisms; complex paragraph-length turns, containing subordinate and prepositional clauses, gerunds and participial clauses referring to complex, abstract, and hypothetical argumentation and relationships.

9 Item Adequacy as a Domain Sample

The ACTFL LPT includes a broad spectrum of genres and topic categories to assure that the test adheres to its construct and consists of topics and language that are relevant for examinees. Each topic is used only once at any one sublevel to provide a representative sample of the language proficiency of examinees across a broad range of topics. Tables 7 and 8 below provide an example

of the genres and topics included in a test. Note that these are open lists that continue to be updated.

Table 7
Task Genres per Sublevel

IL	IM	AL	AM	S
Simple Announcements	Simple Announcements			
Simple Conversations	Simple Conversations			
Short routine telephone or online conversations	Short routine telephone or online conversations			
		Interviews	Interviews	Interviews
		News Items	News Items	News Items
		Narratives	Narratives	Narratives
		Oral Reports	Oral Reports	Oral Reports
			Opinion Pieces	Opinion Pieces
			Short Lectures	Short Lectures
			Debates	Debates
			Technical Discussions	Technical Discussions

Table 8
Task Topics and Subtopics

Topics	Subtopics
Arts	Age
Business & Commerce	Airport
Daily Life	Animals
Education	Brain
Family	Children
Fiction	Cinema
Food	College
Free time	Computer
Government and Politics	Directions
Health & Wellbeing	Drugs
Home	Environment
Law & Crime	Gender
Nature	History
News	Hobbies
Popular culture	Hospital

Science	Hotel
Society	Internet
Sports	Interview
Style	Languages
Technology	Literature
Travel	Living
Work	Love
	Math
	Meeting
	Money
	Moving
	Museum
	Music
	New Job
	People
	Pets
	Plans
	Plants
	Problems
	Recipe
	Religion
	Restaurant
	Routine
	School
	Shopping
	Souvenirs
	Theater
	Trade
	Tradition
	Traffic
	Train
	Transportation
	Trends
	Trips
	TV
	Weather

Subtopics may be subtopics of more than one main topic. Each exam provides a representative sample of the construct by including a broad spectrum of topics, subtopics, genres, and rhetorical organization (text type). The LPT is commonly taken as a two-sublevel test and consists of ten texts, five at each level. The ten texts are chosen to provide a representative statement of the language proficiency of the examinee. In the following, three different examples of two-level tests are presented to show how the texts reflect the *ACTFL Proficiency Guidelines 2012 – Listening and*

how the test ensures selecting a diverse and representative sample of the topics, subtopics, genres, and rhetorical organization of passages listeners are able to understand at each level.

Example 1 represents a test that spans the sublevels NL to IM. Passages and items are at the sublevels IL and IM. NH is defined as responding correctly to 50% of the Intermediate items, NM responding correctly to 40% of the items, and NL to less than 40%. Passage topics, subtopics, genres, and rhetorical organization are based on the ACTFL level descriptions as follows:

Intermediate Low

At the Intermediate Low sublevel, listeners are able to understand some information from sentence-length speech, one utterance at a time, in basic personal and social contexts, though comprehension is often uneven. At the Intermediate Low sublevel, listeners show little or no comprehension of oral texts typically understood by Advanced-level listeners.

Intermediate Mid

At the Intermediate Mid sublevel, listeners are able to understand simple, sentence-length speech, one utterance at a time, in a variety of basic personal and social contexts. Comprehension is most often accurate with highly familiar and predictable topics although a few misunderstandings may occur. Intermediate Mid listeners may get some meaning from oral texts typically understood by Advanced-level listeners.

Table 9 shows the variety and distribution of topics, subtopics, genres and rhetorical organization in a typical NL to IM test.

Table 9

Distribution of Topics, Subtopics, Genres, and Rhetorical Organization in a Typical NL to IM Test

Passage	Topic	Subtopic	Genre	Rhetorical Organization
IL.1	Free Time	Shopping	Announcement	Instruction
IL.2	Food	Restaurant	Simple Conversation	Description
IL.3	Family	People	Telephone Conversation	Description
IL.4	Daily Life	Pets	Simple Conversation	Instruction
IL.5	Arts	Theater	Announcement	Description
IM.1	Daily Life	Routine	Simple Conversation	Instruction
IM.2	Sports	Plans	Virtual Exchange	Description
IM.3	Daily Life	Moving	Simple Conversation	Narration
IM.4	Work	Routine	Simple Conversation	Narration
IM.5	Society	Literature	Announcement	Description
Distribution	3x Daily Life 1x Free Time	1x Shopping 1x Restaurant	3x Announcement 5x Simple Conversation	3x Instruction 4x Description

	1x Food 1x Family 1x Arts 1x Sports 1x Work 1x Society	1x People 1x Pets 1x Theater 2x Routine 1x Plans 1x Moving 1x Literature	1x Telephone Conversation 1x Virtual Exchange	3x Narration
--	---	--	--	--------------

Example 2 represents a test that spans the sublevels IM to AM. Passages and items are at the levels AL and AM. IH is defined as responding correctly to 50% of the Advanced items, and IM as responding correctly to 40% of the items. Responding to less than 40% of the items correctly is defined as Below Range (BR), i.e., as below the lowest sublevel the test is able to assess reliably. Passage topics, subtopics, genres, and rhetorical organization are based on the ACTFL level descriptions as follows:

Advanced Low

At the Advanced Low sublevel, listeners are able to understand short conventional narrative and descriptive passages with a clear underlying structure though their comprehension may be uneven. These passages predominantly contain high-frequency vocabulary and structures. Listeners understand the main ideas and some supporting details. Comprehension may often derive primarily from situational and subject-matter knowledge. Listeners at this level will be challenged to comprehend more complex passages.

Advanced Mid

At the Advanced Mid sublevel, listeners are able to understand conventional narrative and descriptive passages such as expanded descriptions of persons, places, and things and narrations about past, present, and future events. The speech is predominantly in familiar target-language patterns. Listeners understand the main facts and many supporting details. Comprehension derives not only from situational and subject-matter knowledge, but also from an increasing overall facility with the language itself. Listeners at this level may derive some meaning from passages that are structurally and/or conceptually more complex.

Table 10 shows the variety and distribution of topics, subtopics, genres and rhetorical organization in a typical IM to AM test.

Table 10

Distribution of Topics, Subtopics, Genres, and Rhetorical Organization in a Typical IM to AM Test

Task	Topic	Subtopic	Genre	Rhetorical Organization
AL.1	Society	Trends	Simple Story	Narration
AL.2	Daily Life	People	Interview	Description

AL.3	Work	Children	Simple Story	Narration
AL.4	Travel	Money	News Item	Explanation
AL.5	Travel	Trips	Oral Report	Description
AM.1	Society	People	News Item	Narration
AM.2	Education	School	Story	Narration
AM.3	Government/Politics	Plans	Short Lecture	Explanation
AM.4	Arts	Cinema	Interview	Explanation
AM.5	Society	Tradition	Interview	Explanation
Distribu- tion	3x Society 1x Daily Life 1x Work 2x Travel 1x Education 1x Government and politics 1x Arts	1x Trends 2x People 1x Children 1x Money 1x Trips 1x School 1x Plans 1x Cinema 1x Tradition	3x Story 3x Interview 2x News Items 1x Oral Report 1x Short Lecture	4x Explanation 4x Narration 2x Description

Example 3 represents a test that spans the sublevels IH to S. Passages and items are at the levels AM and S. AL is defined as responding correctly to 50% of the AM and S items, and IH as responding correctly to 40% of the items. Responding to less than 40% of the items correctly is defined as Below Range (BR), i.e., as below the lowest sublevel the test is able to assess reliably. Passage topics, subtopics, genres and rhetorical organization are based on the ACTFL level descriptions as follows:

Advanced Mid

At the Advanced Mid sublevel, listeners are able to understand conventional narrative and descriptive passages, such as expanded descriptions of persons, places, and things and narrations about past, present, and future events. The speech is predominantly in familiar target-language patterns. Listeners understand the main facts and many supporting details. Comprehension derives not only from situational and subject-matter knowledge, but also from an increasing overall facility with the language itself. Listeners at this level may derive some meaning from passages that are structurally and/or conceptually more complex.

Superior

At the Superior level, listeners are able to understand speech in a standard dialect on a wide range of familiar and less familiar topics. They can follow linguistically complex extended discourse such as that found in academic and professional settings, lectures, speeches, and oral reports. Comprehension is no longer limited to the listener's familiarity with subject matter, but also comes from a command of the language that is supported by a broad vocabulary, an understanding of more complex structures, and linguistic experience within the target culture. Listeners at the Superior level can understand not only what is said, but sometimes what is left unsaid; that is, they can make inferences.

Superior-level listeners understand texts that use precise, often specialized vocabulary and complex grammatical structures. These texts feature argumentation, supported opinion, and hypothesis, and use abstract linguistic formulations as encountered in academic and professional listening. Such texts are typically reasoned and/or analytic and may frequently contain cultural references.

Superior-level listeners understand speech that typically uses precise, specialized vocabulary and complex grammatical structures. This speech often deals abstractly with topics in a way that is appropriate for academic and professional audiences. It can be reasoned and can contain cultural references.

Table 11 shows the variety and distribution of topics, subtopics, genres and rhetorical organization in a typical IH to S test.

Table 11
Distribution of Topics, Subtopics, Genres, and Rhetorical Organization in a typical IH to S test

Task	Topic	Subtopic	Genre	Rhetorical Organization
AM.1	Society	People	News Item	Narration
AM.2	Education	School	Story	Narration
AM.3	Government and Politics	Plans	Short Lecture	Narration
AM.4	Arts	Cinema	Interview	Explanation
AM.5	Society	Tradition	Interview	Explanation
S.1	Business & Commerce	Money	Debate	Hypothesis
S.2	Government and Politics	Reform	Short Lecture	Argument
S.3	Food	Trends	News Item	Narration
S.4	Technology	Reform	Opinion Piece	Hypothesis
S.5	Science	Problems	Debate	Argument
Distribution	2x Society 1x Education 2x Government/Politics 1x Arts 1x Business & Commerce 1x Food 1x Technology 1x Science	1x People 1x School 1x Plans 1x Cinema 1x Tradition 1x Money 2x Reform 1x Trends 1x Problems	2x News Items 1x Story 2x Short Lecture 2x Interview 2x Debate 1x Opinion Piece	4x Narration 2x Explanation 2x Hypothesis 2x Argument

As these examples show, the tasks in any single exam cover a broad spectrum of topics, subtopics, genres, and rhetorical organization to provide a solid and representative statement of the listening proficiency of examinees. The qualitative and quantitative analyses of the three representative test ranges also provides evidence that the test items represent the domains of knowledge and abilities the test claims it does well. We will refer to this section again in the section on content validity below (Section 18).

10 Item Currency and Representativeness

The *ACTFL Proficiency Guidelines 2012 – Listening* represent the current state of knowledge of second language listening proficiency at various levels of proficiency. Section 7 showed that the number and distribution of topics, subtopics, genres, and rhetorical organization are representative of the proficiency levels identified by the *ACTFL Proficiency Guidelines 2012 – Listening*, while Section 6 showed the same for the items (listening goals), completing the categories used by ACTFL in its level descriptions.

11 Item Sensitivity Panel Review Description

The main sensitivity concerns in second language testing include the kinds of topics and the language used. Neither topics nor language should be offensive toward any examinees. Item writers are instructed to avoid topics such as drugs, sexuality, war, violence, etc. that may engender strong emotional reactions as well as discriminating and linguistically inappropriate content to ensure equal access to the passages for all examinees. It is neither economically feasible nor, indeed, necessary to use panels instead of individual reviewers to review the appropriateness of topics, the situations described, the arguments provided, and the language used in world languages tests, if item sensitivity review is included in all phases of the item development process as it is in the development of ACTFL LPT passages and items, i.e., the writing, revisions, and quality assurance phases (UAT). In addition, an item sensitivity review is part of the item revision process after IRT analyses have been completed, which again results in multiple stages of review involving three or more individuals trained to spot non-inclusive content and language. Within the life cycle of a test form, item sensitivity review is part of the following stages:

1. Test writers are instructed to ensure that the content and language of all passages and items are appropriate.
2. Item reviewers are instructed to flag inappropriate content and/or language. There are two reviews completed by two different reviewers, one focusing on content and style and the other one focusing on level appropriateness and item quality.
3. Before and during UAT 1 (User Assurance Testing), quality assurance reviewers are instructed to flag inappropriate content and/or language. UAT is commonly completed at least twice.
4. In the revision cycle after the IRT analysis, test writers revising flagged items are instructed to listen to all passages and items, not only flagged ones, to ensure the appropriateness of content and language, in addition to flagging outdated content.
5. Revised texts and/or items are reviewed by item specialists who also inspect the appropriateness of the content and the language of the revised passages or items.
6. Before and during UAT 2, quality assurance reviewers are again instructed to flag any inappropriate content or language.

During the first cycle (steps 1-3), therefore, the appropriateness of content and language is checked by four different people, and during the second cycle (steps 4-6) by an additional three different people for a total of seven different people altogether. Great care is taken to ensure an equal distribution of male and female item writers, item reviewers, and quality assurance reviewers as well as of people of various ethnic, social, and regional backgrounds and sexual orientation.

12 Field/Pre-Test Processes and Procedures

All forms go through a rigorous development process (see Item Development Process in Section 3). There is no pretesting. IRT analysis are performed, generally, after approx. 300-400 test administrations, after which a number of items are, generally, revised. To date, all reports have found that each released form showed good psychometric properties with high overall Rasch separation reliability to meet the requirements of a high-stakes test (see Section 13 below).

13 Item Analysis Results

This section presents the item analysis results (e.g., item difficulty, discrimination, correlation with external criteria) (see Appendix 6 – Technical Report and Appendix 7 – Arabic, German, and Spanish LPT Data Reports).

Data reports are completed for all test forms, generally, after 300-400 test administrations when sufficient numbers of examinees have taken each individual sublevel (IL, IM, AL, AM, S). Data reports provide the date on which the report was completed, the name of the test, e.g., Spanish LPT 01, the name of the person completing the report, the date or dates of data collection, and the number of participants. The data are analyzed using item response theory (IRT). The IRT model used is the Rasch model for dichotomous items.

For each item, the data reports provide the number of cases; the item difficulty (measure) reported in logits; the standard error of the mean (SEM) also reported in logits; infit and outfit statistics; and the separation index (point-biserial item-scale correlations) to indicate how well the item discriminates between examinees at various proficiency levels. A comment column completes the item table. In addition, the data reports provide the overall separation reliability and overall model fit, and make recommendations with respect to item difficulty, separation, overall reliability, and construct validity. They also list the nine anchor items and indicate from which form they were derived and their ID numbers. Each report concludes with a general statement as to the quality of the psychometric properties of the test and its usability for high stakes testing.

The item difficulty (measure) of an item expressed in logits should fall within a particular range for each sublevel. These ranges vary from language to language and are determined on the basis of all forms of a particular language. If the item difficulty falls outside one standard deviation of

the sublevel mean but stays within one standard deviation of the adjoining sublevel mean, the item is flagged for inspection (yellow). If it also falls outside one standard deviation of the adjoining sublevel, it is flagged for revision (red).

Fit statistics indicate the degree to which a test item meets the Rasch model expectations. Fit values between 0.5 and 1.5 mean-squares are the most productive values for measurement. Fit values between 1.5 and 2.0 mean-squares are unproductive but not degrading. Fit values larger than 2.0 mean-squares indicate too much variance, degrading the measurement. Whereas infit statistics are sensitive to the competence range for which the test was designed, outfit statistics are sensitive to outliers. Traditionally, infit statistics are considered more important than outfit statistics. Items with infit values above 2.0 are recommended for revision and flagged red. Items with outfit values above 2.0 are recommended for inspection and flagged yellow.

Separation indices should not fall below 0.20. Unlike the Rasch item difficulty estimates, item-scale correlations are sample-dependent. Sampling errors such as participants being more or less proficient than usual, affect the item discrimination parameter. Items with separation indices between 15 and 19 are flagged for inspection (yellow), while items with separation indices below 15 are flagged for revision (red).

The comment column spells out the action recommended (inspection or revision) and the main reason(s) such as inappropriate item difficulty, infit statistic, or separation index.

The overall separation reliability should not be lower than 0.8, and the overall model fit statistics should ideally be between 0.5 and 1.5 but values below 0.5 and between 1.5 and 2.0 are also acceptable. Good overall and item infit statistics, moreover, provide evidence of construct validity because they indicate that the test form measures the proficiency range for which it was designed (see also Section 20).

Results

The item difficulty and discrimination parameters for the LPT are presented for the three selected languages, i.e., Arabic, German, and Spanish. The results of the most recent forms are included below, i.e., Arabic LPT 02, German LPT 02 and Spanish LPT 05 (see Appendix 7 – Arabic, German, and Spanish LPT Data Reports for all Arabic, German, and Spanish forms)².

The item difficulty measure is reported in logits as estimated by the Rasch model for dichotomous items (see Tables 12-14). Probabilistic test theory (Rasch model) yields information that is sample-independent and expresses item difficulty across all proficiency levels on the same metric. The standard error of measurement (SEM) of the difficulty estimate is also reported in logits. Please note that these difficulty parameters cannot be compared directly across languages.

² As of the time of this writing, Arabic LPT 03 and Spanish LPT 06 did not have sufficient test results at the Superior level, while German LPT 03 and Spanish LPT 07 and 08 did not have sufficient test results at any level.

Tables 12-14 show a variety of measures for all of the items in the test. The items are listed in rows. They are coded by level, task, and item. A1 indicates IL, A2 indicates IM, B1 indicates AL, B2 indicates AM, and C1 indicates Superior. The first digit after the sublevel indicates the listening passage, i.e., passages 1 through 5, and the second digit after the sublevel indicates the item, i.e., items 1 through 3. Thus, A1.1.1 indicates IL listening passage 1 item 1.

Column 2 provides the number of examinees (N) responding to a particular item; column 3 provides the item difficulty (measure) in logits; and column 4 the standard error of measurement (SEM), also expressed in logits. Columns 5 and 6 provide the Rasch infit and outfit values in mean-squares (MNSQ). Column 7 provides the separation index (item discrimination) expressed as a point-biserial correlation (r_{pb}); and Column 8 provides the action recommended together with the main reason(s). For each language, the mean difficulty logic of all items was set to 0.

Conspicuous items requiring action are flagged. A yellow flag means that the item needs to be inspected, and revised if needed, while a red flag means that the item needs to be revised. Items with difficulty measures one standard deviation (SD) below or above the mean of the sublevel are flagged yellow when the measure falls within one SD range of an adjoining sublevel and red when the measure falls outside one SD range of the adjoining sublevel. Infit values above 2.0 MNSQ are flagged red and outfit values above 2.0 MNSQ are flagged yellow. Separation indices below 0.15 are flagged red and indices between 0.15 and 0.19 are flagged yellow. Table 12 provides the item characteristics for Arabic LPT 02.

Table 12
Item Characteristics Arabic LPT 02

Item	Number of Cases	Measure	SEM	Infit	Outfit	Separation Index	Comment
A1.1.1	85	-4.68	.62	1.14	.53	.22	
A1.1.2	85	-2.36	.31	1.05	1.69	.37	
A1.1.3	85	-2.36	.31	.79	.53	.57	
A1.2.1	85	-2.90	.35	.97	.62	.44	
A1.2.2	85	-2.01	.29	.88	.83	.53	
A1.2.3	85	-2.18	.30	.83	.73	.55	
A1.3.1	85	-2.27	.30	.87	.62	.54	
A1.3.2	85	-2.09	.29	1.07	1.38	.39	
A1.3.3	85	-2.56	.32	.97	1.22	.42	
A1.4.1	85	-2.01	.29	.86	.73	.55	
A1.4.2	85	-2.46	.31	.87	1.52	.48	
A1.4.3	85	-2.09	.29	.90	.68	.53	
A1.5.1	85	-2.09	.29	1.30	3.75	.19	Separation index below threshold. Inspect.
A1.5.2	85	-2.67	.33	.64	.41	.61	
A1.5.3	83	-4.20	.55	1.01	.32	.33	
A2.1.1	136	-1.95	.24	.91	.79	.41	
A2.1.2	136	-1.38	.21	.96	.89	.41	
A2.1.3	135	-.54	.19	.81	.74	.58	

A2.2.1	136	-1.72	.23	1.04	.94	.32	
A2.2.2	136	-.99	.20	.76	.66	.61	
A2.2.3	133	-2.03	.25	.77	.52	.55	
A2.3.1	136	-.71A	.19	.79	.85	.53	
A2.3.2	135	-.60A	.19	.95	.93	.34	
A2.3.3	136	-.28A	.19	1.00	1.01	.36	
A2.4.1	136	-1.89	.24	.91	.93	.40	
A2.4.2	136	-1.72	.23	.93	1.17	.38	
A2.4.3	136	-.91	.20	.87	.77	.52	
A2.5.1	136	-.61	.19	.90	.89	.49	
A2.5.2	136	.59	.19	1.12	1.39	.25	
A2.5.3	138	-1.10	.20	.94	.92	.45	
B1.1.1	234	-.77A	.17	.98	.90	.54	
B1.1.2	234	-.59A	.16	1.00	1.02	.48	
B1.1.3	232	-1.22A	.18	1.01	.92	.47	
B1.2.1	234	-2.18	.23	1.18	1.84	.25	Item too easy. Revise.
B1.2.2	233	.17	.15	.97	.95	.51	
B1.2.3	233	.29	.15	.95	1.00	.51	
B1.3.1	232	-1.67	.20	1.07	1.45	.38	Item too easy. Revise.
B1.3.2	232	-.95	.17	.88	.86	.56	
B1.3.3	231	-.88	.17	.93	.79	.55	
B1.4.1	233	-.44	.16	.91	.83	.56	
B1.4.2	233	-.01	.15	1.08	1.24	.43	
B1.4.3	231	-.48	.16	.89	.80	.56	
B1.5.1	233	-1.68	.20	.93	.96	.48	Item too easy. Revise.
B1.5.2	232	-.12	.16	1.14	1.21	.40	
B1.5.3	228	1.11	.15	1.31	1.66	.21	
B2.1.1	182	.08	.18	.92	.85	.47	
B2.1.2	181	1.78	.17	1.05	1.35	.29	
B2.1.3	178	.20	.18	.97	.96	.41	
B2.2.1	180	1.87A	.17	.88	.80	.42	
B2.2.2	181	2.63A	.19	.79	.87	.03	Item too difficult. Revise.
B2.2.3	181	.74A	.17	.99	.97	.38	
B2.3.1	182	-.64	.21	.82	.62	.54	
B2.3.2	182	1.29	.16	1.00	1.02	.39	
B2.3.3	182	1.23	.16	1.01	1.12	.36	
B2.4.1	182	-.88	.23	.94	.74	.42	
B2.4.2	182	-.09	.19	.79	.66	.59	
B2.4.3	181	1.12	.16	1.03	1.03	.37	
B2.5.1	182	-1.04	.24	.90	.66	.44	
B2.5.2	182	-.23	.19	.88	.77	.49	
B2.5.3	182	-.38	.20	.78	.62	.58	
C1.1.1	117	.59	.22	.94	.90	.51	
C1.1.2	117	1.95	.21	1.12	1.53	.27	
C1.1.3	117	1.56	.21	1.24	1.66	.19	Separation index below threshold. Inspect.
C1.2.1	118	2.81	.23	.94	1.69	.35	
C1.2.2	118	1.29	.21	1.10	1.20	.34	
C1.2.3	117	.54	.22	.81	.74	.61	

C1.3.1	118	-.73	.28	.81	.75	.55	Item too easy. Revise.
C1.3.2	118	.72	.22	1.14	1.10	.35	
C1.3.3	117	.96	.21	1.04	1.12	.41	
C1.4.1	117	1.90	.21	1.19	1.43	.23	
C1.4.2	117	.83	.21	1.14	1.16	.34	
C1.4.3	117	-.29	.26	.99	.78	.47	
C1.5.1	118	-.90	.30	.75	.43	.61	Item too easy. Revise.
C1.5.2	118	1.37	.21	1.09	1.26	.34	
C1.5.3	118	.28	.23	1.01	.96	.45	

Table 12 shows that the overall item difficulty increases with the sublevels tested as expected. All infit values were between 0.5 and 1.5 and many of them were close to 1.0, indicating that the items fit the model well. All but one outfit values were below 2.0. One item had an outfit value above 2.0, indicating the presence of outliers. This item was flagged for inspection, because it also had a separation index slightly below the threshold.

Table 12 shows that a total of 6 out of 75 items were flagged for revision, either because they were too difficult, too easy, or because they had a separation value below 0.15: 3 AL, 1 AM, and 2 Superior items. One of them was an anchor item, which was replaced for subsequent forms. Additionally, 2 items were flagged for inspection: 1 IL and 1 S item. As a result, a total of 7 items were revised.

Table 13 shows the item characteristics for German LPT 02.

Table 13
Item Characteristics German LPT 02

Item	Number of Cases	Measure	SEM	Infit	Outfit	Separation Index	Comment
A1.1.1	396	-3.47	.19	1.06	1.11	.14	Separation index below threshold. Revise.
A1.1.2	396	-.70	.11	1.05	1.05	.34	
A1.1.3	396	-.99	.11	1.03	1.04	.35	
A1.2.1	396	-2.82	.15	.98	.89	.28	
A1.2.2	396	-2.60	.14	1.01	1.04	.26	
A1.2.3	396	-.01	.11	1.05	1.10	.35	
A1.3.1	396	-3.20	.17	1.08	1.32	.11	Separation index below threshold. Revise.
A1.3.2	396	-1.75	.12	1.12	1.24	.22	
A1.3.3	396	-1.98	.12	.94	.87	.38	
A1.4.1	396	-2.97	.16	.95	.90	.29	
A1.4.2	396	-3.11	.17	1.02	1.06	.21	
A1.4.3	396	-2.58	.14	.90	.86	.36	
A1.5.1	396	-3.33	.18	1.03	.94	.19	Separation index below threshold. Inspect.
A1.5.2	396	-4.34	.28	.97	.66	.19	Item too easy. Revise.
A1.5.3	396	-5.07	.38	.98	.50	.15	Item too easy. Revise.
A2.1.1	522	.74	.11	1.18	1.31	.26	

A2.1.2	522	-2.35	.13	.86	.70	.43	
A2.1.3	522	-1.11	.10	.81	.74	.56	
A2.2.1	522	-1.08A	.10	.80	.72	.52	
A2.2.2	522	-.73A	.10	.89	.85	.49	
A2.2.3	522	-1.87A	.11	.84	.75	.43	
A2.3.1	522	-.79	.10	1.01	.98	.40	
A2.3.2	522	-1.87	.11	1.00	1.18	.33	
A2.3.3	522	-.70	.10	.85	.79	.55	
A2.4.1	522	-1.51	.11	1.02	.98	.36	
A2.4.2	522	-3.86	.21	1.01	1.08	.15	Item too easy. Revise.
A2.4.3	522	-.82	.10	.76	.69	.61	
A2.5.1	522	-1.72	.11	.96	1.02	.38	
A2.5.2	522	1.29	.12	1.15	1.39	.25	Item too difficult. Revise.
A2.5.3	522	-1.29	.10	1.19	1.30	.22	
B1.1.1	870	-.43	.09	.95	.75	.48	
B1.1.2	871	.25	.08	.99	.96	.46	
B1.1.3	868	.31	.08	.89	.82	.54	
B1.2.1	870	.77	.08	.98	1.04	.46	
B1.2.2	869	.91	.08	.94	.99	.50	
B1.2.3	868	-.56	.10	.98	.85	.44	
B1.3.1	869	-.59	.10	.84	.68	.54	
B1.3.2	869	-.39	.09	.85	.68	.55	
B1.3.3	868	-.30	.09	.88	.79	.52	
B1.4.1	870	-.28A	.09	.94	.81	.37	
B1.4.2	870	.71A	.08	.95	.95	.51	
B1.4.3	864	1.24A	.08	.86	.88	.56	
B1.5.1	872	.66	.08	.95	.91	.50	
B1.5.2	870	.32	.08	.87	.90	.55	
B1.5.3	871	.10	.08	.89	.82	.53	
B2.1.1	652	.85	.09	1.11	1.16	.25	
B2.1.2	652	.76	.09	.80	.71	.58	
B2.1.3	652	2.30	.09	.96	1.01	.39	
B2.2.1	652	2.52	.09	1.15	1.43	.15	Separation index below threshold. Inspect.
B2.2.2	652	2.81	.09	1.24	1.42	.07	Separation index below threshold. Revise.
B2.2.3	651	2.97	.10	.99	1.09	.32	Item too difficult. Revise.
B2.3.1	652	3.20	.10	1.07	1.26	.20	Item too difficult. Revise.
B2.3.2	650	.65	.09	.99	.99	.38	
B2.3.3	652	.36	.10	1.10	1.16	.24	
B2.4.1	652	1.30A	.09	1.02	1.03	.36	
B2.4.2	651	1.34A	.09	1.10	1.09	.35	
B2.4.3	651	.30A	.10	1.01	.94	.42	
B2.5.1	652	1.99	.09	.98	.95	.40	
B2.5.2	652	.39	.10	1.01	.98	.34	
B2.5.3	652	1.27	.09	.88	.85	.51	
C1.1.1	608	.69	.09	1.01	.98	.36	
C1.1.2	608	1.14	.09	1.01	1.02	.36	
C1.1.3	604	3.08	.10	.93	.95	.38	
C1.2.1	608	2.37	.09	1.17	1.27	.18	Separation index below threshold. Inspect.

C1.2.2	607	3.15	.10	1.04	1.31	.24	
C1.2.3	608	1.09	.09	1.00	1.01	.37	
C1.3.1	609	2.40	.09	1.18	1.38	.14	Separation index below threshold. Revise.
C1.3.2	608	2.34	.09	1.00	1.13	.34	
C1.3.3	609	2.21	.09	.96	.99	.40	
C1.4.1	609	.28	.10	.99	.98	.35	Item may be too easy. Inspect.
C1.4.2	610	.96	.09	1.06	1.05	.31	
C1.4.3	607	2.21	.09	1.15	1.26	.20	
C1.5.1	608	2.08	.09	1.13	1.18	.23	
C1.5.2	607	1.59	.09	.98	.98	.40	
C1.5.3	600	2.53	.09	1.03	1.09	.31	

Table 13 shows that the overall item difficulty increases with the sublevels tested as expected. All infit values were between 0.5 and 1.5, while many of them were close to 1.0, indicating that the items fit the model well. All outfit values also ranged between 0.5 and 1.5.

Table 13 shows that a total of 10 out of 75 items were flagged for revision, either because they were too difficult, too easy, or because they had a separation index below 0.15: 4 IL, 2 IM, 2 AM, and 3 Superior items. Poor separation values often coincided with poor difficulty values. Additionally, 4 items were flagged for inspection: 2 IL, 1 AM, and 2 Superior items. A total of 12 items were revised during the German LPT 02 revision process.

Table 14 shows the item characteristics for Spanish LPT 05.

Table 14
Item Characteristics Spanish LPT 05

Item	Number of Cases	Measure	SEM	Infit	Outfit	Separation Index	Comment
A1.1.1	75	-2.51	.51	1.02	1.23	.33	
A1.1.2	76	.79	.27	1.13	1.24	.43	Item may be too difficult. Inspect.
A1.1.3	76	-1.56	.39	1.39	2.25	.20	
A1.2.1	77	-.10	.30	1.31	1.33	.36	
A1.2.2	77	-.58	.32	.94	.86	.57	
A1.2.3	77	-3.62	.74	.93	.20	.32	
A1.3.1	77	-2.30	.47	.75	.25	.55	
A1.3.2	77	-3.16	.62	.78	.16	.44	
A1.3.3	77	-.38	.31	1.33	1.42	.34	
A1.4.1	77	-1.02	.35	.71	.50	.68	
A1.4.2	77	-1.28	.37	1.07	1.13	.44	
A1.4.3	77	-1.15	.36	1.05	.99	.48	
A1.5.1	77	-.58	.32	.93	.82	.57	
A1.5.2	77	.91	.27	1.39	1.86	.22	Item may be too difficult. Inspect.
A1.5.3	77	-1.90	.42	1.10	.95	.39	

A2.1.1	111	1.23	.21	1.33	1.74	.12	Separation index below threshold. Revise.
A2.1.2	112	-.31	.24	.94	.87	.46	
A2.1.3	112	-.37	.24	1.03	1.10	.38	
A2.2.1	114	.54A	.21	.94	.89	.46	
A2.2.2	112	-.34A	.24	1.27	1.42	.34	
A2.2.3	112	.01A	.23	.62	.51	.54	
A2.3.1	114	-.22	.24	1.05	1.07	.37	
A2.3.2	114	-.34	.24	1.09	1.28	.31	
A2.3.3	113	-2.14	.41	.78	.37	.45	Item too easy. Revise.
A2.4.1	113	-2.77	.52	.91	.45	.30	Item too easy. Revise.
A2.4.2	113	-.35	.24	1.09	1.03	.33	
A2.4.3	112	-.78	.27	.93	1.25	.40	
A2.5.1	114	1.01	.21	1.05	1.12	.40	Item may be too difficult. Inspect.
A2.5.2	114	.38	.22	1.17	1.30	.28	
A2.5.3	113	-1.07	.29	.97	.86	.41	
B1.1.1	325	-2.30	.29	.77	.72	.42	Item too easy. Revise.
B1.1.2	335	-.84	.17	.92	.98	.41	
B1.1.3	336	-1.72	.23	.90	.62	.41	Item too easy. Revise.
B1.2.1	335	-1.94	.25	.78	.44	.47	Item too easy. Revise.
B1.2.2	336	-1.83	.24	.93	.71	.37	Item too easy. Revise.
B1.2.3	336	-.53	.16	.90	.70	.48	
B1.3.1	335	.51A	.13	1.16	1.24	.42	
B1.3.2	333	1.03A	.12	.93	.99	.45	
B1.3.3	332	.53A	.13	.84	.79	.43	
B1.4.1	336	-1.72	.23	.96	.70	.36	Item too easy. Revise.
B1.4.2	335	-1.57	.22	.77	.48	.52	Item may be too easy. Inspect.
B1.4.3	334	-.43	.15	.95	.77	.45	
B1.5.1	334	.71	.12	1.20	1.34	.21	
B1.5.2	334	-1.43	.21	.98	.71	.38	Item may be too easy. Inspect.
B1.5.3	333	-2.06	.26	.80	.60	.44	Item too easy. Revise.
B2.1.1	326	2.09	.12	.96	1.17	.38	
B2.1.2	323	.17	.14	.95	.93	.40	
B2.1.3	323	.89	.12	1.14	1.19	.22	
B2.2.1	324	-1.10	.20	.82	.71	.46	Item may be too easy. Inspect.
B2.2.2	324	1.35	.12	.93	.96	.44	
B2.2.3	323	.74	.13	.96	1.02	.39	
B2.3.1	324	.53	.13	1.07	1.11	.28	
B2.3.2	323	.41	.13	.99	.97	.37	
B2.3.3	323	-1.39	.22	.92	.93	.32	Item too easy. Revise.
B2.4.1	323	.69A	.13	1.34	1.47	.25	
B2.4.2	323	1.55A	.12	1.02	1.00	.34	
B2.4.3	323	.49A	.13	.72	.65	.37	
B2.5.1	324	1.37	.12	.92	.89	.45	
B2.5.2	322	.98	.12	.92	.89	.45	
B2.5.3	322	.28	.13	.85	.77	.51	
C1.1.1	193	1.84	.16	1.08	1.41	.27	
C1.1.2	182	1.51	.16	.93	.91	.45	
C1.1.3	179	.79	.17	1.05	1.09	.33	

C1.2.1	182	1.46	.16	1.08	1.33	.28	
C1.2.2	180	3.85	.24	.94	2.21	.25	Item may be too difficult. Inspect.
C1.2.3	179	3.63	.22	1.17	3.00	-.04	Separation index below threshold. Revise.
C1.3.1	182	.83	.17	1.04	1.03	.36	
C1.3.2	182	2.24	.16	1.16	1.72	.12	Separation index below threshold. Revise.
C1.3.3	178	-.36	.22	1.05	1.05	.31	Item may be too easy. Inspect.
C1.4.1	180	1.52	.16	1.05	1.05	.34	
C1.4.2	182	-.31	.22	.83	.67	.55	Item may be too easy. Inspect.
C1.4.3	181	.94	.17	1.02	1.04	.38	
C1.5.1	182	2.66	.17	.97	1.38	.33	
C1.5.2	181	1.78	.16	.96	1.11	.40	
C1.5.3	181	1.72	.16	.96	.93	.42	

Table 14 shows that the overall item difficulty increases with the sublevels tested as expected. All infit values were between 0.5 and 1.5 and many of them were close to 1.0, indicating that the items fit the model well. Three outfit values were above 2.0, indicating the presence of outliers. Only one of them was flagged for revision, because it also had a separation index below the threshold.

Table 14 shows that a total of 15 out of 75 items were flagged for revision, either because they were too difficult, too easy, or because they had a separation index below 0.15: 3 IM, 6 AL, 1 AM, and 2 Superior items. Additionally, 11 items were flagged for inspection: 3 IL, 2 IM, 2 AL, 1 AM, and 3 Superior items.

All 15 items flagged for revision will be revised in the upcoming Spanish LPT 05 revision process. A number of additional items will most likely be revised as well.

14 Internal Consistency Reliability

To measure the internal consistency of the five sublevels of each Arabic, German, and Spanish form, Cronbach's Alpha was computed for all examinees who took the complete test, i.e., who completed all five sublevels (Version H). Cronbach's Alpha provides an overall reliability estimate and is considered to be a measure of scale reliability. A value above 0.8 suggests that the items have high internal consistency. Table 15 shows Cronbach's Alpha for the examinees who took all five sublevels.

Table 15
Scale Reliability of all Arabic, German, and Spanish LPTs

Language	<i>N</i>	Cronbach's Alpha
Arabic LPT 01	219	0.863*
Arabic LPT 02	42	0.870*
German LPT 01	21	0.947*
German LPT 02	15	0.909*
Spanish LPT 01	140	0.891*
Spanish LPT 02	14	0.862*
Spanish LPT 03	335	0.877*
Spanish LPT 04	123	0.882*
Spanish LPT 05	50	0.883*

* $p < 0.5$

Table 15 shows that Cronbach's Alpha was above 0.8 for all forms of all three languages, indicating high internal consistency of the items.

This conclusion is corroborated by the overall Rasch item fit statistics for all forms of the three languages in Table 16 (see Section 13 for item fit statistics for individual items).

Table 16
Overall Rasch Fit Statistics

	<i>N</i>	Rasch Item Infit (MNSQ)	Rasch Item Outfit (MNSQ)
Arabic LPT 01	545	0.99	1.04
Arabic LPT 02	280	0.96	1.00
German LPT 01	443	1.00	1.06
German LPT 02	1,156	0.99	0.99
Spanish LPT 01	1,658	1.01	1.07
Spanish LPT 02	1,038	1.00	1.05
Spanish LPT 03	1,223	1.03	1.08
Spanish LPT 04	628	0.99	1.02
Spanish LPT 05	338	0.99	1.02

Table 16 shows that the items generally produce exactly the same amount of infit variance that would be expected from the Rasch model. Outfit values are equally close to the ideal variance range. The overall Rasch fit statistics, thus, add another piece of evidence to support the conclusion that the measurement functions as desired.

15 Equivalence of Exam Forms Evidence

There are several measures in place to ensure equivalence of test forms: the training and monitoring of item writers and reviewers; the use of anchor items; and the revision of test forms on the basis of the IRT analysis.

Item writers and reviewers are rigorously trained and monitored throughout the text and item writing process (see *Item Development Process* in Section 3). They are provided with a very detailed Item Writing Manual and Item Checklists (see Appendix 8 – Item Writing Manual and Appendix 9 – Item Checklists). The same item writers and/or reviewers are commonly involved in several test forms. Because the texts and items are reviewed and revised at least twice and because there are at least three experienced item writers and reviewers involved in every single test form, there is a precise and deeply shared understanding of what the ACTFL levels and sublevels involve.

Three anchor texts and nine anchor items of one form are used for each subsequent form, i.e., three anchor items at IM, three at AL, and three at AM. These anchor items are carefully selected on the basis of the IRT analysis and exhibit the best difficulty measures and separation indices of that particular form. By means of common item equating using the WINSTEPS software, the difficulty of new test items is determined with high precision.

IRT analyses are completed for all forms generally after 300-500 test administrations. Items with conspicuous values are inspected and revised, if necessary (see Section 13). This is a mandatory part of the item development cycle. Revised forms become part of the form pool and will be inspected and revised on the basis of additional IRT analyses further down the road. These revisions ensure even greater form equivalency.

Figures 1, 2, and 3 show logit boxplots of the two Arabic, two German, and five Spanish forms available at present. Note that these boxplots represent forms that had not yet been revised. (All but Spanish LPT 05 were revised and replaced the previous versions as of this writing.)

Figure 1
Logit Distributions of Arabic LPT 01 and LPT 02

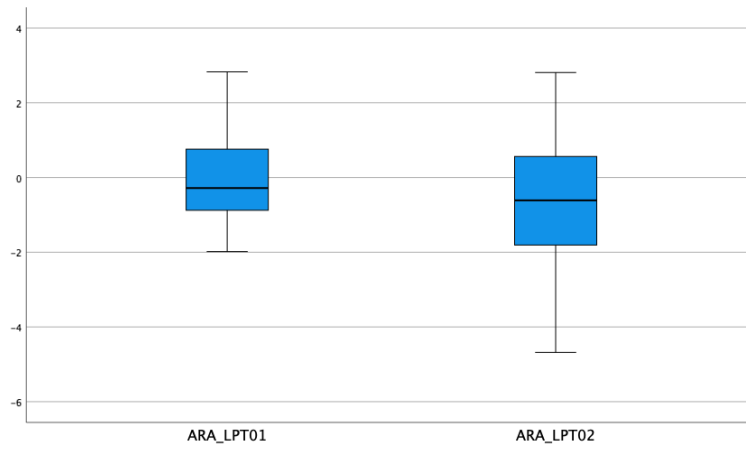


Figure 2
Logit Distributions of German LPT 01 and LPT 02

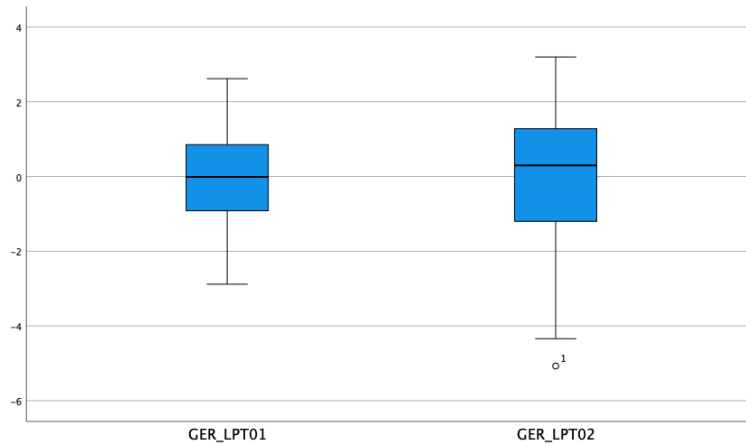
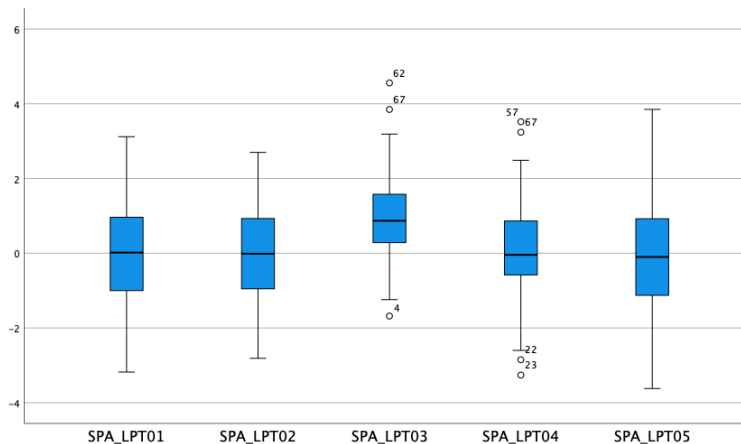


Figure 3
Logit Distributions of Spanish LPT 01 through LPT 07



Figures 1, 2, and 3 show very similar distributions for all forms of each language except for Spanish LPT 03. The medians were similar and the interquartile ranges (IRQ) (boxes) and full ranges (whiskers) were also quite similar even before the first mandatory revision, i.e., on the basis of the quality of the item development process alone.

Table 17 shows the number of items, logit medians, means, standard errors of the mean (SEM), and standard deviations of all seven test forms.

Table 17
Descriptive Statistics of two Arabic, two German, and five Spanish LPT forms

	N	Median	Mean	SEM	SD
Arabic LPT 01	75	-0.28	0.00	0.13	1.11
Arabic LPT 02	75	-0.61	-0.58	0.18	1.52
German LPT 01	75	-0.01	0.00	0.14	1.23
German LPT 02	75	0.30	-0.04	0.23	1.97
Spanish LPT 01	75	0.02	0.00	0.15	1.31
Spanish LPT 02	75	-0.01	0.01	0.15	1.26
Spanish LPT 03	75	0.87	0.92	0.14	1.17
Spanish LPT 04	75	-0.04	0.12	0.15	1.32
Spanish LPT 05	75	-0.10	-0.06	0.18	1.52

The medians and means for the initial (non-revised) Arabic test forms varied by approx. a half logit, indicating that Arabic LPT 02 was slightly less difficult on average than Arabic LPT 01. The two German initial test forms were almost identical on average as were most of the Spanish test forms except LPT 03, which was close to one logit more difficult than all other forms. With the exception of Spanish LPT 03, the statistics in Table 17, thus, support the claim that test forms are more or less equivalent for all current forms of the Arabic, German, and Spanish listening proficiency test. With the exception of Spanish LPT 05, which was very similar to Spanish LPT 01 even initially, all forms were revised and are expected to be much closer in difficulty level with each other as a result.

16 Scorer Reliability

Not applicable

17 Cut-Score Classification Errors

Cut scores were determined empirically through a side-by-side study between the LPT and NATO's Benchmark Advisory Test – Listening. Because the LPT is a high stakes test, false positive classification decisions were considered to be relatively more serious than false negative classification errors. Therefore, cut scores were set at the upper end of the cut score range determined by the calibration study (See Section 19). The reasonableness and appropriateness of the cut scores were reconfirmed in two posterior studies: a standard-setting workshop relying on expert judgments in German (see Section 20) and an analysis of the means of the two sublevels that are rated together for all Spanish tests (see Section 30). These three sources of evidence provide evidence that the cut-scores established for the LPT do not exhibit any serious classification errors (see Section 30 for more information).

18 Content-Related Validity

Each exam provides a representative sample of the construct by including a broad spectrum of topics, subtopics, genres, and rhetorical organization (text type). The LPT is commonly taken as a two-sublevel test and consists of ten passages, five at each level. The ten passages are chosen to provide a representative statement of the language proficiency of the examinee. In Section 8, three examples of different two-level tests were presented to show how the passages reflect the *ACTFL Proficiency Guidelines 2012 – Listening*, and how the test ensures selecting a diverse and representative sample of the topics, subtopics, genres, and rhetorical organization of passages listeners need to be able to understand to be rated a particular proficiency level. These examples showed that the tasks in any single exam cover a broad spectrum of topics, subtopics, genres and

rhetorical organization and provide a solid and representative statement of the listening proficiency of examinees.

19 Criterion-Related Validity

The ACTFL LPT was externally validated by a side-by-side study with NATO's Benchmark Advisory Test – Listening (BAT-L) (see Appendix 6 – Technical Report). The present section describes 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 and Instruments

The subjects were students of English at the University of Leipzig ranging from beginning to very advanced levels (Bärenfänger & Tschirner, 2013). Both the ACTFL LPT and NATO's BAT-L were administered to a total of 88 examinees. The BAT-L measures listening proficiency using the STANAG 6001 scale, which is derived from the ILR scale, the scale used by U.S. government agencies. The ILR scale was also used as the basis for the ACTFL scale. Both, the ILR and ACTFL scales continue to be commensurate, which means that there are precise correspondences between ACTFL and ILR levels.

To ensure 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 relatively rare, the proportion of participants with beginning proficiency in English was smaller than that of participants with more advanced proficiency.

Research Design

Both, the LPT and BAT-L were administered to the same group of students in a split test design. Half the participants took the 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 LPT sublevels IL, IM, and AL and BAT-L levels 1 and 2. Mid-level proficiency students took LPT sublevels AL and AM and BAT-L levels 1 and 2. High-level proficiency students took LPT sublevels AL, AM, and S and BAT-L levels 2 and 3. Participants were given 75 minutes for the three-sublevel LPT and the BAT-L and 50 minutes for the two-sublevel LPT. Tests were computer-scored according to the LPT's standard internal scoring algorithm. For the three-sublevel 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 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. 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 LPT.

To gain insights into the *external validity* of the ACTFL LPT, raw percentages of agreement between the LPT and BAT-L were cross-tabulated, and the following correlation values were computed: Raw percentage of agreement; Pearson's correlation; Spearman's *rho*; Kendall's *tau*; and Goodman and Kruskal's *gamma*.

Data Analysis

Table 19 displays all measures that were computed to establish the ACTFL LPT's *external validity*. It contains four parameters, which describe the relationship between the ACTFL LPT and the BAT-L. Two correlation and two agreement measures were computed. Both correlation parameters, Pearson's *r* and Spearman's *rho* show high interdependence between the two tests. As for the agreement measures, Kendall's *tau* is affected by bindings in the data and thus somewhat lower than Goodman-Kruskall's *gamma*. Both indicators support, however, the conclusion that there is high agreement between the ratings of both tests.

Table 19
Correlation and Agreement Measures Between Final Ratings of the ACTFL LPT and the BAT-L

<i>N</i>	Pearson's <i>r</i>	Spearman's <i>rho</i>	Kendall's <i>tau</i>	Goodman-Kruskall's <i>gamma</i>
88	.842*	.833*	.753*	.898*

*Correlations were significant at $p < 0.01$.

To confirm that the results of the LPT show the correct correspondences between ACTFL and ILR levels, the frequency distribution between the two sets of results was examined. Table 20 presents the frequency of agreement in final ratings between the LPT and the BAT-L.

Table 20
Frequency of Agreement in Final Ratings between the LPT and the 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)		
	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.

Table 20 shows that the following correspondences between the results of the two tests had the greatest proportion of agreement: IL and ILR 1 (60%); IM and ILR 1 (57%); AL and ILR 2 (66%); S and ILR 3 (62%). This represents the relationship between ACTFL and ILR well. The established correspondences between ACTFL and ILR are as follows: IL corresponds to ILR 1; IM (rarely IH) corresponds to ILR 1+; AL corresponds to ILR 2; AM (rarely AH) corresponds to ILR 2+; and baseline Superior corresponds to ILR 3.

The finding that IL agrees with 0+ (40%) and ILR 1 (67%), i.e., the lower ILR 1 ranges, and that IM agrees with ILR 1 (57%) and ILR 1+ (21%), i.e., the higher level 1 ranges, is consistent with the relationship between ACTFL and ILR as established above. Similarly, the finding that AL corresponds to ILR 1+ (23%) and ILR 2 (66%), i.e., the lower ILR 2 ranges, and AM corresponds to ILR 2 (14%) and ILR 2+ (29%) and even to ILR 3 (43%), i.e., the higher level 2 ranges, is also consistent with the established relationship between ACTFL and ILR. Superior, finally, clearly corresponds to ILR 3 (62%). The results of this study, therefore, provide external validity evidence, i.e., criterion-related validity evidence.

20 Construct-Related Validity

There are two pieces of evidence to support the construct validity of the LPT: The results of a standard-setting workshop and the Rasch model fit.

Standard-Setting Workshop

Another piece of evidence comes from a two-day standard-setting workshop, which was conducted with the German LPT 01 in July 2015. Eight experts with a college degree in German as a

Agreement	1.00	1.00	0.63	1.00	0.38	0.88	0.75	0.75	1.00
SD	0.00	0.00	0.52	0.00	0.52	0.35	0.46	0.46	0.00
	A2.2.1	A2.2.2	A2.2.3	A2.3.1	A2.3.2	A2.3.3	A2.4.1	A2.4.2	A2.4.3
N	8	8	8	8	8	8	8	8	8
Agreement	0.75	1.00	0.38	0.75	0.75	1.00	0.88	1.00	0.50
SD	0.46	0.00	0.52	0.46	0.46	0.00	0.35	0.00	0.53
	A2.5.1	A2.5.2	A2.5.3	B1.1.1	B1.1.2	B1.1.3	B1.2.1	B1.2.2	B1.2.3
N	8	8	8	8	8	8	8	8	8
Agreement	0.75	1.00	0.88	0.88	0.38	0.50	0.75	0.13	0.88
SD	0.46	0.00	0.35	0.35	0.52	0.53	0.46	0.35	0.35
	B1.3.1	B1.3.2	B1.3.3	B1.4.1	B1.4.2	B1.4.3	B1.5.1	B1.5.2	B1.5.3
N	8	8	8	8	8	8	8	8	8
Agreement	0.88	0.75	0.63	0.88	1.00	0.63	0.38	1.00	0.50
SD	0.35	0.46	0.52	0.35	0.00	0.52	0.52	0.00	0.53
	B2.1.1	B2.1.2	B2.1.3	B2.2.1	B2.2.2	B2.2.3	B2.3.1	B2.3.2	B2.3.3
N	8	8	8	8	8	8	8	8	8
Agreement	1.00	0.88	0.50	0.75	0.75	0.88	0.50	0.50	0.25
SD	0.00	0.35	0.53	0.46	0.46	0.35	0.53	0.53	0.46
	B2.4.1	B2.4.2	B2.4.3	B2.5.1	B2.5.2	B2.5.3	C1.1.1	C1.1.2	C1.1.3
N	8	8	8	8	8	8	8	8	8
Agreement	0.88	0.75	0.88	1.00	0.75	0.75	1.00	0.63	0.75
SD	0.35	0.46	0.35	0.00	0.46	0.46	0.00	0.52	0.46
	C1.2.1	C1.2.2	C1.2.3	C1.3.1	C1.3.2	C1.3.3	C1.4.1	C1.4.2	C1.4.3
N	8	8	8	8	8	8	8	8	8
Agreement	0.88	1.00	0.38	0.63	0.75	0.75	1.00	0.75	0.75
SD	0.35	0.00	0.52	0.52	0.46	0.46	0.00	0.46	0.46
	C1.5.1	C1.5.2	C1.5.3						
N	8	8	8						
Agreement	0.75	0.88	1.00						
SD	0.46	0.35	0.00						

Rater agreement of 0.5 and higher indicates that the majority of raters believed that the item matches the test construct of a particular sublevel. As Table 21 shows, there were only 3 out of 75 cases, where the raters judged an item too difficult for the targeted proficiency level; in all other cases, raters agreed with the level the item was supposed to target. This finding provides evidence of the alignment of the test with the construct matrix and proficiency scale.

Rasch Model Fit

The second piece of evidence of the construct validity of the LPT comes from Rasch measurement. Rasch statistics impose a theoretical model – in this case the Rasch model for dichotomous items – on empirical data. When the observed data fit the theoretical model, this may be interpreted as an indication of the validity of the model, i.e., of construct validity. Rasch person infit and outfit

values for each test form was provided in Table 16 in Section 14. For ease of reference, it is repeated in Table 22. A value of 1.0 implies a perfect fit, while values between 0.5 and 1.5 are considered to be an acceptable fit.

Table 22
Rasch Person Infit and Outfit Values

	N	Rasch Item Infit (MNSQ)	Rasch Item Outfit (MNSQ)
Arabic LPT 01	545	0.99	1.04
Arabic LPT 02	280	0.96	1.00
German LPT 01	443	1.00	1.06
German LPT 02	1,156	0.99	0.99
Spanish LPT 01	1,658	1.01	1.07
Spanish LPT 02	1,038	1.00	1.05
Spanish LPT 03	1,223	1.03	1.08
Spanish LPT 04	628	0.99	1.02
Spanish LPT 05	338	0.99	1.02

As Table 22 shows, the data fit the model impressively well. All infit values fall within 0.04 MNSQ of a perfect fit of 1.0; all outfit values fall within 0.08 MNSQ of the perfect fit of 1.0. All test forms, therefore, are highly predictive of examinees' performance. This provides strong evidence of the construct validity of the test.

21 Item Bias and Differential Item Functioning

Two main aspects for possible item bias are gender-based and culture-based bias. The item writing manual and the two check lists require writers and reviewers to keep these sources of bias in mind when writing and reviewing texts and items. Topics and items are developed to have equal appeal to all genders, and they are developed and reviewed equally by female and male authors to avoid gender-based bias.

To avoid discrimination of certain cultures, causing cultural-based item bias, emotionally charged topics such as sexuality, religion, war, or violence as well as topics that are culture-specific are avoided, as is the use of inappropriate language.

Because LTI does not request nor collect personal information from examinees for privacy reasons (see Section 25), it is not possible to calculate differential item functioning (DIF) statistics. The steps outlined in Section 11, therefore, have been put in place to avoid including biased test items before operational testing.

22 Time Limit Appropriateness

To determine if time limits are appropriate and the exam is not unduly speeded, the time it took examinees to finish the test was examined for the time period in question (1/1/20 to 1/1/23). The five-sublevel adaptive test (NL to Superior) was selected, because it was for all three languages the version taken most often (versus the two-sublevel, three-sublevel, and five-sublevel non-adaptive versions). 67% of the Arabic tests, 53% of the German tests, and 30% of the Spanish tests³ were five-sublevel adaptive tests. To avoid an artificial lowering of the mean, examinees who speeded through the test were removed. Speeding was defined as spending less than ten minutes. Moreover, examinees who were not at least Intermediate Low (IL) were also removed. The maximum amount of time provided to examinees for the five-sublevel adaptive test is 75 minutes. Table 23 shows the minimum, maximum, mean, standard error of the mean (SEM), and standard deviation (SD) of the time in minutes it took examinees to take the test per language. In addition, Table 23 shows the percentage of examinees who used the full 75 minutes.

Table 23

Number of Test-Takers by Language, Minimum, Maximum, Mean, and Standard Deviation of Time it Took to Complete the Test, and Percentage of Test-takers who took the full 75 minutes

Language	N	Minimum	Maximum	Mean	SEM	SD	75 min
Arabic	444	23	74	43.99	0.57	12.07	0%
German	72	20	57	31.44	0.95	8.07	0%
Spanish	441	20	73	35.04	0.51	10.76	0%

Table 23 shows that no one used up the full 75 minutes. The average time it took examinees to take the test was 44 minutes in Arabic, 31 minutes in German, and 35 minutes in Spanish. This may be taken as evidence that the time limits are appropriate and that the test is not unduly speeded.

23 Exam Administration Standardization Provisions

This section summarizes the provisions for standardizing the administration of the examination (see Appendix 1 – Familiarization Manual and Appendix 10 – Examinee Handbook). Impartial treatment of examinees during all aspects of the administration of the LPT from registering for the assessment to taking the assessment is ensured by the following regulations:

- Individuals have equal access to information about the LPT content and procedures.

³ For Spanish, the three-level test was actually taken the most often, albeit by only 12 test administrations. The adaptive test was selected for Spanish, too, however, to use the same test version for all three languages.

- Individuals have equal access to the LPT, in terms of cost, location, and familiarity with conditions and equipment.
- Individuals have equal opportunity to demonstrate the ability to be assessed.

Examinees may access information about the test and download the LPT Familiarization Manual and the Examinee Handbook from the official homepage of Language Testing International (LTI), the ACTFL Testing Office.

The LPT is delivered over the Internet using the same test algorithm every single time and it is accessible to examinees in any part of the world where there is reliable Internet availability.

The LPT is a machine-scored test administered online. Official ACTFL LPT ratings are assigned to LPTs by LTI. Persons supervising the test are required to treat all examinees impartially following procedures described in the Examinee Handbook.

24 Exam Security Provisions

Language Testing International (LTI), ACTFL's test administration office, has built test registration, scheduling, test management, and delivery test processing platforms that meet the high security standards for encrypting personal information and hosting tests on Amazon Web Services (AWS). Data is securely backed up in redundant locations in order to ensure 24/7 performance and data security.

At the completion of every test, answers are immediately streamed to a secured cloud datacenter, preventing the possibility of any response being stored. All servers are hardened for security and are also part of a high-availability cloud cluster. Cloud servers are managed and monitored by the data center, in conjunction with LTI, for performance and security events. Responses are backed up daily and the data is stored in a secure environment.

LTI's Client Site, a part of the aforementioned test management system, is a web-based portal that provides those who are registering for an ACTFL test with various options to register and monitor progress throughout the testing process, from pre-test to post-test administration. Access to LTI's Client Site is privilege-based and restricts modules' access to users based on their accounts' configuration. Users can: (1) request language tests; (2) view all of the tests that have been completed along with their results; (3) generate certificates of proficiency for relevant tests; and (4) update billing information.

All records are stored electronically in a secure environment. Examinees' names and assessment results are stored securely in LTI's database repository. All personally identifiable information is digitally encrypted to prevent unauthorized access. LTI's production servers are located in an SOC 2 compliant datacenter where access is secured using biometric access controls.

LTI intentionally uses only the minimum amount of data needed to take a test. LTI will not disclose any customer identifiable information (CII), such as customer name, home or email address, or phone number unless directed by the customer. LTI may use anonymous, aggregated information about its customers for internal research, or to update and/or maintain its systems. However, LTI does not sell, rent, or loan any CII to any third parties that are not authorized service providers, or who are not clients with whom LTI has signed Confidentiality Agreements concerning the use of Customer Information. LTI's full privacy statement is located at: <https://www.languagetesting.com/privacy>.

25 IRT Scaling Models Used

The IRT model used is the Rasch model for dichotomous items. All items are dichotomously scored as correct or incorrect. The Rasch model was selected because it allows person ability and item difficulty measures to be put on the same scale and because it works well with responses that consist of yes/no answers (correct/incorrect). The full model is used for scoring purposes in the ACTFL Listening and Listening Computer-Adaptive Test (L&Rcat). For the LPT (and RPT), the model is used for scaling new items on the current scale with the help of anchor items (see Section 15).

26 IRT Model's Evidence of Fit

See *Rasch Model Fit* in Section 20 for evidence that the items of the 9 LPT forms for Arabic, German, and Spanish fit the Rasch model to a very high degree.

27 Evidence that New Items/Tests Fit the Current Scale Used

To ensure that test results with new items (new forms) have the same meaning and interpretation as the previous form, a total of nine anchor items are used. See Section 15 for evidence that the new items for each subsequent form fit both the IRT model and scale previously adopted and used.

28 Exposure Rate of Items and Operational Test Item

New forms are generally introduced after 300 to 500 test administrations. Repeat test takers will never be exposed to any items previously seen. They have to wait for 90 days before they can retake the test. LTI keeps track of which form an examinee took so that a different but equivalent form of the test can be used when they retake the test. Currently, there are three different but

equivalent forms for Arabic, three for German, and eight for Spanish. New forms are developed continually.

29 Equivalency Between Hardcopy and Computer Administration

The ACTFL LPT was designed as a computer-administered test from the beginning. There are no paper-and-pencil versions.

30 Cut-Score Rationale, Reasonableness, and Appropriateness

Cut scores were determined empirically through a side-by-side study between the LPT and NATO's Benchmark Advisory Test – Listening (BAT-L) (See Section 19 and Appendix 6 – Technical Report). The BAT-L rates listening proficiency using the STANAG 6001 scale, which is derived from the ILR scale, the scale used by U.S. government agencies. The ILR scale was also used as the basis for the ACTFL scale resulting in precise correspondences between ACTFL and ILR levels.

The BAT-L uses a percentage system to convert scores to levels: 1-30% is considered to be a random effect (may, e.g., be achieved by guessing); 31-50% is considered emerging proficiency; 51-70% is called developing proficiency; and a score above 70% is considered as evidence of a proficiency level. The side-by-side study revealed that for the LPT, the percentages that aligned best with the results of the BAT-L were 40%, 60%, and 80%. Scores below 40%, i.e., below 12, were found to be *random*, i.e., they indicated no evidence of any level; scores between 60% and 79%, i.e., between 18 and 23, were found to provide evidence of the examinee being at the lower of the two levels considered; and scores of 24 and above were found to provide evidence of the examinee having reached the higher of the two levels considered.

Because the LPT is a high stakes test, false positive classification decisions were considered to be relatively more serious than false negative classification errors. Therefore, cut scores were set at the upper end of the cut score range determined by the calibration study. (See Section 19 for more information on the study and the way cut-scores were determined).

These cut-scores were verified in a later study using another type of empirical data, the results of a standard-setting workshop relying on expert judgments (see Section 20). Table 24 displays the mean agreement of the expert judges across all items of the main proficiency sublevels of the test.

Table 24
Mean Rater Agreement on the Cut-Scores of the German LPT 01

	N	Cut-Score IL	Cut-Score IM	Cut-Score AL	Cut-Score AM	Cut-Score S

German	10	.80 (SD* = .28)	.75 (SD = .35)	.85 (SD = .29)	.80 (SD = .33)	.79 (SD = .33)
--------	----	-----------------	----------------	----------------	----------------	----------------

*SD = Standard Deviation

As Table 24 shows, the cut-scores as estimated in the standard-setting workshop were consistently in the range of 0.75 and 0.85. Because it seems safe to assume that examinees have to answer at least 70% of the items of any proficiency sublevel correctly to be placed at that sublevel, these expert judgments provided further evidence of the reasonableness and appropriateness of the cut-scores recommended on the basis of the side-by-side study.

A third piece of evidence that the cut-scores are reasonable and appropriate comes from an analysis of the means of the two sublevels that are rated together. Because the algorithm simply counts the number correct of both sublevels, it is important to know which sublevel contributes most to a rating. While it may be safe to assume that it is not very relevant to distinguish between IL and IM (or AL and AM) items, which are relatively similar to begin with, when determining if an examinee is IL or IM (or AL or AM), and that correct responses of both sublevels may simply be added together, this approach may need to be supported more persuasively for test versions that combine two main levels such as version B, which combines IM and AL items, and version D, which combines AM and S items (see Section 1). Table 25 shows the number of test administrations, the median score, the mean score, and the standard error of the mean (SEM) of the two sublevels rated together of all Spanish tests separately for the lower and the upper rating. The lower rating for Version A is IL and the upper rating is IM. For Version B, the lower rating is IM and the upper rating is AL. For version C, the lower rating is AL and the upper rating is AM. For Version D, the lower rating is AM and the upper rating is S.⁴

Table 25
Number of Test Administrations, Median Score, Mean Score, and SEM of the two Sublevels Rated Together for all Spanish Tests by Rating

Version		Lower Rating				Upper Rating			
		N	Median	Mean	SEM	N	Median	Mean	SEM
A	IL	572	11	11.12	0.07	271	12	11.30	0.14
	IM	572	9	8.78	0.07	271	11	11.34	0.10
B	IM	529	11	10.69	0.07	428	13	13.06	0.07
	AL	529	8	8.83	0.07	428	12	12.04	0.07
C	AL	865	12	11.89	0.04	444	13	12.87	0.08
	AM	865	8	8.06	0.06	444	11	10.59	0.08
D	AM	453	10	10.07	0.07	113	13	12.84	0.11
	S	453	9	9.22	0.08	113	12	12.23	0.13

⁴ the following, the sublevel scores that were rated together of all Spanish tests were used, including the adaptive, the three-level, and the five-level versions, i.e., the results of Versions E through H are included in Table 25 and Figures 4-7 in those instances, in which the respective rating was based on the sublevels under discussion.

Table 25 shows that the mean scores are consistently higher for the lower level than for the higher level, and higher for the upper rating than for the lower rating. The latter is not surprising because one needs a higher score to be rated at the higher level. The former, however, is significant, because it means that examinees do, indeed, get a higher score at the lower level and a lower score at the higher level. Take Version B as an example. Note that Version B is one of the two versions that spans two main levels (Version D is the other one). For the lower rating, i.e., IM, the average examinee score for IM items was 10.69 and for AL items, it was 8.83. This means that examinees generally had close to 11 out of 15 items correct at IM when their final rating was IM, while they had close to 9 out of 15 items correct at AL. This shows that IM items contributed more to the IM rating than AL items. Looking at the upper rating, i.e., AL, one sees that the mean IM score for all examinees was 13.06, while it was 12.04 for AL. This means that AL examinees had, on average, all but two IM items correct, and 12 out of 15 of the AL items. Figures 4-8 show boxplots of all Spanish scores for the two versions under discussion (B and D).

Figure 4
IM and AL Scores for Spanish Version B Examinees Rated IM

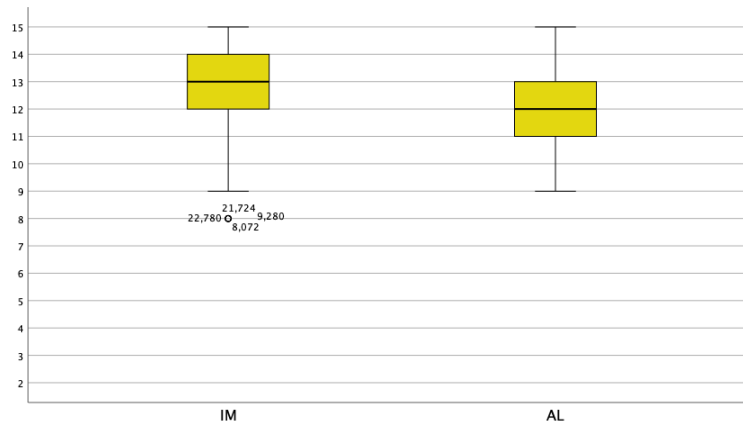
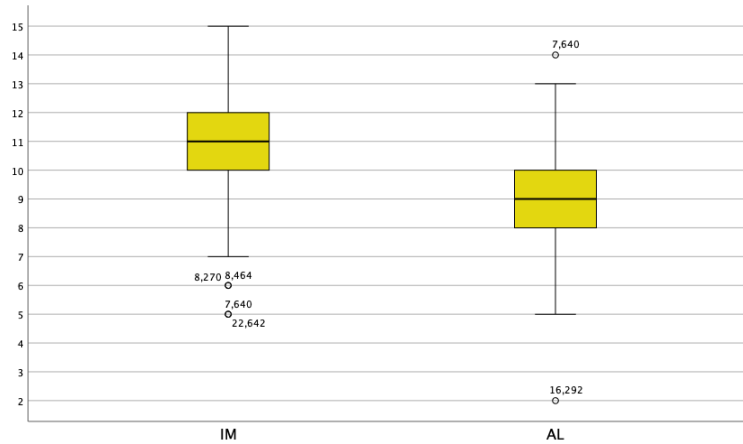


Figure 5
IM and AL Scores for Spanish Version B Examinees Rated AL



Figures 4 and 5 exhibit substantial differences between IM and AL scores of examinees rated IM and AL. Figure 4 shows a median score of 13 and an IQR of 12-14 for IM items and a median score of 12 and an IQR of 11-13 for AL items. Figure 5 shows a median score of 11 and an IQR of 10 to 12 for IM items and a median score of 9 and an IQR of 8-10 for AL items. An paired samples t -test (two-tailed) found that the difference between IM and AL ratings was statistically significant for both IM ($t = 13.72$, $p = 0.000$, $df = 528$) and AL items ($t = 9.71$, $p = 0.000$, $df = 427$).

Figure 6
AM and S Scores for Spanish Version D Examinees Rated AM

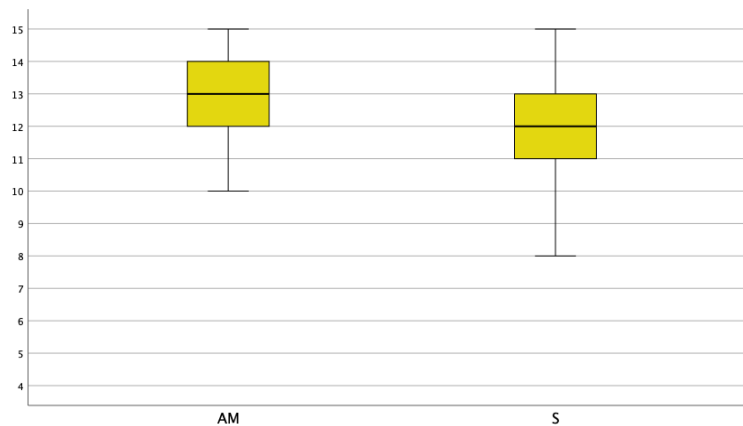
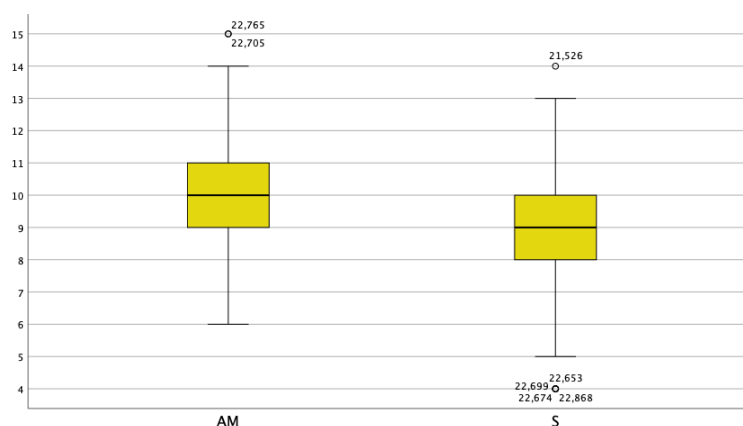


Figure 7
AM and S Scores for Spanish Version D Examinees Rated S



Figures 6 and 7 also show clear differences between AM and S scores of examinees rated AM and S. Figure 6 shows a median score of 13 for IM items and a median score of 12 for AL items with slightly overlapping IQRs. Figure 7 shows a median score of 10 and an IQR of 9-11 for AM ratings and a median score of 9 and an IQR of 8-10 for S ratings. A paired samples *t*-test (two-tailed) found that the difference between AM and S ratings was statistically significant for both AM ($t = 5.99$, $p = 0.000$, $df = 452$) and S items ($t = 2.96$, $p = 0.004$, $df = 112$).

These three sources of evidence: the results of the side-by-side study; the results of the standard-setting workshop; and the analysis of mean scores of 3675 Spanish test administrations collectively provide a great deal of evidence that the cut-scores recommended on the basis of the original side-by-side study are reasonable and appropriate.

31 Procedures Recommended to Users for Establishing their own Cut-Scores

LPT raw scores are converted to ACFTL proficiency levels depending on the version of the test, e.g., IL to IM, AL to AM, etc. The same raw scores, therefore, have different meanings depending on the ranges considered. This means, that raw scores cannot be used to recommend college credit. Instead, ACTFL proficiency levels may be used. Because ACTFL proficiency levels follow the same logic across the modalities of speaking, writing, reading, and listening, it is recommended to use the same ACTFL sublevels for listening (and reading) as for speaking and writing. Table 26 shows the recommendations for granting college credit for each ACTFL proficiency level.

Table 26
Recommendations for Granting College Credit

Official ACTFL LPT Rating	Category I French, Italian, Spanish, Portuguese	Category II German	Category III Russian	Category IV Arabic, Chinese, Japanese, Korean
Novice High/Intermediate Low	2 LD*	2 LD	3 LD	3 LD
Intermediate Mid	4 LD	4 LD	6 LD	6 LD
Intermediate High/Advanced Low	6 LD	6 LD	8 LD	8 LD
Advanced Mid	8 LD + 2 UD**	8 LD + 3 UD	6 LD + 4 UD	6 LD + 5 UD
Advanced High / Superior	8 LD + 2UD	8 LD + 3 UD	6 LD + 6 UD	6 LD + 6 UD

*LD = Lower division baccalaureate/associate degree category

**UD = Upper division baccalaureate degree category

These recommendations are supported by the results of a nation-wide study examining listening proficiency levels of college students (cf. Tschirner, 2016, and the studies collected in Winke & Gass, 2018). Table 27 shows average listening proficiency ratings of college students after having completed two, four, six, or eight semesters of Chinese⁵, German, or Spanish.

Table 27
Mean Listening Proficiency Levels of Chinese, German, and Spanish Students at U.S. Colleges and Universities with Numbers of Tests in Parentheses

Semester	Chinese		German		Spanish	
	Speaking	Listening	Speaking	Listening	Speaking	Listening
2	NH (55)	NL (53)			NH (342)	NM (344)
4	NH (68)	NM (64)	IL-IM (194)	NH-IL (312)	IL (436)	NH (418)
6	IM (43)	IL (33)	IM (36)	IH (34)	IM (501)	IM (456)
8	IM (26)		IH-AL (45)	IH (67)	IH (233)	AL (154)

Table 27 shows that proficiency levels of college students are lower for listening than for speaking after two and four semesters and lower for Chinese than for German or Spanish and that they are broadly comparable across modalities and languages after six semesters. They revert to being lower for Chinese than for German or Spanish after eight semesters. These results provide additional evidence for the credit recommendations in Table 26 above.

⁵ The study did not yield sufficient students of Arabic to compare speaking and listening proficiency levels. Therefore, the results for Chinese, another Category IV language, are presented (s. Tschirner et al., unpublished manuscript).

32 Information on Norms and Normative Groups (If Appropriate)

Not applicable

References

- American Educational Research Association, American Psychological Association, National Council on Measurement in Education (2014). *Standards for Educational and Psychological Testing*. Washington, DC: AERA.
- Bärenfänger, O., & Tschirner, E. (2013). *Assessing Evidence of Validity of the ACTFL CEFR Listening Proficiency Test (LPT)* (Technical Report 2013-US-PUB-2). Leipzig: Institut für Testforschung und Testentwicklung.
- Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement*, 34, 353–366.
- Tschirner, E., Bärenfänger, O., & Wisniewski, K. (2015). *Assessing Evidence of Validity of the ACTFL CEFR Reading and Listening Proficiency Tests (RPT and LPT) Using a Standard-Setting Approach* (Technical Report 2015-EU-PUB-2). Leipzig: Institut für Testforschung und Testentwicklung.
- Tschirner, E. (2016). Listening and listening proficiency levels of college students. *Foreign Language Annals*, 49, 201-223.
- Tschirner, E., Gass, S., Hacking, J., Rubio, F., Sonesson, D., & Winke, P. (unpublished manuscript). *The Role of Listening in the Growth of Speaking Ability*.
- Winke, P. & Gass, S. M. (Eds.). (2018). *Foreign Language Proficiency in Higher Education* (Vol. 37). Springer.
- Weir, C., & Khalifa, H. (2008). A cognitive processing approach towards defining reading comprehension. *Cambridge ESOL Research Notes*, 31, 2–10.

Appendices

- Appendix 1: Familiarization Manual
- Appendix 2: Assessment Use Argument
- Appendix 3: Design Statement
- Appendix 4: Blueprint
- Appendix 5: Construct Matrix
- Appendix 6: Technical Report
- Appendix 7: Arabic, German, and Spanish LPT Data Reports
- Appendix 8: Item Writing Manual
- Appendix 9: Item Checklists
- Appendix 10: Examinee Handbook
- Appendix 11: Certificate