Language Research and Development, Inc.

Erwin Tschirner, PhD
Geibelstr. 64a
04129 Leipzig
Germany

September 25, 2015

# Assessing Evidence of Validity and Reliability of the

# ACTFL Listening Proficiency Test (LPT)

**Technical Report 2015/3-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

# 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 three languages for which there exist sufficient data, i.e., Spanish, French, and German.

## 1. 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. The ACTFL LPT is a carefully constructed assessment based on the *ACTFL Proficiency Guidelines 2012 – Listening* that evaluates Novice to Superior levels of listening ability. It is delivered by computer via the Internet. The test can assess a specific range of proficiency. The available ranges are shown in Table 1 below. These options ensure that the test administered targets the range of the test-taker's listening ability and is economical in terms of time and effort.

**Number of parts:**

One

**Number of Tasks per Part:**

There are five task proficiency sublevels: Intermediate Low (IL); Intermediate Mid (IM); Advanced Low (AL); Advanced Mid (AM); and Superior (S). The number of tasks per part depends on the range of proficiency to be assessed (see Table 1 below). There are two-sublevel (A-D), three-sublevel (E-F) and full-range tests (G-H). There are five listening passages (tasks) per sublevel, each followed by three multiple-choice items (15 items per level) with four options, of which only one is correct. Version A includes five IL and five IM tasks; Version B includes five IM and five AL tasks; Version C includes five AL and five AM tasks, and Version D includes five AM and five S tasks. Version E includes IL, IM, and AL tasks, Version F includes AL, AM, and S tasks, and Version H includes IL, IM, AL, AM, and S tasks. Version G is a semi-adaptive version of the test, which starts at Advanced Low, and based on the candidate's responses, moves to higher or lower level tasks. Depending on the candidate's proficiency, it includes between 10 and 15 tasks. If the candidate is at least IM or at best AM, the test contains ten tasks (five IM and five AL, or five AL and five AM tasks, respectively). If the candidate is below IM or better than AM, the test includes 15 tasks (five tasks each at IL, IM, and AL, or five tasks each at AL, AM, and S, respectively).

**Table 1. Test Versions and Ranges Assessed**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Superior | | | | D | | F | G | H |
| Advanced High | | | | ▨ | | ▨ | ▨ | ▨ |
| Advanced Mid | | | C | ▨ | | ▨ | ▨ | ▨ |
| Advanced Low | | B | ▨ | ▨ | E | ▨ | ▨ | ▨ |
| Intermediate High | | ▨ | ▨ | ▨ | ▨ | ▨ | ▨ | ▨ |
| Intermediate Mid | A | ▨ | ▨ | | ▨ | ▨ | ▨ | ▨ |
| Intermediate Low | ▨ | ▨ | | | ▨ | | ▨ | ▨ |
| Novice High | ▨ | ▨ | | | ▨ | | ▨ | ▨ |
| Novice Mid | ▨ | | | | ▨ | | ▨ | ▨ |
| Novice Low | ▨ | | | | ▨ | | ▨ | ▨ |

There are four different item types: Global, Detail, Selective, and Inference (see Table 2 for number of item types per sublevel):

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

**Table 2. Number of Item Types per Sublevel.**

| **Level** | **IL** | **IM** | **AL** | **AM** | **S** |
|---|---|---|---|---|---|
| Number of questions | Selective: 5<br>Detail: 10 | Global: 5<br>Detail: 10 | Global: 5<br>Detail: 10 | Global: 5<br>Detail: 5<br>Inference: 5 | Global: 5<br>Detail: 5<br>Inference: 5 |

## Sequence of tasks

All tasks are listed on the left-hand side of the screen and ordered from easier to more difficult: five tasks at a lower level are presented first and are followed by tasks at the next higher level. The tasks within one level may appear in random order, however, within a task, the order or sequence of questions (items) remains unchanged because they follow a logical order. They are sequenced according to the parts of text that contain the answer.

There are four timed phases to every LPT test item. In the first phase, the title of the passage and the questions pertaining to the passage appear. This gives the test-taker time to become familiar with the topic and the questions before he or she hears the passage. The passage is then played, and, while it is played, the test-taker is instructed to take notes on a notepad that is provided within the test. When the passage ends, there is a five-second period to review notes before the multiple-choice options to the questions appear. Test-takers cannot click through any of these stages and must wait until they are advanced to the next stage by the program.

### Relative importance of parts and tasks

All tasks are equally important.

### 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.

Before the listening passage begins, test-takers have 30 seconds to preview the three test questions without the multiple-choice answers. They remain on the screen during all phases of the task.

The test-takers can hear a passage just once. While they are listening, it is highly recommended that they take notes on the notepad provided within the test. Then they are given 5 seconds to organize and review their notes. Finally, the four multiple choice answers to each question appear below each question and the test-takers are given 2 to 2.5 minutes to answer the question based on the difficulty of the task. When the test-takers finish the task they may click “Next Task” to start the next listening task. When the time is up, the test-takers will automatically be brought to the next listening task. They cannot go back to previous tasks once they have opened a new task.

## 2. Rationale and Purpose of the Examination *(See Appendix 4 – Blueprint)*

The LPT measures how well a person understands spoken language extemporaneously when presented with discourse types and listening tasks as described in the *ACTFL Proficiency Guidelines 2012 – Listening*. Listening skills are evaluated without candidates having time and/or access to dictionaries or grammar references.

The items focus on global, detail or selective understanding or on making inferences. Item types are operationalized differently depending on the sublevel tested (see Table 3).

**Table 3. Item Types at All Sublevels.**

| Level | IL | IM | AL | AM | S |
|---|---|---|---|---|---|
| **Global** | | 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. | Ability to understand the main idea depends on comprehending supporting details. Test-taker 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 speaker intent. | 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 speaker or speakers are intending to say. Speaker intent is clearly signaled.<br>. | Fully able to understand the main argument and all supporting facts. It is the main argument the speaker or speakers are making. The correct answer is spread out over different parts of the passage. Distractors refer to other arguments the speaker(s) is/are making or to an argument they could be making based on statements contained in the passage. |
| **Detail** | Able to comprehend simple single facts. These facts are the easiest to understand aurally and do not necessarily have to be important for the passage as a whole. Distractors must be viable passage-based options that are clearly false, however. | Able to comprehend single straightforward facts. These facts contribute to the gist of the passage. Still, their comprehension only requires understanding single utterances. Distractors must be viable passage-based options. Key must use synonyms or paraphrases that consist of highly frequent or shared international vocabulary. | Able to understand explicitly mentioned facts and thoughts. They go beyond single utterance-based facts. Their understanding is dependent on understanding the gist of the passage. They usually require understanding more than one utterance. Distractors focus on other relevant facts mentioned in the passage. Key must use synonyms or paraphrases that contain general vocabulary. | Able to understand explicitly mentioned facts, thoughts, and argument. Their understanding is dependent on understanding the gist of the passage. They usually require understanding more than one utterance. Keys and distractors focus on explicitly mentioned facts or argument. Key must use synonyms and paraphrases that contain a broad general vocabulary. | Able to understand argument, finer points of detail and abstraction. They require understanding complete subsections of the passage. 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 commonly contain precise, specialized and low-frequency vocabulary and |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | complex structure. |
| **Selective** | 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 passage. | | | | |
| **Inference** | | | | Able to identify the main conclusions in clearly signaled argumentative discourse and to make straightforward inferences. Items refer to the complete passage and focus on something that is clearly understood but not explicitly mentioned. | Able to infer attitude, mood, and intentions; able to infer implied as well as stated opinion; 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 did. |

## 3. Name(s) and institutional Affiliations of the Principle Author(s) or Consultant(s)

### Principal Authors

- Dr. Erwin Tschirner, Gerhard Helbig Professor of German as a Foreign Language, University of Leipzig, Germany
- Dr. Olaf Bärenfänger, Director of Language Learning Center, University of Leipzig, Germany

## 4. Specifications That Define the Domain(s) of Content, Skills, and/or Developed Abilities That the Exam Samples (See Appendix 2 – Assessment Use Argument and Appendix 3 – Design Statement)

Based on the *ACTFL Proficiency Guidelines*, the construct matrix defines the domains of content, skills and developed abilities that the exam measures. The target language use (TLU) task that was selected as a basis for developing assessment tasks is listening in general, i.e. retrieving information from a variety of spoken texts in daily life, at work, university or school etc., indicating different aspects of comprehension (global, selective, detail understanding, or making inferences). Tasks are described in terms of function, content, context, text type, vocabulary, grammar and culture at all major ACTFL levels (see Table 4 for the proficiency levels represented by test tasks).

**Table 4. Summary of Task Descriptors at the Proficiency Levels Represented by Test Tasks**

| | Function | Content | Context | Text Type | Vocabulary | Grammar | Culture |
|---|---|---|---|---|---|---|---|
| **Superior** | Argumentation<br>Supported Opinion<br>Hypothesis | Familiar and unfamiliar abstract topics | Professional<br>Academic<br>Literary | Complex, lengthy texts | Broad<br>Precise<br>Specialized | Complex structures | Cultural references<br>Aesthetic properties |
| **Advanced** | Description<br>Narration<br>Exposition<br>Explanation | Concrete current and general interest topics | Public<br>Education<br>Work<br>News | Paragraph-based connected texts with a clear predictable structure | Broad general vocabulary | Sequencing<br>Time frames<br>Chronology | Most common cultural patterns |
| **Intermediate** | Convey basic information | Highly familiar everyday content | Highly familiar everyday contexts | Simple, predictable, loosely connected texts | High frequency vocabulary | Simple sentence patterns and strings of sentences | Some of the most common cultural patterns |

- The term *function* refers to the different purposes texts may have such as instruction, description, narration, explanation, or argumentation.
- The term *content* refers to the content areas that the listener can understand in the language.
- The term *context* refers to the different domains in which the passages are used such as the public, educational, or work domain.
- The term *text type* refers to the quantity, quality and organization of texts 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 can understand in the language.
- The term *culture* refers to the range of idiomatic expressions and cultural references the listener can understand in the language.

## 5. Statement of the Exam's Emphasis on Each of the Content, Skill, and/or Ability Areas

The tested contents, skills and ability areas are based on the construct of the *ACTFL Proficiency Guidelines 2012 - Listening (see Construct Matrix in Appendix 1)*. Each exam contains items for at least two sublevels. Thus, at least 10 texts and 30 items form the basis of a rating. This allows to test a representative sample of real-life topics and to make a meaningful statement about the listening proficiency of a test-taker. Depending on the sublevels tested, 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 literary. Taking the example of an exam that tests the sublevels Advanced Mid and Superior, the 10 texts of the exam represent the functions of both levels: description, narration, explanation and exposition at the Advanced level and argumentation, supported opinions and hypothesis at the Superior level. The same distribution goes for the content and the genre. The exam consists of texts of concrete, current and general interest topics and familiar and unfamiliar abstract topics such as news coverage, articles and reports concerned with contemporary social problems, biographical accounts, short stories and Op/ed pieces, analyses and commentaries, detailed technical reports, and/or literary texts.

## 6. Information About Each Task (Item) Included in The Exam *(See Appendix 3 – Design Statement And Appendix 4 – Blueprint)*

### Item Types

There are four item types: Global (IM-S), Detail (all levels), Selective (IL only), and Inference (AM-S). Depending on the level, these item types are defined differently (see below).

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

Global

- 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.
- AL: Ability to understand the main idea depends on comprehending supporting details. Test-taker 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 speaker or speakers are intending to say. Speaker intent is clearly signaled.
- S: Fully able to understand the main argument and all supporting facts. It is the main argument the speaker or speakers are making. The correct answer is spread out over different parts of the passage. Distractors refer to other arguments the speaker(s) is/are making or to an argument they could be making based on statements contained in the passage.

Detail

- IL: Able to comprehend simple single facts. These facts are the easiest to understand aurally and do not necessarily have to be important for the passage as a whole. Distractors must be viable passage-based options that are clearly false, however.
- IM: Able to comprehend single straightforward facts. These facts contribute to the gist of the passage. Still, their comprehension only requires understanding single 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 single utterance-based facts. Their understanding is dependent on understanding the gist of the passage. They usually require understanding more than one utterance. Distractors focus on other relevant facts mentioned in the passage. 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 usually require understanding more than one utterance. Keys and distractors focus on explicitly men-

tioned 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. 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 commonly contain precise, specialized and low-frequency vocabulary and complex structure.

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 passage.

Inference

- AM: Able to identify the main conclusions in clearly signaled argumentative discourse 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; infer implied as well as stated opinion; 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 did.

**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 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 argument and relationships.

## 7. Information About the Adequacy of the Items on the Exam as a Sample From the Domain(s)

Task topics are relevant and interesting to test-takers. Topics such as drugs, sexuality, war, violence, etc. that may engender strong emotional reactions as well as discriminating and linguistically inappropriate content are avoided to ensure equal access to the texts for all test-takers.

In addition, the test 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 test-takers. Each topic is used once at any one level to provide a representative sample of the language proficiency of test-takers across a broad range of topics. Tables 5 and 6 below provide an example of the genres and topics included in a test. Note that these are open lists that are constantly updated.

**Table 5: Task Genres per Sublevel**

| IL | IM | AL | AM | S |
|---|---|---|---|---|
| Advertisement | Advertisement | Advertisement | Advertisement | |
| Business Correspondence | Business Correspondence | Business Correspondence | Business Correspondence | |
| Giving Advice | Giving Advice | Giving Advice | | |
| Personal Correspondence | Personal Correspondence | Personal Correspondence | | |
| Simple Text | Simple Text | | | |
| | Encyclopedia entry | Encyclopedia entry | Encyclopedia entry | Encyclopedia entry |
| | Report | Report | Report | Report |
| | Notice | Notice | | |
| | News Item | News Item | News Item | News Item |
| | Narrative | | | Narrative |
| | | | Op-Ed | Op-Ed |
| | | | Journal Article | Journal Article |
| | | | Review | |

**Table 6: 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 |
| | Reform |
| | Religion |
| | Restaurant |
| | Routine |
| | School |
| | Shopping |
| | Souvenirs |
| | Theater |
| | Trade |
| | Tradition |
| | Traffic |
| | Train |
| | Transportation |

| | |
|---|---|
| | Trends |
| | Trips |
| | TV |
| | Weather |

Ensuring the adequacy of the items is a prominent goal of the training of test authors and reviewers as well as of the multi-stage process of item development, review, and quality assurance.

## 8. Information About Whether and/or How the Items Were Pretested Before Inclusion Into the Final Form *(See Appendix 2 – Assessment Use Argument and Appendix 3 – Design Statement)*

All forms go through a rigorous pilot study process. All tests are usually taken by at least 100 participants ideally with 20 participants at each of the five sublevels. Data reports are completed for all pilot tests. The data report provides the date on which the report was completed, the name of the test, e.g., Spanish LPT 01A, the name of the person completing the report, the date or dates of data collection and the number of participants. The data report provides both a classical item analysis as well as a Rasch analysis.

The classical item analysis provides Cronbach's alpha for two adjoining sublevels, i.e., IL/IM, IM/AL, AL/AM, and AM/S as well as for all sublevels combined. Cronbach's alpha reflects the degree to which the items of two adjoining levels discriminate reliably between test participants of different degrees of ability. Its value for all five levels is an indicator of the overall reliability of the test. Cronbach's alpha should not be lower than .8. In addition to Cronbach's alpha, the report also provides difficulty and separation indices for each item. Difficulty indices should be close to .5, not lower than .1 and not larger than .9. Separation indices should not be lower than .25.

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

The model fit values are calculated by comparing empirical answer patterns with the patterns predicted by the Rasch model in the form of a residual analysis. Whereas infit statistics refer to the randomness of the data and thus to threats to the validity of the model with respect to the data, outfit statistics yield information on outliers (Eckes, 2009). Generally, infit statistics are considered more important than outfit statistics (Bond & Fox, 2007; Eckes, 2009). Both infit and outfit mean-square values range from 0 to infinity. An infit value of 1.0 indicates that the amount of variance in the data is exactly the amount that is predicted by the model. Mean-square values below 1.0 represent less variance in the data than predicted and mean-square values larger than 1.0 represent more variance. While mean-square values below 0.5 or between 1.5 and 2.0 are considered to be less productive but not degrading, mean-square values above 2.0 distort or degrade the measurement system (Linacre, 2012). For this reason, items with fit values above 2.0 are recommended for revision. The closer fit values are to 1.0, the better the model fits the data. Fit values may also be computed for individual items. Again, an item should ideally have infit and outfit values close to one and should not exceed 2.0.

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

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.

A test is released only when both measures of classical and probabilistic test theory point to a high degree of internal validity. Released tests meet all requirements of a standardized high stakes test *(see Appendix 6 – Technical Report).*

## 9. Item Analysis Results (e.g., Item Difficulty, Discrimination, Correlation With External Criteria)

The item difficulty and discrimination parameters for the LPT are presented for the three selected languages, i.e., Spanish, French, and German. These languages were chosen because they had the greatest number of tests. Spanish items were taken between 276 (S) and 1,128 (AL) times; French items were taken between 93 (S) and 584 (IM) times; and German items were taken between 10 (S) and 183 (IM) times. In general, Superior (S) items were taken the least often, while IM and AL items were taken the most often.

**Item difficulty** is reported in logits as estimated by the Rasch model for dichotomous items (see Tables 7-9). 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 of the difficulty estimate is also reported in logits. Please note that these difficulty parameters cannot be compared directly across languages.

Item-scale correlations (point-biserial correlations) are used for **item discrimination**. According to Oller (1979), separation indices should not fall below .25. Unlike the Rasch item difficulty estimates, item-scale correlations are sample-dependent. Sampling errors, e.g. if participants are too strong or too weak, affect the item discrimination parameter.

To gain further insights into the quality of each item, Rasch infit and outfit measures are reported. Fit statistics indicate the degree to which a test item meets the Rasch model expectations.

Fit values between .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.

Tables 7-9 show a variety of measures for all of the items in the test. The items are listed in columns. 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 task, i.e. tasks 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 task 1 item 1.

Row 2 provides the number of test-takers (*N*) taking a particular item; row 3 provides the **item difficulty** in logits; and Row 4 the Standard Error of Measurement (SEM), also expressed in logits. Row 5 provides the **item discrimination** expressed as a point-biserial correlation ($r_{pb}$); and Rows 6 and 7 provide the Rasch infit and outfit values in mean-squares (MNSQ).

A few comments summarizing the data in the tables follow after all three tables.

**Table 7. Item Characteristics Spanish**

| | **A1.1.1** | **A1.1.2** | **A1.1.3** | **A1.2.1** | **A1.2.2** | **A1.2.3** | **A1.3.1** | **A1.3.2** | **A1.3.3** | **A1.4.1** | **A1.4.2** | **A1.4.3** | **A1.5.1** | **A1.5.2** | **A1.5.3** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ***N*** | 735 | 735 | 735 | 735 | 735 | 735 | 735 | 735 | 735 | 735 | 735 | 735 | 735 | 735 | 735 |
| **Difficulty (logits)** | -1.01 | -2.16 | -1.18 | -1.18 | -.58 | -2.32 | -1.52 | -1.51 | -1.56 | -.60 | -3.22 | .37 | -2.00 | -1.22 | .44 |
| **SEM (logits)** | .08 | .10 | .08 | .08 | .08 | .10 | .08 | .08 | .09 | .08 | .13 | .09 | .09 | .08 | .09 |
| **Discrimination (rpb)** | .42 | .36 | .47 | .47 | .49 | .44 | .48 | .36 | .37 | .49 | .30 | .23 | .37 | .38 | .32 |
| **Rasch infit (MNSQ)** | .96 | .95 | .90 | .91 | .91 | .88 | .89 | .99 | .98 | .90 | .92 | 1.15 | .96 | .99 | 1.05 |
| **Rasch outfit (MNSQ)** | .94 | .92 | .89 | .86 | .88 | .70 | .80 | 1.00 | .98 | .89 | 1.00 | 1.20 | .90 | .95 | 1.13 |
| | **A2.1.1** | **A2.1.2** | **A2.1.3** | **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** | **A2.5.1** | **A2.5.2** | **A2.5.3** |
| ***N*** | 1046 | 1046 | 1046 | 1046 | 1046 | 1046 | 1046 | 1046 | 1046 | 1046 | 1046 | 1046 | 1046 | 1046 | 1046 |
| **Difficulty (logits)** | -.87 | -1.12 | .60 | -.36 | 2.07 | .57 | 1.65 | 2.57 | 1.10 | -.96 | -.86 | -.64 | .29 | -.01 | -.09 |
| **SEM (logits)** | .07 | .07 | .07 | .07 | .10 | .07 | .09 | .12 | .08 | .07 | .07 | .07 | .07 | .07 | .07 |
| **Discrimination (rpb)** | .56 | .37 | .46 | .54 | .39 | .46 | .29 | -.08 | .33 | .41 | .34 | .58 | .26 | .33 | .36 |
| **Rasch infit (MNSQ)** | .84 | 1.03 | .95 | .87 | .92 | .94 | 1.05 | 1.28 | 1.06 | 1.00 | 1.09 | .82 | 1.18 | 1.11 | 1.07 |
| **Rasch outfit (MNSQ)** | .76 | 1.00 | .95 | .85 | .97 | .96 | 1.37 | 3.00 | 1.14 | .94 | 1.07 | .76 | 1.29 | 1.16 | 1.10 |
| | **B1.1.1** | **B1.1.2** | **B1.1.3** | **B1.2.1** | **B1.2.2** | **B1.2.3** | **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*** | 1128 | 1128 | 1128 | 1128 | 1128 | 1128 | 1128 | 1128 | 1128 | 1128 | 1128 | 1128 | 1128 | 1128 | 1128 |
| **Difficulty (logits)** | -1.94 | 1.46 | -1.21 | .52 | .99 | .46 | -1.56 | -.56 | 1.23 | -1.42 | -1.51 | -1.99 | -2.97 | -.35 | -1.08 |
| **SEM (logits)** | .09 | .07 | .08 | .07 | .07 | .07 | .08 | .07 | .07 | .08 | .08 | .10 | .14 | .07 | .07 |
| **Discrimination (rpb)** | .39 | .02 | .36 | .33 | .33 | .52 | .24 | .31 | .48 | .35 | .45 | .33 | .23 | .52 | .48 |
| **Rasch infit (MNSQ)** | .90 | 1.34 | .98 | 1.09 | 1.06 | .87 | 1.03 | 1.07 | .90 | .98 | .88 | .96 | .98 | .86 | .87 |
| **Rasch outfit (MNSQ)** | .65 | 1.70 | .92 | 1.11 | 1.14 | .89 | 1.28 | 1.12 | .93 | .88 | .67 | .78 | .79 | .82 | .74 |
| | **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** | **B2.4.1** | **B2.4.2** | **B2.4.3** | **B2.5.1** | **B2.5.2** | **B2.5.3** |
| ***N*** | 945 | 945 | 945 | 945 | 945 | 945 | 945 | 945 | 945 | 945 | 945 | 945 | 945 | 945 | 945 |
| **Difficulty (logits)** | .93 | 3.10 | -.19 | -.39 | 1.13 | -.03 | -.56 | -.36 | .65 | .55 | .87 | 1.48 | .82 | 1.55 | .63 |
| **SEM (logits)** | .07 | .12 | .07 | .08 | .07 | .07 | .08 | .07 | .07 | .07 | .07 | .08 | .07 | .08 | .07 |
| **Discrimination (rpb)** | .32 | .15 | .40 | .51 | .36 | .48 | .39 | .53 | .41 | .50 | .44 | .45 | .30 | .30 | .18 |
| **Rasch infit (MNSQ)** | 1.09 | 1.08 | .99 | .88 | 1.04 | .92 | 1.00 | .85 | 1.01 | .90 | .96 | .92 | 1.12 | 1.08 | 1.24 |
| **Rasch outfit (MNSQ)** | 1.11 | 1.54 | 1.00 | .81 | 1.09 | .88 | .96 | .78 | 1.01 | .88 | .44 | .96 | 1.14 | 1.18 | 1.32 |
| | **C1.1.1** | **C1.1.2** | **C1.1.3** | **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** | **C1.5.1** | **C1.5.2** | **C1.5.3** |
| ***N*** | 276 | 276 | 276 | 276 | 276 | 276 | 276 | 276 | 276 | 276 | 276 | 276 | 276 | 276 | 276 |
| **Difficulty (logits)** | 2.09 | .14 | .08 | -.14 | -.10 | 1.30 | .10 | .99 | 2.14 | 1.04 | -.34 | 1.96 | 2.02 | 1.41 | 2.33 |
| **SEM (logits)** | .15 | .14 | .14 | .14 | .14 | .14 | .14 | .13 | .15 | .13 | .15 | .15 | .15 | .14 | .16 |
| **Discrimination (rpb)** | .21 | .43 | .59 | .38 | .54 | .38 | .55 | .23 | .24 | .20 | .49 | .13 | .24 | .24 | .24 |
| **Rasch infit (MNSQ)** | 1.05 | 1.01 | .83 | 1.05 | .88 | 1.02 | .88 | 1.20 | 1.06 | 1.22 | .92 | 1.17 | 1.06 | 1.16 | 1.03 |
| **Rasch outfit (MNSQ)** | 2.07 | 1.02 | .76 | 1.14 | .82 | 1.06 | .84 | 1.30 | 1.35 | 1.43 | .95 | 1.76 | 1.49 | 1.29 | 1.44 |

**Table 8. Item Characteristics French**

| | **A1.1.1** | **A1.1.2** | **A1.1.3** | **A1.2.1** | **A1.2.2** | **A1.2.3** | **A1.3.1** | **A1.3.2** | **A1.3.3** | **A1.4.1** | **A1.4.2** | **A1.4.3** | **A1.5.1** | **A1.5.2** | **A1.5.3** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ***N*** | 363 | 363 | 363 | 363 | 363 | 363 | 363 | 363 | 363 | 363 | 363 | 363 | 363 | 363 | 363 |
| **Difficulty (logits)** | -1.38 | -.74 | -2.30 | -2.20 | -2.28 | -.23 | -1.16 | .03 | 1.09 | -2.12 | -.68 | -2.34 | -3.52 | -1.69 | .46 |
| **SEM (logits)** | .12 | .11 | .13 | .13 | .13 | .12 | .11 | .12 | .16 | .13 | .11 | .13 | .19 | .12 | .13 |
| **Discrimination (rpb)** | .51 | .49 | .39 | .56 | .48 | .31 | .51 | .48 | .29 | .49 | .25 | .45 | .32 | .52 | .40 |
| **Rasch infit (MNSQ)** | .88 | .90 | .95 | .80 | .86 | 1.07 | .88 | .88 | .98 | .86 | 1.13 | .90 | .93 | .86 | .92 |
| **Rasch outfit (MNSQ)** | .83 | .88 | .92 | .66 | .76 | 1.08 | .84 | .92 | 1.28 | .77 | 1.23 | .77 | .76 | .79 | 1.12 |
| | **A2.1.1** | **A2.1.2** | **A2.1.3** | **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** | **A2.5.1** | **A2.5.2** | **A2.5.3** |
| ***N*** | 584 | 584 | 584 | 584 | 584 | 584 | 584 | 584 | 584 | 584 | 584 | 584 | 584 | 584 | 584 |
| **Difficulty (logits)** | .69 | -.46 | -1.61 | -3.26 | -2.27 | .61 | -.70 | -.49 | .39 | -1.24 | 1.33 | .29 | -.58 | -.41 | .94 |
| **SEM (logits)** | .11 | .09 | .10 | .15 | .11 | .10 | .09 | .09 | .10 | .09 | .12 | .10 | .09 | .09 | .11 |
| **Discrimination (rpb)** | .14 | .43 | .29 | .40 | .41 | .37 | .17 | .40 | .49 | .30 | .23 | .46 | .27 | .38 | .29 |
| **Rasch infit (MNSQ)** | 1.16 | .97 | 1.07 | .87 | .92 | .98 | 1.23 | .99 | .86 | 1.09 | 1.08 | .92 | 1.13 | 1.02 | 1.05 |
| **Rasch outfit (MNSQ)** | 1.58 | .97 | 1.17 | .61 | .84 | 1.07 | 1.27 | 1.04 | .94 | 1.10 | 1.26 | .89 | 1.17 | 1.03 | 1.23 |
| | **B1.1.1** | **B1.1.2** | **B1.1.3** | **B1.2.1** | **B1.2.2** | **B1.2.3** | **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*** | 464 | 464 | 464 | 464 | 464 | 464 | 464 | 464 | 464 | 464 | 464 | 464 | 464 | 464 | 464 |
| **Difficulty (logits)** | .14 | .04 | -1.53 | -.74 | -.06 | .43 | -.15 | 1.02 | -.11 | -2.11 | -1.29 | -.40 | .81 | .85 | .46 |
| **SEM (logits)** | .10 | .10 | .12 | .10 | .10 | .10 | .10 | .11 | .10 | .14 | .11 | .10 | .11 | .11 | .10 |
| **Discrimination (rpb)** | .35 | .38 | .28 | .44 | .44 | .07 | .47 | .40 | .50 | .31 | .29 | .50 | .18 | .59 | .33 |
| **Rasch infit (MNSQ)** | 1.03 | 1.00 | 1.01 | .93 | .94 | 1.27 | .91 | .96 | .89 | .95 | 1.02 | .88 | 1.16 | .79 | 1.03 |
| **Rasch outfit (MNSQ)** | 1.02 | 1.00 | 1.09 | .89 | .93 | 1.36 | .89 | .97 | .85 | 1.05 | 1.12 | .85 | 1.24 | .73 | 1.06 |
| | **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** | **B2.4.1** | **B2.4.2** | **B2.4.3** | **B2.5.1** | **B2.5.2** | **B2.5.3** |
| ***N*** | 281 | 281 | 281 | 281 | 281 | 281 | 281 | 281 | 281 | 281 | 281 | 281 | 281 | 281 | 281 |
| **Difficulty (logits)** | -.60 | .80 | .87 | .00 | -.32 | 1.46 | 2.06 | -.73 | .91 | 1.68 | .96 | 2.49 | -.16 | .13 | 1.17 |
| **SEM (logits)** | .13 | .13 | .13 | .13 | .13 | .15 | .17 | .14 | .13 | .16 | .13 | .20 | .13 | .13 | .14 |
| **Discrimination (rpb)** | .15 | .46 | .19 | .48 | .52 | .34 | .29 | .40 | .27 | .03 | .26 | .08 | .40 | .40 | .38 |
| **Rasch infit (MNSQ)** | 1.13 | .91 | 1.11 | .89 | .85 | .97 | .98 | .94 | 1.07 | 1.20 | 1.06 | 1.09 | .96 | .96 | .97 |
| **Rasch outfit (MNSQ)** | 1.19 | .88 | 1.19 | .86 | .81 | .98 | 1.04 | .89 | 1.08 | 1.45 | 1.10 | 1.45 | .92 | .94 | .93 |
| | **C1.1.1** | **C1.1.2** | **C1.1.3** | **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** | **C1.5.1** | **C1.5.2** | **C1.5.3** |
| ***N*** | 93 | 93 | 93 | 93 | 93 | 93 | 93 | 93 | 93 | 93 | 93 | 93 | 93 | 93 | 93 |
| **Difficulty (logits)** | .55 | .75 | 1.76 | 1.07 | 1.07 | .31 | 1.62 | 1.42 | 1.55 | 1.07 | 1.30 | 1.12 | .65 | 2.09 | 1.42 |
| **SEM (logits)** | .22 | .22 | .27 | .23 | .23 | .22 | .26 | .25 | .26 | .23 | .24 | .24 | .22 | .30 | .25 |
| **Discrimination (rpb)** | .17 | .34 | .26 | .22 | .00 | .32 | -.19 | .21 | .32 | -.03 | .25 | .33 | .01 | .08 | .38 |
| **Rasch infit (MNSQ)** | 1.10 | .98 | 1.02 | 1.05 | 1.21 | 1.00 | 1.30 | 1.06 | .95 | 1.23 | 1.03 | .99 | 1.19 | 1.08 | .92 |
| **Rasch outfit (MNSQ)** | 1.11 | .97 | 1.01 | 1.11 | 1.27 | .97 | 1.59 | 1.04 | 1.02 | 1.30 | 1.04 | .95 | 1.28 | 1.37 | .95 |

**Table 9. Item Characteristics German**

| | **A1.1.1** | **A1.1.2** | **A1.1.3** | **A1.2.1** | **A1.2.2** | **A1.2.3** | **A1.3.1** | **A1.3.2** | **A1.3.3** | **A1.4.1** | **A1.4.2** | **A1.4.3** | **A1.5.1** | **A1.5.2** | **A1.5.3** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ***N*** | 109 | 109 | 109 | 109 | 109 | 109 | 109 | 109 | 109 | 109 | 109 | 109 | 109 | 109 | 109 |
| **Difficulty (logits)** | -1.29 | -.87 | -.74 | -2.78 | -1.19 | .35 | -1.43 | 2.40 | -1.53 | -2.88 | -2.52 | -.74 | .40 | -1.80 | .11 |
| **SEM (logits)** | .22 | .21 | .21 | .31 | .22 | .22 | .22 | .39 | .23 | .32 | .28 | .21 | .22 | .24 | .22 |
| **Discrimination (rpb)** | .33 | .40 | .51 | .34 | .22 | .24 | .14 | .43 | .30 | .27 | .31 | .38 | .40 | .37 | .39 |
| **Rasch infit (MNSQ)** | 1.03 | .97 | .85 | .89 | 1.12 | 1.15 | 1.15 | .96 | 1.02 | .95 | .94 | .99 | 1.00 | .95 | 1.01 |
| **Rasch outfit (MNSQ)** | .96 | .94 | .79 | .83 | 1.51 | 1.37 | 1.89 | .82 | 1.09 | .88 | .90 | .97 | 1.02 | .82 | .98 |
| | **A2.1.1** | **A2.1.2** | **A2.1.3** | **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** | **A2.5.1** | **A2.5.2** | **A2.5.3** |
| ***N*** | 183 | 183 | 183 | 183 | 183 | 183 | 183 | 183 | 183 | 183 | 183 | 183 | 183 | 183 | 183 |
| **Difficulty (logits)** | .22 | -1.60 | .22 | 1.17 | 1.29 | .83 | -2.19 | 1.33 | .22 | -.91 | -.66 | -1.89 | -1.43 | .67 | -1.43 |
| **SEM (logits)** | .17 | .19 | .17 | .19 | .20 | .18 | .21 | .20 | .17 | .17 | .17 | .20 | .18 | .18 | .18 |
| **Discrimination (rpb)** | .64 | .45 | .49 | .48 | .57 | .29 | .19 | .25 | .38 | .57 | .45 | .46 | .31 | .38 | .44 |
| **Rasch infit (MNSQ)** | .76 | .87 | .94 | .95 | .83 | 1.21 | 1.13 | 1.24 | 1.07 | .78 | .95 | .84 | 1.06 | 1.09 | .90 |
| **Rasch outfit (MNSQ)** | .67 | .73 | .92 | 1.01 | .76 | 1.29 | 1.17 | 1.42 | 1.16 | .68 | .92 | .64 | 1.10 | 1.15 | .83 |
| | **B1.1.1** | **B1.1.2** | **B1.1.3** | **B1.2.1** | **B1.2.2** | **B1.2.3** | **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*** | 122 | 122 | 122 | 122 | 122 | 122 | 122 | 122 | 122 | 122 | 122 | 122 | 122 | 122 | 122 |
| **Difficulty (logits)** | .06 | 1.05 | -.48 | .27 | -1.26 | -.21 | -1.45 | -.53 | -.25 | -.12 | 1.18 | 1.62 | -.98 | 1.32 | .61 |
| **SEM (logits)** | .21 | .21 | .22 | .21 | .24 | .21 | .26 | .22 | .21 | .21 | .21 | .23 | .23 | .22 | .21 |
| **Discrimination (rpb)** | .58 | .57 | .05 | .35 | .05 | .61 | .40 | .66 | .47 | .55 | .57 | .61 | .17 | .23 | .51 |
| **Rasch infit (MNSQ)** | .85 | .86 | 1.42 | 1.17 | 1.32 | .80 | .98 | .81 | .98 | .88 | .87 | .76 | 1.27 | 1.28 | .96 |
| **Rasch outfit (MNSQ)** | .76 | .88 | 2.10 | 1.16 | 2.96 | .71 | .74 | .66 | .91 | .82 | .84 | .77 | 1.63 | 1.58 | .95 |
| | **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** | **B2.4.1** | **B2.4.2** | **B2.4.3** | **B2.5.1** | **B2.5.2** | **B2.5.3** |
| ***N*** | 56 | 56 | 56 | 56 | 56 | 56 | 56 | 56 | 56 | 56 | 56 | 56 | 56 | 56 | 56 |
| **Difficulty (logits)** | -1.12 | .50 | .86 | -.85 | -.28 | .03 | -1.27 | .03 | 2.12 | -.28 | .22 | .95 | 1.04 | 1.41 | .41 |
| **SEM (logits)** | .38 | .30 | .30 | .36 | .33 | .31 | .40 | .31 | .33 | .33 | .31 | .30 | .30 | .31 | .30 |
| **Discrimination (rpb)** | .52 | .42 | .39 | .63 | .67 | .42 | .03 | .49 | .36 | .42 | .73 | -.13 | .49 | .49 | .49 |
| **Rasch infit (MNSQ)** | .88 | 1.05 | 1.09 | .75 | .73 | 1.03 | 1.28 | .97 | 1.04 | .98 | .66 | 1.67 | .91 | .92 | .95 |
| **Rasch outfit (MNSQ)** | .61 | 1.03 | 1.06 | .54 | .67 | 1.05 | 2.53 | .90 | .89 | 1.23 | .60 | 1.91 | 1.13 | .90 | .98 |
| | **C1.1.1** | **C1.1.2** | **C1.1.3** | **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** | **C1.5.1** | **C1.5.2** | **C1.5.3** |
| ***N*** | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 |

**Comments**

For all languages, the mean difficulty logic of all items was set to 0. Table 7 shows the item characteristics for Spanish. It demonstrates that the overall item difficulty increases with the sublevels tested as expected. The precision of the item difficulty parameter is high, as suggested by the low SEMs, varying from .08 to .13 at the IL level, from .07 to .12 at the IM level, from .07 to .14 at the AL level, from .07 to .12 at the AM level, and from .13 to .16 at the Superior level.

12 out of 75 items are below the threshold of .25 of item discrimination: one at the IL level, one at IM, two at AL, two at AM, and six at S. Infit statistics for all 12 of these items, however, are between .5 and 1.5, suggesting that the low discrimination values may be due to the fact that the test-takers were too homogenous. Five of the outfit statistics are above 1.5, suggesting that some of the low discrimination values may also be caused by outliers. All infit values are between .5 and 1.5, and many of them are close to 1.0, indicating a good overall item fit.

Table 8 shows the item characteristics for French. It reveals that the overall item difficulty increases with the sublevels tested as expected. The precision of the item difficulty parameter is high, as suggested by the low SEMs, varying from .11 to .19 at the IL level, from .09 to .15 at the IM level, from .10 to .14 at the AL level, from .13 to .20 at the AM level, and from .22 to .30 at the Superior level. At the Superior level, the SEMs are higher due to the smaller sample size.

16 out of 75 items are below the threshold of .25 of item discrimination: three at the IM level, two at the AL level, three at the AM level, and eight at the Superior level. Infit statistics for all of these 16 items, however, are between .5 and 1.5, suggesting that the low discrimination values may be due to the fact that the test-takers were too homogenous. Only two of the outfit statistics are above 1.5, and only barely so, indicating that some of the low discrimination values may also be caused by outliers. All infit values are between .5 and 1.5, and many of them are close to 1.0, indicating a good overall item fit.

Table 9 shows the item characteristics for German. There were only 10 test-takers at the Superior level, too few to calculate any meaningful statistics. Item characteristics, therefore, are provided only for the 60 items covering the levels from IL to AM. Table 9 shows that the overall item difficulty increases with the sublevels tested as expected. The precision of the item difficulty parameter, again, is high, as suggested by the low SEMs, varying from .21 to .39 at the IL level, from .17 to .21 at the IM level, from .21 to .26 at the AL level, and from .30 to .40 at the AM level. The higher values at the AM level are due to the smaller sample size.

10 out of 60 items are below the threshold of .25 of item discrimination: three at the IL level, one at IM, four at AL, and two at AM. Infit statistics for nine of the ten items are between .5 and 1.5, suggesting that the low discrimination values may be due to the homogeneity of the test-takers. One infit item is slightly above 1.5, i.e. 1.67, suggesting that a revision of the item should be considered. Eight of the outfit statistics are above 1.5, and three of these are above 2.0, in-

dicating that the low discrimination values may also be caused by outliers. 74 out of 75 infit values are between .5 and 1.5, and many of them are close to 1.0, indicating a good overall item fit.

## 10. Reliability Information

As suggested by AREA/APA/NCME (2014: 46), both, the overall and conditional standard errors of measurement (SEM) are considered central indicators of test reliability. In the current section, the overall SEM is reported for the whole test, while Rasch person ability estimates as well as conditional SEMs are reported for the four score options of two sublevels each that may be tested independently (see Table 10).

The Rasch person separation reliability was calculated for the whole test as another reliability measure. The Rasch person separation reliability can be considered equivalent to Cronbach's alpha. The Rasch person separation reliability, however, is sample independent and tends to underestimate the true reliability, whereas classical measures such as Cronbach's alpha tend to overestimate the true reliability.

As suggested by AREA/APA/NCME (2014: 38), the Rasch test information function for each modality and language is reported as further evidence of test reliability (see Figure 1 below).

Since no information on specific subgroups of test-takers is available, reliability estimates could not be computed for subgroups.

**Table 10. Reliability Estimates of the ACTFL Listening Proficiency Test (LPT)**

| | *N* | Overall SEM | Rasch Separation Reliability | Conditional SEM | | | |
|---|---|---|---|---|---|---|---|
| | | | | A1/A2 | A2/B1 | B1/B2 | B2/C1 |
| Spanish | 1884 | .43 | .83 | .45<br>(*N* = 735) | .46<br>(*N* = 419) | .45<br>(*N* = 816) | .43<br>(*N* = 276) |
| French | 816 | .42 | .81 | .45<br>(*N* = 363) | .43<br>(*N* = 270) | .43<br>(*N* = 241) | .42<br>(*N* = 93) |
| German | 239 | .45 | .84 | .46<br>(*N* = 109) | .47<br>(*N* = 75) | .45<br>(*N* = 47) | n.a. * |

* Not enough cases to calculate a meaningful SEM or meaningful difficulty estimates.

Table 10 shows that the overall Rasch person separation reliability is very high for all languages. The large majority of test-takers took tests consisting of 30 items. The smallest SEM value possible for a test with 30 items is .37. The observed overall SEMs are only marginally higher than that, indicating a high degree of reliability for the number of items used. The conditional SEMs are equally low. All measures reported in this table, therefore, provide evidence that the RPT has a high degree of reliability.

This conclusion is corroborated by the overall Rasch item fit statistics in Table 11 (see Section 9 for item fit statistics for individual items).

**Table 11. Overall Rasch Fit Statistics**

| | ***N*** | **Rasch Item Infit (MNSQ)** | **Rasch Item Outfit (MNSQ)** |
|---|---|---|---|
| Spanish | 1884 | 1.00 | 1.07 |
| French | 816 | 1.00 | 1.03 |
| German | 239 | 1.00 | 1.05 |

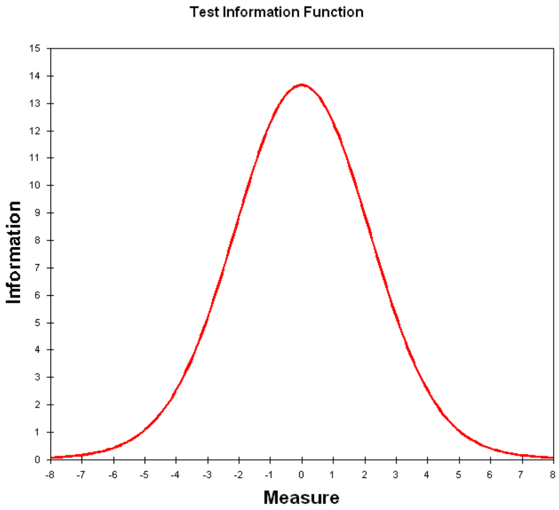
Table 11 shows that the items generally produce exactly the same amount of infit variance that is expected from the Rasch model. Outfit values are slightly higher than the infit values, yet still very close to the ideal variance range. The Rasch fit statistics, thus, add another piece of evidence that the measurement functions as desired.

Further evidence comes from an analysis of the test information function for each language. The test information is the aggregated Fischer information of the test across all items. The Fischer information of an item is equal to the probability that a person with a given ability level will answer this item correctly multiplied by its counter-probability. The optimal information possible is reached when probability and counter-probability are 1:1, or 50% each, respectively. In that case, the information is .25. The following test information functions show for which competence ranges the test yields the most information (see Figure 1).

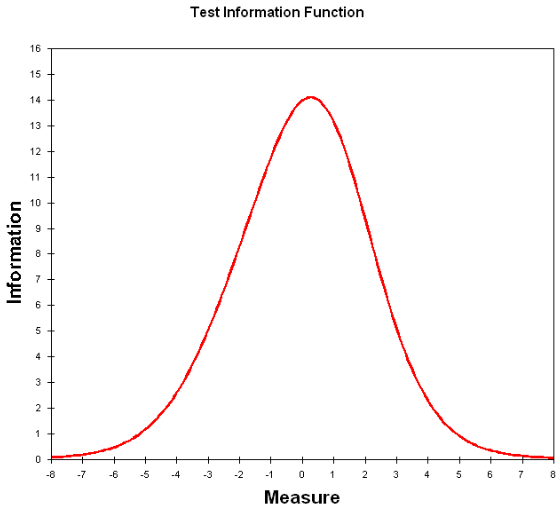
Figure 1 shows that all three test information functions have their peak in the middle of the competence range, i.e., close to the person ability of .0 logits. The most information is generally collected in the range between –3 and +3 logits. This is exactly the ability range for which the test was designed. Therefore, the test information function, too, supports the conclusion that the ACTFL Listening Proficiency Test provides reliable results.

### Figure 1. Test Information Functions for Spanish, French, and German

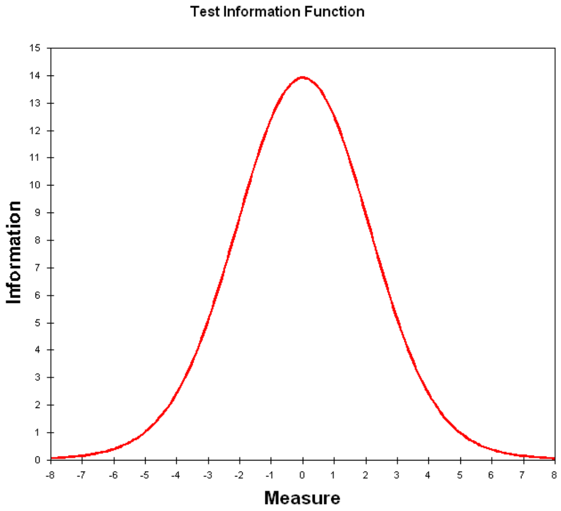
### Spanish Listening Proficiency Test



### French Listening Proficiency Test



### German Listening Proficiency Test



## 11. Scorer Reliability for Essay Items

Not applicable

## 12. Errors of Classification When Single or Multiple Cut-scores Are Used

Table 12 shows the logits and their respective SEM of all cut-scores distinguishing between the ACTFL levels of the LPT (see Appendix 7 for logits and SEMs for all scores from 1 to 75 for Spanish, French, and German). Cut-score logits and SEM are based on the assumption of a test-taker responding to all 75 items of a complete test.

**Table 12. Cut-score Logits and SEMs for All ACTFL Levels by Language**

| | | Spanish | | French | | German | |
|---|---|---|---|---|---|---|---|
| **ACTFL** | **Cut-score** | **Logit** | **SEM** | **Logit** | **SEM** | **Logit** | **SEM** |
| NL | below 12 | | | | | | |
| NM | 12 | -2.19 | .35 | -2.18 | .36 | -2.15 | .35 |
| NH | 15 | -1.85 | .32 | -1.82 | .33 | -1.81 | .32 |
| IL | 18 | -1.55 | .31 | -1.50 | .31 | -1.52 | .31 |
| IM | 24 | -1.02 | .29 | -.96 | .29 | -1.00 | .28 |
| IH | 37 | -.03 | .27 | .03 | .27 | -0.03 | .27 |
| AL | 48 | .79 | .28 | .82 | .27 | .77 | .28 |
| AM | 54 | 1.28 | .30 | 1.28 | .29 | 1.26 | .29 |
| AH | 67 | 2.75 | .41 | 2.66 | .39 | 2.70 | .40 |
| S | 69 | 3.12 | .45 | 3.00 | .44 | 3.06 | .45 |

Table 12 shows that the SEM is low for all sublevel cut-points and languages. The logits are very similar across all three languages. The most important levels for assigning college credits are the ACTFL levels IL to AM because these are the proficiency levels of the great majority of high school and college students. At these levels, SEMs range from .27 to .31 for Spanish, .27 to .31 for French, and .27 to .31 for German. The largest SEM for all three languages is .45 at the Superior level.

## 13. Evidence of Validity: Content-related

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 test-takers. In the following, three examples of different two-level tests are presented to show how the texts reflect the *ACTFL Proficiency Guidelines*, and how the test ensures the selection of a diverse and representative sample of the topics, subtopics, genres, and rhetorical organization of texts listeners are able to comprehend at each level.

**Example 1** represents a test that spans the sublevels NL to IM. Texts 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%. Text topics, subtopics, genres and rhetorical organization are based on the ACTFL level descriptions (see below for IL and IM). Table 13 shows the variety and distribution of topics, subtopics, genres and rhetorical organization in a typical NL to IM test.

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

| Task | Topic | Subtopic | Genre | Rhetorical Organization |
|---|---|---|---|---|
| IL.1 | Free Time | Shopping | Advertisement | Instruction |
| IL.2 | Food | Restaurant | Simple Text | Description |
| IL.3 | Family | People | Personal Correspondence | Description |
| IL.4 | Daily Life | Pets | Simple Text | Instruction |
| IL.5 | Arts | Theater | Advertisement | Description |
| IM.1 | Daily Life | Routine | Report | Description |
| IM.2 | Sports | Plans | News Item | Narration |
| IM.3 | Daily Life | Moving | Narrative | Narration |
| IM.4 | Work | Routine | Narrative | Narration |
| IM.5 | Society | Literature | Advertisement | Description |
| Distribution | 3x Daily Life<br>1x Free Time<br>1x Food<br>1x Family<br>1x Arts<br>1x Sports<br>1x Work<br>1x Society | 1x Shopping<br>1x Restaurant<br>1x People<br>1x Pets<br>1x Theater<br>2x Routine<br>1x Plans<br>1x Moving<br>1x Literature | 3x Advertisement<br>2x Simple Text<br>1x Personal Correspondence<br>1x Report<br>1x News Item<br>2x Narrative | 2x Instruction<br>5x Description<br>3x Narration |

**Example 2** represents a test that spans the sublevels IM to AM (see Table 14). Texts 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. Text topics, subtopics, genres and rhetorical organization are based on the ACTFL level descriptions (see below for AL and AM). Table 14 shows the variety and distribution of topics, subtopics, genres and rhetorical organization in a typical IM to AM test.

### Advanced Low

At the Advanced Low sublevel, listeners are able to understand short conventional narrative and descriptive texts with a clear underlying structure though their comprehension may be uneven. The listener understands the main facts and some supporting details. Comprehension may often derive primarily from situational and subject-matter knowledge.

### Advanced Mid

At the Advanced Mid sublevel, listeners are able to understand conventional narrative and descriptive texts, 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.

**Table 14. 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 | News Item | Narration |
| AL.2 | Daily Life | People | Report | Narration |
| AL.3 | Work | Children | Personal Correspondence | Narration |
| AL.4 | Travel | Money | Giving Advice | Explanation |
| AL.5 | Travel | Trips | Personal Correspondence | Description |
| AM.1 | Society | People | Report | Exposition |
| AM.2 | Education | School | Report | Exposition |
| AM.3 | Government and Politics | Plans | Op-Ed | Argument |
| AM.4 | Arts | Cinema | Op-Ed | Argument |
| AM.5 | Society | Tradition | Report | Exposition |
| Distribution | 3x Society<br>1x Daily Life<br>1x Work | 1x Trends<br>2x People<br>1x Children | 1x News Item<br>2x Personal Correspondence | 1x Explanation<br>3x Narration<br>1x Description |

| | | | | |
|---|---|---|---|---|
| | 2x Travel<br>1x Education<br>1x Government and politics<br>1x Arts | 1x Money<br>1x Trips<br>1x School<br>1x Plans<br>1x Cinema<br>1x Tradition | 4x Report<br>1x Giving Advice<br>2x Op-Ed | 3x Exposition<br>2x Argument |

**Example 3** represents a test that spans the sublevels IH to S (see Table 15). Texts 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. Text topics, subtopics, genres and rhetorical organization are based on the ACTFL level descriptions (see below for AM and S). Table 15 shows the variety and distribution of topics, subtopics, genres and rhetorical organization in a typical IH to S test.

**Advanced Mid**

At the Advanced Mid sublevel, listeners are able to understand conventional narrative and descriptive texts, 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.

**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 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. Superior listeners can understand not only what is said, but sometimes what is left unsaid; that is, they can make inferences.

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 15: 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 | Journal Article | Exposition |
| AM.2 | Education | School | Report | Narration |
| AM.3 | Government and Politics | Plans | Op-Ed | Argument |
| AM.4 | Arts | Cinema | Op-Ed | Argument |
| AM.5 | Society | Tradition | Report | Exposition |
| S.1 | Business & Commerce | Money | Advertisement | Exposition |
| S.2 | Government and Politics | Reform | News Item | Argument |
| S.3 | Food | Trends | Advertisement | Narration |
| S.4 | Technology | Reform | Review | Exposition |
| S.5 | Science | Problems | Report | Argument |
| Distribution | 2x Society<br>1x Education<br>2x Government and Politics<br>1x Arts<br>1x Business & Commerce<br>1x Food<br>1x Technology<br>1x Science | 1x People<br>1x School<br>1x Plans<br>1x Cinema<br>1x Tradition<br>1x Money<br>2x Reform<br>1x Trends<br>1x Problems | 1x Jounral Article<br>3x Report<br>2x Op-Ed<br>2x Advertisement<br>1x News Item<br>1X Review | 4x Exposition<br>4x Argument<br>2x Narration |

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 test-takers.

## 14. Evidence of Validity: Criterion-related (*See Appendix 6 – Technical Report*)

The ACTFL LPT is based on standardized criteria taken from the *ACTFL Proficiency Guidelines 2012 – Listening*. The test was externally validated by a side-by-side study of the ACTFL LPT with NATO's Benchmark Advisory Test – Listening (BAT-L). That study also summarizes and explains the internal validity studies completed for every single form.

This 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

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

**Design**

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

**Statistical Analysis**

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

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 16 displays all measures that were computed to establish the ACTFL LPT's *external validity*. It contains four parameters, which indicate the relationship between the ACTFL LPT and the BAT-L. Two correlation and two agreement measures were computed. Both correlation parameters, Pearson's $r_s$ and Spearman's *rho*, show a high interdependence between the two tests. As for the agreement measures, Kendall's *tau* is obviously affected by bindings in the data and 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 16. Correlation and Agreement Measures Between Final Ratings of the ACTFL LPT and the BAT-L**

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

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

The frequency distribution in Table 17 below also points to a strong relationship between the two tests and corroborates the correlation parameters and agreement measures reported in Table 16.

**Table 17. Frequency of Agreement in Final Ratings of the ACTFL 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.

As to the nature of the correspondence between LPT and BAT-L, Table 17 above shows the following: IL corresponds to STANAG/ILR 0+ 40% and to STANAG/ILR 1 60% of the time. IM corresponds to STANAG/ILR 1 57% and to STANAG/ILR 1+ or higher 42% per cent of the time. AL corresponds to STANAG/ILR 1+ or lower 32% and to STANAG/ILR 2 66% of the time. AM corresponds to STANAG/ ILR 2 or lower 28% and to STANAG/ILR 2+ or higher 72% of the time. S corresponds to STANAG/ ILR 3 62% of the time.

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

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

## 15. Evidence of Validity: Construct

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

The first piece of evidence comes from a two-day standard-setting workshop, which was conducted with the German LPT in July 2015. Eight experts with a college degree in German as a Foreign Language and with broad experience teaching and testing German as a Foreign Language participated in the study, one of them male and seven female. The experts were asked to judge each of the 75 items of one form of the German LPT whether a borderline candidate at a specific competence level would be able to answer test items at his or her competence level correctly (Modified Angoff Technique).

The standard-setting workshop consisted of three different phases: the familiarization phase, the calibration phase, and the standard-setting phase itself. In the initial familiarization phase, the experts ordered relevant competence descriptors in small groups and discussed their results. They subsequently discussed the salient features of the relevant proficiency levels. The overall aim of the familiarization phase, which lasted 90 minutes, was to create a shared understanding of the proficiency scale and the test construct.

In the calibration phase, participants individually applied their understanding of the listening proficiency construct to ten listening tests of German as a Foreign Language with calibrated difficulties (the tests included tests from the Goethe Institute, The European Language Certificates/telc, and Test-DaF). In the concluding discussion, participants provided an account of their judgments. There was high agreement among the participants with respect to the proficiency levels of the tests rated. The calibration phase lasted 90 minutes. The results of the calibration phase provided ample evidence of the rating reliability and agreement of the experts being sufficiently high to provide reliable judgments in the standard-setting phase.

The standard-setting phase lasted 240 minutes. Participants were first asked to listen to a passage from the German LPT and read its related items. They then judged whether a borderline candidate would be able to answer each of the three items correctly. Participants were also asked to indicate on a four-point Likert scale how confident they were of their rating. At the bottom of their rater sheets, they had ample space to comment on the passage, the items, and the rating process. Passages and items were ordered in two different ways: one set started with the easiest passages and continued to the more difficult ones, and the other set started with the most difficult passages and continued to the easier ones. This was intended to mitigate ordering effects. After the participants had judged all 75 test items, they were asked to comment on the rating process on a separate sheet.

Table 18 presents the results of the standard setting for each individual item. The first line contains the mean participant agreement on whether a borderline candidate would answer the item correctly ("yes" was coded "1", "no" was coded "0"). The second line represents the standard deviation of the agreement measures.

A rater agreement of .5 and higher indicates that the majority of raters believed that the item matches the test construct of a particular sublevel. As Table 18 shows, there were 8 out of 72 cases, where the expert raters judged an item to be too difficult for the targeted proficiency level; in all other cases, the raters agreed with the level the item was supposed to target. This finding can be considered as clear evidence of the alignment of the test with the construct matrix and proficiency scale.

**Table 18. Results of the Standard-setting Workshop of the German LPT**

| | **A1.1.1** | **A1.1.2** | **A1.1.3** | **A1.2.1** | **A1.2.2** | **A1.2.3** | **A1.3.1** | **A1.3.2** | **A1.3.3** | **A1.4.1** | **A1.4.2** | **A1.4.3** | **A1.5.1** | **A1.5.2** | **A2.1.3** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ***N*** | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 |
| **Agreement** | 0.63 | 0.75 | 0.50 | 0.75 | 1.00 | 0.63 | 1.00 | 0.88 | 0.38 | 1.00 | 1.00 | 0.63 | 1.00 | 0.38 | 0.88 |
| **Standard Deviation** | 0.52 | 0.46 | 0.53 | 0.46 | 0.00 | 0.52 | 0.00 | 0.35 | 0.52 | 0.00 | 0.00 | 0.52 | 0.00 | 0.52 | 0.35 |
| | **A2.1.1** | **A2.1.2** | **A2.1.3** | **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** | **A2.5.1** | **A2.5.2** | **A2.5.3** |
| ***N*** | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 |
| **Agreement** | 0.75 | 0.75 | 1.00 | 0.75 | 1.00 | 0.38 | 0.75 | 0.75 | 1.00 | 0.88 | 1.00 | 0.50 | 0.75 | 1.00 | 0.88 |
| **Standard Deviation** | 0.46 | 0.46 | 0.00 | 0.46 | 0.00 | 0.52 | 0.46 | 0.46 | 0.00 | 0.35 | 0.00 | 0.53 | 0.46 | 0.00 | 0.35 |
| | **B1.1.1** | **B1.1.2** | **B1.1.3** | **B1.2.1** | **B1.2.2** | **B1.2.3** | **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 | 8 | 8 | 8 | 8 | 8 | 8 |
| **Agreement** | 0.88 | 0.38 | 0.50 | 0.75 | 0.13 | 0.88 | 0.88 | 0.75 | 0.63 | 0.88 | 1.00 | 0.63 | 0.38 | 1.00 | 0.50 |
| **Standard Deviation** | 0.35 | 0.52 | 0.53 | 0.46 | 0.35 | 0.35 | 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** | **B2.4.1** | **B2.4.2** | **B2.4.3** | **B2.5.1** | **B2.5.2** | **B2.5.3** |
| ***N*** | 8 | 8 | 8 | 8 | 8 | 8 | 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 | 0.88 | 0.75 | 0.88 | 1.00 | 0.75 | 0.75 |
| **Standard Deviation** | 0.00 | 0.35 | 0.53 | 0.46 | 0.46 | 0.35 | 0.53 | 0.53 | 0.46 | 0.35 | 0.46 | 0.35 | 0.00 | 0.46 | 0.46 |
| | **C1.1.1** | **C1.1.2** | **C1.1.3** | **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** | **C1.5.1** | **C1.5.2** | **C1.5.3** |
| ***N*** | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 |
| **Agreement** | 1.00 | 0.63 | 0.75 | 0.88 | 1.00 | 0.38 | 0.63 | 0.75 | 0.75 | 1.00 | 0.75 | 0.75 | 0.75 | 0.88 | 1.00 |
| **Standard Deviation** | 0.00 | 0.52 | 0.46 | 0.35 | 0.00 | 0.52 | 0.52 | 0.46 | 0.46 | 0.00 | 0.46 | 0.46 | 0.46 | 0.35 | 0.00 |

### Rasch Model Fit

The second piece of evidence for 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 can be interpreted as an indicator of the validity of the model, i.e., construct validity. Table 19 provides Rasch person infit and outfit values for each of the languages.

**Table 19. Rasch Person Infit and Outfit Values**

| | *N* | Rasch Person Infit (MNSQ) | Rasch Person Outfit (MNSQ) |
|---|---|---|---|
| Spanish | 1884 | .99 | 1.02 |
| French | 816 | .99 | 1.01 |
| German | 239 | 1.00 | 1.05 |

As Table 19 shows, the data fit the model impressively well. This provides evidence that the test allows predicting a test-taker's performance in the test to a very high degree. This provides strong evidence for the construct validity of the test.

## 16. Rationale for the Particular Cut-score Recommended *(See Appendix 2 – Assessment Use Argument and Appendix 6 – Technical Report)*

Because the LPT is a high stakes test, false positive classification decisions are considered to be relatively more serious than false negative classification errors. Cut-scores were determined empirically. To avoid false positive classification decision errors, cut-scores were set at the upper end of the cut-score range determined by the calibration study.

## 17. Evidence for the Reasonableness and Appropriateness of the Cut-score Recommended

Two sources of evidence for the reasonableness and appropriateness of the cut-score are available: a side-by-side study between the LPT and NATO's Benchmark Advisory Test – Listening (BAT-L), providing evidence for the alignment of the test with an external criterion; and the results of a standard-setting workshop relying on expert judgments.

In the side-by-side study, cut-scores were verified empirically, using an external criterion calibrated against the same competence scale (see Section 14 above).

In addition, cut-scores were verified using another type of empirical data, the results of a standard-setting workshop relying on expert judgment (see Section 15 above). Table 20 displays the mean agreement of the expert judges across all items of the main proficiency sublevels of the test.

**Table 20. Mean Rater Agreement on the Cut-scores of the German RPT**

| | *N* | Cut-Score IL | Cut-Score IM | Cut-Score AL | Cut-Score AM | Cut-Score S |
|---|---|---|---|---|---|---|
| German | 8 | .76 (SD* = .32) | .81 (SD = .30) | .68 (SD = .39) | .73 (SD = .39) | .79 (SD = .34) |

*SD = Standard Deviation

As Table 20 shows, the cut-scores as estimated in the standard-setting workshop using the Modified Angoff approach are consistently in the range between .73 and .81 except for AL where the cut-score is slightly below .70. Because it seems safe to assume that a test-taker has to answer at least 70% of the items of any proficiency sublevel correctly to be placed at this sublevel, the expert judgments as provided by the standard-setting workshop provide further evidence of the reasonableness and appropriateness of the cut-scores recommended on the basis of the side-by-side study.

## 18. Procedures Recommended to Users for Establishing Their Own Cut-scores (e.g., Granting College Credit)

The ACTFL LPT is used for classification purposes. In line with the college credit recommendations for the ACTFL OPI/OPIc and WPT, the following cut-score ACTFL sublevels are recommended for granting college credit (see Table 21).

**Table 21. Cut-score Recommendation for Granting College Credit**

| Official ACTFL LPT Rating | **Category I**<br>English, French, Italian, Spanish, Portuguese | **Category II**<br>German | **Category III**<br>Russian | **Category IV**<br>Arabic, Japanese, Korean, Mandarin |
|---|---|---|---|---|
| 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

## 19. Possible test bias

Two main aspects for possible test bias are gender-based and culture-based bias. Therefore, great care is given to use topics and develop items that have equal appeal to both genders. Items are developed and reviewed equally by female and male authors to avoid gender-based bias.

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

## 20. Information on Norms and Normative Groups (If Appropriate)

Not applicable

## 21. Evidence that the Time Limits are Appropriate and That the Exam is not Unduly Speeded

To determine if the time limits are appropriate and the exam is not unduly speeded, the time it took test-takers to finish the test was examined. The maximum amount of time provided to test-takers for the standard two-sublevel test is 50 minutes. Table 22 shows the minimum, maximum, mean, and standard deviation of the time in minutes it took test-takers to take the test per language. In addition it shows the maximum time in minutes it took 95 per cent of the test-takers and the percentage of test-takers who used the full 50 minutes.

**Table 22. 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 50 minutes.**

| Language | *N** | Min. | Max. | Mean | SD | 50 min |
|---|---|---|---|---|---|---|
| Spanish | 1776 | 19 | 50 | 28.53 | 5.29 | .2% |
| French | 753 | 20 | 50 | 31.00 | 5.62 | .4% |
| German | 238 | 20 | 50 | 27.50 | 5.63 | .0% |

*The *N* is slightly different from the *N* in other tables such as Table 19 due to the fact that some students took tests that spanned more than two sublevels and because the cut dates of the periods examined were slightly different.

Table 22 shows that very few test-takers took the full 50 minutes. Less than 99 per cent of test-takers in French and Spanish took the full 50 minutes and no one did so in the case of German. This can be taken as evidence that the time limits are appropriate and that the test is not unduly speeded.

## 22. Provisions for Standardizing Administration of the Examination *(See Appendix 2 – Assessment Use Argument and Appendix 10 – Proctor Manual)*

Impartial treatment of test-takers during all aspects of the administration of the LPT from registering for the assessment to taking the assessment is ensured by the strict adherence to the regulations below.

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

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

The LPT is delivered over the Internet based on the same test algorithm each time and is accessible to test-takers in any part of the world where there is reliable Internet availability.

The LPT is a machine-scored test performed on the computer. Official ACTFL LPT ratings are assigned to those LPTs that are conducted under the supervision of LTI. Persons supervising the test treat all test-takers impartially following the procedures described in the Proctor Manual.

## 23. Directions for Scoring *(See Appendix 4 – Blueprint)*

The ACTFL LPT is scored automatically, using the cut-scores discussed below. To assign ratings, 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 a 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. The ratings are assigned as follows (see Table 23):

**Table 23. Scoring Algorithm**

| Sublevels | 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 23 shows what ratings are assigned to what scores. Two levels are rated together. When more than two levels are administered, the highest two levels that have at least 18 points (or 11 points, respectively) are used. BR (Below Rating) is assigned to a total score of 0-11, because such scores may be achieved by guessing. Novice Low is assigned to a total score of 0-11 when evaluating the sublevels IL and IM.

## 24. Provisions for exam security

The proctor, nominated by the organizing agency, will sign a form and provide it to LTI in advance of the assessment, undertaking to guarantee the identity of the candidate and the conditions under which the test is taken.

To ensure connectivity and full operational status, the System Check page ensures that the computer over which the test will be delivered is set to support the test. After the System Check page, there is a Login page requiring a login and password. LPT logins and passwords are created by a proctor on a secure LTI client site. The client/proctor also chooses the range of the test (and corresponding length of the test). Once created, the login and password is valid for two weeks, after which time the login expiration date can be extended by the proctor on the LTI client site. If the date is not extended, the login and password will become invalid and a test-taker trying to enter an expired login and password will receive an "invalid login" message.

Test-takers should not try to open any other windows, browsers or pop-ups while in the test. If a test-taker clicks outside of the test, the test will automatically shut down and the test-taker will need to log in to the test again. Test-takers are allowed three attempts to access the test; further login attempts will fail.

## 25. Information on the Currency and Representativeness of the Exam's Items

One way of ensuring the currency of the exam's items is the way texts and items are written. They are written in such a way that they cannot be easily dated. In addition, the item life cycle is carefully monitored. Items are regularly reviewed and outdated items are updated or retired.

The representativeness of the texts and items in a test is guaranteed by providing a diversity of topics, subtopics, genres, domains and rhetorical organization so that the test can provide ample evidence of the listening proficiency of the test-taker across a broad spectrum of target language use domains (see Section 7, Tables 5 and 6).

## 26. Scoring keys

Not applicable

## 27. Equivalence of Forms

All tasks and items are calibrated on the same metric using Rasch statistics (model for dichotomous items). Fifteen anchor items from the first test form are used in all subsequent forms. By means of common item equating using the WINSTEPS software, the difficulty of new test items is determined with high precision.

In addition, the equivalence of forms is ensured by the use of a comprehensive construct matrix, the rigorous training of test authors, and revisions informed by extensive psychometric analyses. Item difficulty is continually monitored to provide evidence for comparable difficulty levels across languages.

## 28. Other relevant information

### Item Development Process and Training of Test Authors and Reviewers

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, training and certification process as well as ongoing quality assurance appropriate for high stakes testing.

The training of test authors and reviewers constitutes an integral part of the Item Development Process. The Institute for Test Research and Test Development Leipzig regularly arranges Item Writing Workshops consisting of several training sessions (one and two day workshops). The objective of the workshops is to train test authors and calibrate them with calibrated texts and items. Workshop facilitators are ACTFL-trained and certified Tester Trainers. During these workshops participants are familiarized with the Construct Matrix, the Item Writing Manual, and the Item Checklist while working individually and in groups. The workshop agenda includes the following activities: Sort the ACTFL Listening Proficiency Descriptors according to their proficiency levels; Complete the Construct Matrix with missing descriptors; Take an LPT to become familiar with the test; Get introduced to the Item Writing Manual and to Item Writing Do's and Don'ts, Get calibrated by benchmarking calibrated tests individually and in small groups, Write first drafts of items; and Take part in group discussions. After the workshop, there is a practice round and a certification round, in which participants author at least two texts and two sets of items at each sublevel, receive feedback on them, and get certified after passing the final review by a senior test development officer.

Items are developed in multiple stages in a controlled process. Certified authors and native speakers of the target language develop texts 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 structured, 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. The main focus is on the level appropriateness of the passages 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. The items are checked for spelling and punctuation before uploading the test to the LTI Assessment System. A final spelling/typing and functionality check is conducted once the test is online. The test then enters the piloting phase with at least 100 test-takers at all proficiency levels taking the test. Detailed data reports are developed using traditional (Cronbach's alpha, difficulty and separation indexes) and Rasch analysis (separation reliabilities, model fit, misfitting items). Any misfitting items are revised or discarded. If the report determines that the test form follows all requirements of a high stakes test, the test will be moved to operational testing.