Empirical Validation of Listening Proficiency Guidelines

Troy L. Cox  
*Brigham Young University*, troyc@byu.edu

Ray Clifford  
*Brigham Young University*

Follow this and additional works at: [https://scholarsarchive.byu.edu/facpub](https://scholarsarchive.byu.edu/facpub)

Part of the Arts and Humanities Commons

**Original Publication Citation**

**BYU ScholarsArchive Citation**
[https://scholarsarchive.byu.edu/facpub/5879](https://scholarsarchive.byu.edu/facpub/5879)

This Peer-Reviewed Article is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in Faculty Publications by an authorized administrator of BYU ScholarsArchive. For more information, please contact ellen_amatangelo@byu.edu.
Empirical Validation of Listening Proficiency Guidelines

Troy L. Cox
Brigham Young University

Ray Clifford
Brigham Young University

Abstract: Because listening has received little attention and the validation of ability scales describing multidimensional skills is always challenging, this study applied a multistage, criterion-referenced approach that used a framework of aligned audio passages and listening tasks to explore the validity of the ACTFL and related listening proficiency guidelines. Rasch measurement and statistical analyses of data generated in seven separate language studies resulted in significant differences in listening difficulty between the proficiency levels tested and confirmed the validity of the ACTFL proficiency assessment for listening.

Keywords: ACTFL Proficiency Guidelines, multistage assessment, proficient listening, scale validation, testing listening

Introduction
The Interagency Language Roundtable (ILR) language proficiency scales (2012) grew out of the U.S. government’s pragmatic need to appropriately assign language learners with varying ability levels to jobs of varying linguistic difficulty. By using a standard set of ability descriptions, the personnel office could thus make overseas assignments and know that a Level 1 speaker (Intermediate) would have sufficient language skills to survive in the target culture, a Level 2 (Advanced) speaker would have limited working proficiency and could be employed in areas in which he or she would be interacting with the target culture, and a Level 3 (Superior) would be able to function professionally with educated speakers in the target culture. One of the characteristics of the scales was to functionally describe, in a way that laypeople could understand, what language learners could do in real-world contexts.

This practical view to language learning and eventual language use led to the first widespread deployment of speaking assessments that were used to match civil servants with various assignments throughout the world. However, language...
consists of more than speaking, and it soon became apparent that there was a need to describe language use for the other skills, particularly because progress toward native-like proficiency in listening does not always progress in the same way and at the same rate. For example, in situations where learners have rehearsed the language they may need, they can produce utterances that they would not necessarily understand when listening to speakers using natural language in the target culture. Similarly, in other situations, learners can understand language at a much higher level than the speech they can produce. Thus, while closely related to proficiency scales and assessments for speaking, the listening proficiency scale was developed to describe listener functions apart from speaking in a hierarchical, criterion-referenced manner. The ability to measure listening has important implications ranging from the pragmatic need to monitor communication for national security to the lofty ideal of understanding another culture as goodwill ambassadors.

However, the listening proficiency scale, along with other proficiency scales, has long been under attack in academia. The genesis of these criticisms can be found in the fact that the scales were never intended to model language acquisition, are agnostic in their view of language theories, and focus instead on real-world functional language use. Thus, academics intent on developing assessments that operationalize theoretical constructs of acquisition have concluded that the scales violate a tenet of sound assessment practice. Namely, they believe that not having a theoretical framework of language learning that defines a language ability construct a priori to constructing a test weakens any validity claims the resulting scale might have (Bachman, 1990; Bachman & Palmer, 1996). Even those sympathetic to using the proficiency scale with the productive skills of speaking and writing may harbor some hesitations in applying the scales to the receptive skills. This reluctance often comes from experiences in which researchers observe the apparent confounding affect that prior content knowledge has on comprehension.

These concerns are addressed in this study in the following ways. First, an alternative, criterion-referenced validity framework is introduced in which item development can be used as the basis for validating a construct. Then, listening proficiency scales are discussed, followed by a careful operationalization of the listening construct as the basis for item development. Finally, an overview of how this process was used in seven different languages is presented in an attempt to answer the following research question: “To what extent do test items ascend in a hierarchy of difficulty levels when both the listening passage and the comprehension question are based on the ACTFL listening proficiency guidelines?”

Background

Constructing Measures

When operationalizing a construct such as listening proficiency, it is often helpful to look at the process as consisting of four building blocks: construct, items, item scores, and measures (Wilson, 2005). When a well-defined construct exists, one can validate the construct assumptions by writing items based on the construct, scoring the items, and then converting those item scores into measures (see Figure 1) through the use of a psychometric model that has the quality of invariant measurement. In this study, the psychometric model used to determine if the proficiency guidelines ascend hierarchically is the Rasch model of Item Response Theory (IRT).

Instead of starting the validation process with a theory of how language is acquired, this model assumes a construct map (see Figure 2) that has the quality of invariant measurement (Wilson, 2009). Along the map is a single, vertical axis that progresses from no ability to a great amount of ability. One side of the axis represents people and their ability level. There is the assumption that the trait being measured progresses
linearly, with people having more ability located higher on the axis and subsuming all the skills of those who are lower on the axis. The other side of the axis represents items. There is the assumption that some items are more difficult than other items. The items that are higher on the axis require competency on all the lower items in order for a person to successfully complete the task. The measures used to describe the location of the people and items on the axis are referred to as logits, or log odds ratios. Instead of being curvilinear, like probability, logits are interval data and thus put people and items on the same ruler or scale. When a person and item have the same logit value, it indicates that the probability of a correct answer is 50/50.

Thus the construct validation model presented in Figure 1 hypothesizes a construct that progresses in a hierarchical model of difficulty from which items can be written. Those items are more likely to successfully be performed by people who are located at the same level or higher than by people who are located lower on the axis. To determine if the items progress in the hypothesized hierarchy, the items need to be administered to a group of people who represent the full range of ability. From the responses of the test-takers, item scores can be calculated.

Item scores must be converted to measures because raw scores often are group dependent and do not have the characteristics of interval data. A further discussion of
this issue as applied to language assessment can be found in Clifford and Cox (2013). Rasch measurement, on the other hand, is mathematically modeled to be group independent; thus, the item parameter measures function independently of the examinees. Furthermore, as noted earlier, logits are interval data. Those measures can therefore inform whether the items have functioned as intended, i.e., as defined in the construct independent of the test-takers. If they do not, then the construct may need revision. This property makes Rasch measurement ideal for criterion-referenced testing, which is based on external criteria rather than on how test-takers compare to each other. If the items function as predicted by the construct, then there is evidence that the construct’s hierarchical structure has merit.

When considering how language learners progress in their ability, it may be helpful to visualize clouds and ladders to discuss growth. Often language is represented in comic books as clouds above a character’s head, and the size of the cloud is proportional to the amount of language the character needs. As learners develop in their language progression, the height, width, breadth, and depth of what is required to accomplish the language functions increase in multiple directions—that is, volume is added to the cloud. While unidimensionality is one of the assumptions needed to perform IRT, Clifford and Cox (2013) argued that true unidimensionality rarely exists in the social sciences. With language in particular, unidimensionality can only be achieved through careful a priori definitions of what is being measured. These definitions create specific combinations of elements, and each combination may be equated to a separate step in the hypothesized hierarchy. The next challenge is to operationalize and score the construct. Many test developers have simply constructed a test with an array of item difficulties, such as different heights or rungs on a ladder, and then have attempted to set a cut score that indicates in which proficiency cloud the language learner can function. Cizek and Bunch (2007) defined the setting of such conversion formulas as the “concrete activity of deriving cut points along a score scale” (p. 14).

As shown in Figure 3, a traditional test design is like a long ladder. Because this design produces a single score for each test-taker, various standard-setting procedures are used to determine how to convert those scores into proficiency levels. This process is not without its challenges. Because there is only one score for each test-taker, and that score is compensatory, e.g., a person with well-established Intermediate skills and some emerging Advanced and Superior skills could have the same total score as a test-taker with solidly established Intermediate and solid Advanced skills but very limited Superior skills. With both test-takers having the same score, a conversion table would either give both individuals an Advanced rating or give neither of them that rating. A further complication, which can reduce the accuracy of conversion tables, is that a test-taker’s correct guesses on items from all of the test’s proficiency levels are included in his or her total score. Furthermore, this method provides no direct information about the construct being tested.

By contrast, the series of shorter ladders in Figure 4 depicts the approach used when criterion-referenced testing principles are applied. Although the process can be complex, the essential elements of criterion-referenced testing can be summarized with just three requirements:

- Each level is treated as a separate construct or criterion.
- Each level has separate test specifications.
- Each level-specific test is scored separately, with either classical test theory or invariant Rasch measurement.

By applying these three criterion-referenced principles, many of the rating deficiencies associated with traditional test designs can be avoided. For instance, it is no longer necessary to estimate how guessing at levels beyond the test-takers’ actual ability level may have influenced their total scores, nor is it necessary to convert a range
of multilevel compensatory scores into categorical proficiency ratings. As Luecht (2003) pointed out, a complete alignment of the construct, test specifications, and scoring procedures is crucial when measuring a set of hierarchical, multidimensional standards like those in the ILR/ACTFL proficiency guidelines. How the items on each
level’s ladder function provides evidence of whether the construct on which the items were based is valid.

**Listening Proficiency Guidelines**

As with the speaking guidelines, the ACTFL listening proficiency guidelines grew out of the ILR guidelines that were established in the 1950s to create a “system that was objective, applicable to all languages and all Civil Service positions, and unrelated to any particular language curriculum” (Herzog, n.d., para. 3). While the ILR scale originally interpreted language proficiency as a unitary construct, it was later divided into the four skill areas of speaking, writing, listening, and reading. The intent was to provide a scale that stakeholders in various branches of government, with little or no language training, could use in making personnel assignments. In addition to serving as the foundation for the ACTFL scales, the ILR scale was also the basis for the NATO Standardization Agreement (STANAG) 6001 scales (2010). While the major levels of the three scales have remained constant, the sublevels are different. STANAG and ILR may apply a “plus” sublevel rating, whereas ACTFL divides major levels into three parts: Low, Mid, and High. The ACTFL guidelines were recently updated and sublevels were also defined for the Advanced Level. The introduction to the new guidelines states:

The ACTFL Proficiency Guidelines 2012—Listening describe five major levels of proficiency: Distinguished, Superior, Advanced, Intermediate, and Novice. The description of each major level is representative of a specific range of abilities. Together these levels form a hierarchy in which each level subsumes all lower levels. The major levels Advanced, Intermediate, and Novice are divided into High, Mid, and Low sublevels. The subdivision of the Advanced Level into High, Mid, and Low is new. This makes the Listening descriptions parallel to the other skill-level descriptions.

Listening is an interpretive skill. Listening comprehension is largely based on the amount of information listeners can retrieve from what they hear and the inferences and connections that they can make. By describing the tasks that listeners can perform with different types of oral texts and under different types of circumstances, the Listening Proficiency Guidelines describe how listeners understand oral discourse. The Guidelines do not describe how listening skills develop, how one learns to listen, nor the actual cognitive processes involved in the activity. Rather, they are intended to describe what listeners understand from what they hear.

These Guidelines apply to listening that is either Interpretive (non-participative or overheard) or Interpersonal (participative). (ACTFL, 2012, p. 15)

While all types of listening involve similar audiological processes, the ability to comprehend a speaker’s intent differs from the ability to negotiate meaning and ask for clarification. Thus, interpersonal communication acknowledges the participative, give-and-take nature of listening when both parties are present. Often, though, listening for meaning occurs without the possibility to ask for clarification. For instance, media consumption, lectures in large classes, and announcements over public address systems are some of the many ways in which a language learner must attend to speech and interpret its content. As participatory listening is indirectly assessed in Oral Proficiency Interviews, this study focuses on nonparticipatory listening.

**Proficient Listening**

Proficient listening is defined as the active, automatic, far-transfer process of using one’s internalized language and culture expectancy system to efficiently comprehend an authentic spoken passage for the purpose for which it was produced (see Figure 5). This
Definition of proficient listening can be applied to each ACTFL proficiency level, which allows the definition of multistage manifestations of the unidimensional trait that conform to the hierarchical theory upon which the guidelines are based.

Following Luecht’s (2003) recommendation that one should define major stages or levels of progress along the continuum of language ability, each stage or step can be defined as a separate performance standard with its own combination of aligned communication expectations. These specifications are summarized in Figure 6, where the speaker’s purpose, based on Child’s (1987) text modes, is aligned with the types of texts that are typically used to accomplish those tasks and where the listener’s task is aligned with the speaker’s purpose in communicating.

Note that the Novice level is not included in Figure 6 as it is characterized by the inability to sustain the Intermediate level. On the other hand, the Distinguished level is included to provide a ceiling for the Superior level, although there have been, to date, no requirements to test at that level.

It is also important to point out that as with the other skill modalities described in the ACTFL Proficiency Guidelines, the listening proficiency ratings are noncompensatory; that is, to qualify for a given rating, the individual must consistently meet all of the construct criteria for that level (Swender & Vicars, 2012, p. 5). In a listening proficiency assessment, this requirement means that the listener must be able to consistently comprehend speech of the specified type for the purpose for which it was created. Stated another way, listeners must successfully accomplish those comprehension tasks that are aligned with the speaker’s purpose. Blended or nonaligned combinations of listener and speaker purpose are possible, and when they occur, they may be useful in assigning sublevel ratings. However, nonaligned combinations are not useful when assigning major-level proficiency ratings because consistent performance across the aligned factors for each level is both expected and required. Thus, testing whether the listener can perform lower-level listening tasks on higher-level text passages provides insufficient information to assign the higher proficiency rating: For example, understanding the main idea (an Intermediate-level task) of a Superior-level speech would not qualify the listener for a Superior or even an Advanced proficiency rating. Again, the expectation of the listening guidelines is that the speaker’s purpose and the listeners’ task are congruent. If the speaker’s purpose is to narrate a story, then the listeners’ purpose is to comprehend the details of that narration. It should be noted that this restriction of the tested tasks to those that align with the speaker’s purpose distinguishes assessment practices from teaching practices, because proficiency testing places the focus on whether listeners can comprehend spoken passages rather than on the instructional focus of how listeners comprehend different text types and speaker purposes at each of the tested levels.

**Background Summary**

The listening proficiency scales have been in use for a number of years as a practical way to describe overall language ability levels, yet there is little research to validate their use. Some research with the other receptive modality—reading—has seemed to contradict the practical findings of those who use the proficiency guidelines as the basis for their tests. The failure of these reading studies to have their items align hierarchically could come from a number of factors, including a misconception of the scale, an inability to define the construct unidimensionally, applying an inappropriate
measurement theory, and not having a clear construct definition that spans all levels of the scale. By contrast, other research has validated the use of such scales by carefully defining the construct (Clifford & Cox, 2013). Thus, by extension, applying the same alignment and measurement principles from the Clifford and Cox (2013) reading validation study to this listening validation study, similar results should be found. In order to determine the extent to which test items for which both the passage and the question are based on the ACTFL listening proficiency guidelines, ascending in a hierarchy of difficulty levels, the factors of task, condition, and accuracy would need to be controlled.
Method
To investigate the validity of the listening scales and to determine the extent to which test items carefully based on the ACTFL listening proficiency guidelines ascend in a hierarchy of difficulty levels, listening test items targeting the major levels of Intermediate and higher were created in a number of different languages. No test items targeted the Novice level, as the failure to correctly answer Intermediate items is indicative of Novice proficiency. The test development process was conducted first in English for listening tests that assessed through the Superior level for use by NATO. Subsequently, other grants and projects necessitated the development of additional proficiency-based listening tests through the Superior level, including, in order of development, Arabic and Russian, as well as listening tests through the Advanced level, in order of development, in Turkish, Chinese, Spanish, and French. This article reports data for test items in these seven languages by the number of levels tested and the chronology of their development. A description of each instrument, the number of questions targeting each major proficiency level, and the subjects who took the tests are presented in the Results section. Data analysis procedures were chosen based on the type of data obtained from the administrations of each test.

Training the Item Writers
To see if test items that aligned with the ACTFL/ILR/NATO listening proficiency guidelines could be developed, item writers were first taught about the construct of listening proficiency that would be applied. Second, they were presented with the proficiency scales and shown how the task, conditions, and accuracy criteria were manifest at each level. Third, they were taught to select appropriate listening passages for each level, to write questions that aligned the listener’s task with the interlocutor’s purpose, and to create answers and plausible response distractors that aligned with each level’s expectations.

Internal Validation
Prior to administering the test items to examinees, the items went through another review process where teams of three to five item writers were asked to ensure that the questions being asked were aligned with the interlocutor’s communicative purpose, that the listening passage was aligned with the typical characteristics of spoken passages produced for that purpose, and that the intended difficulties of that combination corresponded to a level in the proficiency scale. As a vehicle to structure that review, the item writers were asked to look at each item and to estimate each item’s:

- difficulty for learners at that targeted level.
- difficulty for those with proficiency at one level lower than the item.
- difficulty for learners at one level higher than the item.
- discrimination index across levels.

Any item that was judged to lack at-level alignment was revised. The items were then added to the pool of items to be included in the study and were subsequently administered to learners of varying ability levels.

Data Analysis Procedures
The items were placed into multiple test forms of which a minimum of 20% of the items acted as anchors or overlap items and were administered to groups ranging from 93 (for the Turkish study) to 587 (for the Arabic study) participants depending on the language. The results were then analyzed using the software program Winsteps (Linacre, 2012), which is based on the Rasch measurement. As noted earlier, the data from Rasch measurements are interval level and can be analyzed using parametric statistics. Demonstrating that the proficiency levels in the guidelines represent a hierarchy of stages or steps of increasing
difficulty required that three conditions be met:

- The items designed to measure specific proficiency levels should cluster together at their intended difficulty levels.
- The mean difficulty values of each of those item clusters should be statistically different from the mean difficulty values of the other clusters of items.
- The mean values of those item clusters should be arrayed in a hierarchy of increasing difficulty.

Most test development processes involve piloting items with students and then eliminating poorly functioning items. Because the item difficulties of the empirically tested items were the dependent variable, indicating how the construct definition was functioning, all items needed to be included in the analysis. If this were a test development project, those items would be eliminated prior to the final version of the test. To validate the scale, however, it was important to see if item writers trained on the scale with the aligned construct definition could create items that ascended hierarchically. Thus, all the items that were tested empirically were analyzed together.

Each test was analyzed for reliability through person and item separation statistics obtained with Rasch measures. As with most statistical procedures, the greater the sample size, the greater the confidence that the results are not biased by aberrant behavior. While exploratory analyses can be conducted with as few as 30 people, researchers typically aim for a minimum of 100 to achieve stable parameter estimates (Linacre & Wright, 2009). While Rasch reliability uses the same ratio (0 to 1), it is reported somewhat differently than classical test theory in that it signifies the reproducibility of relative measure location (Linacre & Wright, 2009). Therefore, the closer the person separation reliability is to 1, the more likely it is that persons with high ability actually do have higher measures than people with lower ability. The same is true with items. The closer the item separation reliability is to 1, the more likely it is that items with high difficulty values actually do have higher measures than items with lower difficulty values. To interpret the reliability, then, one must determine if the test can divide the sample into enough distinct values for the needed purpose. The tentative guideline for this division is the following: 0.9 = three or four levels, 0.8 = two or three levels, and 0.5 = one or two levels.

After reporting the reliability, the item difficulty values based on the logits were used as the dependent variable. For data from the Turkish, Chinese, Spanish, and French tests that only went to the Advanced level, an independent-samples t test that evaluated the 95% confidence interval (CI) of the mean difference and its effect size was used. For English, Arabic, and Russian tests that went through the Superior level, the statistical test used to determine whether the intended item difficulty, as measured in logits, differed by intended proficiency levels was a one-way ANOVA. The comparisons between levels were reported using a Bonferroni posthoc test (Larson-Hall, 2010) that defined the mean difference, its 95% CI, and its effect size. To see how important the intended level of item difficulty was on the differences between the mean difficulty of the item clusters, the effect size was calculated and reported with Cohen's $d$, which indicates the number of standard deviations that separate the means being compared.

**Results**

Results are presented here for tests in seven languages: Arabic, Chinese, English, French, Russian, Spanish, and Turkish. Because the tests were created for different projects with different requirements, the structure of each of the tests was slightly different. However, the item creation process was consistent across all of the languages. These tests were then administered to different groups of examinees, and the Rasch measurement statistics were calculated for each language test. The results are presented by the highest
proficiency level at which examinees can be assessed (Advanced or Superior) using each test and by the chronological order of item development.

Languages Tested Through Advanced Turkish Study
The Turkish test was originally created as part of a project sponsored by the National Middle East Language Resource Center (NMELRC) at Brigham Young University (BYU) to develop assessments in some of the less commonly taught languages. The test consisted of two ACTFL proficiency levels: Intermediate, with 52 items, and Advanced, with 46 items. Items were drawn at random, and each examinee took 20 of the 52 items at the Intermediate level and 20 of 46 items at the Advanced level. The test was administered to a total of 93 university undergraduate students from various institutions who were studying Turkish as a foreign language.

The level of internal consistency of the Turkish test was 0.77, as indicated by the Rasch IRT person separation reliability. This level of reliability confirmed that examinees could be reliably divided into two statistically distinct groups (Linacre & Wright, 2009). The item reliability of 0.84 was moderately high and indicated that the items functioned distinctly for at least two levels of difficulty (Linacre & Wright, 2009). The item logit means at each of the intended difficulty levels progressed monotonically (see Table 1). These results were quite promising because they were attained despite both the small number of subjects and the large number of items in the test bank, which resulted in an average item exposure of just 33 times per item. Because Rasch analyses typically need more examinees (e.g., >100) while the reliability statistics were lower than hoped for, the trend was similar to that found for the other languages and led to the belief that, with a greater number of subjects and thus greater exposure of each test item, the reliability estimate would increase.

Levene’s test for homogeneity of variances on the Turkish data (F = 2.31, p = 0.13) indicated that equal variances could be assumed and an independent-samples t test was appropriate. The t test between the Intermediate- and Advanced-level items was conducted to determine if the two groups of items differed in item difficulty. The 95% CI (see Figure 7) for the difference in the means was between −3.12 and −2.07 (t = −5.02, p < 0.001, df = 96). Thus, the items’ intended proficiency level had a strong effect (d = 1.02) on the empirical item difficulty.

Chinese Study
This test was initially designed for university undergraduate students studying Chinese as a foreign language. The test had 82 items that were divided into the two proficiency levels: Intermediate (n = 35) and Advanced (n = 47). Two separate test forms were created with shared anchor items across the forms. The tests were subsequently administered to 143 students studying Chinese at four different institutions.

The test had a high internal consistency with a Rasch person reliability of 0.91,
indicating that the examinees could be reliably divided into three or four statistically distinct groups (Linacre & Wright, 2009). The item reliability of 0.94 was very high and indicated that the items functioned at distinctly separate levels of difficulty. The means of each of the intended difficulty levels progressed monotonically (see Table 2).

Levene’s test for homogeneity of variances on the Chinese data (F = 2.63, p = 0.11) indicated that equal variances could be assumed and an independent-samples t test was appropriate. The t test between the Intermediate- and Advanced-level items was conducted to determine if the two groups of items differed in item difficulty. The 95% CI (see Figure 8) for the difference in the means was between −2.09 and −0.28 (t = −2.61, p < 0.05, df = 80), and it can be seen that the items’ intended proficiency level had a strong effect (d = 3.74) on the empirical item difficulty.

**Spanish Study**
This test was initially designed for students who were studying Spanish as a foreign language at the postsecondary level and was designed to measure only through the Advanced level. The subjects were 197 students from four different universities. The test had 74 items that were divided into the two proficiency levels: Intermediate and Advanced. The test was administered in two forms (A and B) that each had 34 items, including 14 Intermediate-level items and 20 Advanced-level items with each form sharing five Intermediate items and five Advanced items for a total of 10 anchor items on both forms.

The test had a Rasch person reliability of 0.85, indicating that there was a relatively high level of internal consistency and that the examinees could be reliably divided into two to three statistically distinct groups. The item reliability of 0.97 was very high and indicated that the items functioned at distinctly separate levels of difficulty. The mean of the Advanced items was much higher than the mean of the Intermediate items (see Table 3).

Levene’s test for homogeneity of variances on the Spanish data (F = 2.63, p = 0.11) indicated that equal variances could be assumed and an independent samples t test was appropriate. The t test between the
Intermediate- and Advanced-level items was conducted to determine if the two groups of items differed in item difficulty. The 95% CI (see Figure 9) for the difference in the means was between $-2.55$ and $-1.36$ ($t = -6.60$, $p < 0.001$, $df = 72$), and it can be seen that the items' intended proficiency level had a strong effect ($d = 1.54$) on the empirical item difficulty.

**French Study**

This test was initially designed for students who were studying French as a foreign language at the postsecondary level and was designed to measure only through the Advanced level. The subjects were 214 students from one university. The test had 63 items that were divided into two proficiency levels: Intermediate and Advanced. The test was administered in four forms that covered two different tiers. The lower tier was for the first- and second-year students (Forms A and B). These test forms had 34 items total and consisted of 26 Intermediate items and eight Advanced items. The higher tier was for the students in the third year or higher (Forms C and D). These test forms had 37 items in total and comprised eight Interme-

<table>
<thead>
<tr>
<th># of items</th>
<th>Mean logit value</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intermediate</td>
<td>35</td>
<td>$-1.12$</td>
</tr>
<tr>
<td>Advanced</td>
<td>47</td>
<td>0.07</td>
</tr>
<tr>
<td>Total</td>
<td>82</td>
<td></td>
</tr>
</tbody>
</table>

**TABLE 2**

**Descriptive Statistics of Chinese Item Difficulty by Intended Proficiency Level**

**FIGURE 8**

95% CI of Chinese Item Difficulty by Intended Proficiency Level

Language: Chinese

Intermediate

Advanced

Level
diate items and 29 Advanced items, with each form sharing anchor items.

The test had a Rasch person reliability of 0.80, indicating that there was a relatively high level of internal consistency and that the examinees could be reliably divided into two to three statistically distinct groups. The item reliability of 0.96 was very high and indicated that the items functioned at distinctly separate levels of difficulty. The mean of the Advanced items was much higher than the mean of the Intermediate items (see Table 4).

Levene’s test for homogeneity of variances for the French data ($F = 0.31, p = 0.58$) indicated that equal variances could be assumed and an independent samples $t$ test was appropriate. The $t$ test between the Intermediate- and Advanced-level items was conducted to determine if the two groups of items differed in item difficulty. The 95% CI (see Figure 10) for the difference in the means was between $-2.25$ and $-1.19$ ($t = -6.56, p < 0.001, df = 61$). The null hypothesis that the difference in the means was zero could be rejected, and it can be seen that the items’ intended proficiency level had a strong effect ($d = 1.86$) on the empirical item difficulty.

<table>
<thead>
<tr>
<th>TABLE 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Descriptive Statistics of Spanish Item Difficulty by Intended Proficiency Level</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Intermediate</td>
</tr>
<tr>
<td>Advanced</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

FIGURE 9

95% CI of Spanish Item Difficulty by Intended Proficiency Level

Language: Spanish

![Graph showing 95% CI of Spanish Item Difficulty by Intended Proficiency Level](image-url)
Languages Tested Through Superior English Study

The English test was originally created for NATO to assess whether personnel had sufficient English skills to succeed in a multinational work environment where English was the common language. The test consisted of three ACTFL proficiency levels: Intermediate with 20 items, Advanced with 38 items, and Superior with 24 items. The items were drawn at random, and each examinee took all 20 of the items at the Intermediate level, 20 of the 38 items at the Advanced level, and 20 of the 24 items at the Superior level. The test was administered to a total of 188 examinees who were personnel associated with NATO.

The level of internal consistency of the English test was 0.87, as indicated by the Rasch IRT person separation reliability. This level of reliability confirmed that the examinees could be reliably divided into three or four statistically distinct groups (Linacre & Wright, 2009). The item reliability of 0.99 was very high and indicated that the items functioned at distinctly separate levels of difficulty (Linacre & Wright, 2009). The means of each of the intended difficulty

### TABLE 4

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intermediate</td>
<td>31</td>
<td>-0.88</td>
<td>1.05</td>
</tr>
<tr>
<td>Advanced</td>
<td>32</td>
<td>0.85</td>
<td>1.00</td>
</tr>
<tr>
<td>Total</td>
<td>63</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The table above lists the descriptive statistics for French item difficulty by intended proficiency level. The mean difficulty for the Intermediate level is -0.88, for the Advanced level is 0.85, and for the total is not specified. The standard deviation for each level is also provided.
levels progressed monotonically (see Table 5).

Levene’s test for homogeneity of variances of the English data ($F = 1.44, p = 0.58$, $df = 2, 79$) indicated that equal variances could be assumed, so an ANOVA with its subsequent posthoc Bonferroni contrast was an appropriate test. Comparisons using the Bonferroni contrasts found statistical differences between the Intermediate and Advanced items (mean difference $= -1.47$ logits, a 95% CI between $-1.83$ and $-1.10$, and $p < 0.001$) and between the Advanced and Superior items (mean difference $= -1.23$ logits, a 95% CI between $-1.61$ and $-0.84$, and $p < 0.001$). Figure 11 shows the differences between the levels. Although some of the items did not cluster as expected, as a whole the items by level had different difficulty levels. The effect size among the three levels was very strong. For Intermediate items vs. Advanced items, Cohen’s $d = 2.26$, and for the Advanced items vs. Superior items, $d = 2.09$. Thus, the intended proficiency level of the items had a strong effect on the empirical item difficulty level.

Arabic Study

The Arabic test was originally created as part of a project to develop assessments in some

<table>
<thead>
<tr>
<th># of items</th>
<th>Mean logit value</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intermediate</td>
<td>20</td>
<td>$-1.13$</td>
</tr>
<tr>
<td>Advanced</td>
<td>38</td>
<td>0.02</td>
</tr>
<tr>
<td>Superior</td>
<td>24</td>
<td>.90</td>
</tr>
<tr>
<td>Total items</td>
<td>82</td>
<td></td>
</tr>
</tbody>
</table>

**TABLE 5**

**Descriptive Statistics of English Item Difficulty by Intended Proficiency Level**

**FIGURE 11**

95% CI of English Item Difficulty by Intended Proficiency Level
of the less commonly taught languages sponsored by BYU’s NMELRC. The test consisted of three ACTFL proficiency levels: Intermediate with 36 items, Advanced with 36 items, and Superior with 28 items. Items were drawn at random, and each examinee took 20 of the 36 items at the Intermediate level, 20 of the 36 items at the Advanced level, and 20 of the 28 items at the Superior level. The test was administered to a total of 587 university undergraduate students who were studying Arabic as a foreign language from various institutions.

The Arabic test had a Rasch IRT person reliability of 0.80, indicating a moderately high level of internal consistency. This level of reliability confirmed that the examinees could be reliably divided into two or three statistically distinct groups (Linacre & Wright, 2009). The item reliability of 0.98 was very high and indicated that the items functioned at distinctly different levels of difficulty (Linacre & Wright, 2009). The means of each of the intended difficulty levels progressed monotonically (see Table 6).

Levene’s test for homogeneity of variances of the Arabic data (F = 0.31, p = 0.73, df = 2,97) indicated that equal variances could be assumed, thus an ANOVA with a subsequent posthoc Bonferroni contrast was selected as an appropriate test. Comparisons using the Bonferroni contrasts found statistical differences between the Intermediate and Advanced items (mean difference = −1.08 logits, a 95% CI between −1.63 and −0.51, and p < 0.001) and between the Advanced and Superior items (mean difference = −0.83 logits, a 95% CI between −1.44 and −0.21, and p < 0.01). Figure 12 shows the differences between the levels. Although some of the items fell between the major levels in difficulty logits, as a whole the items had different difficulty levels for each targeted level. The effect size among the three levels was very strong between the levels as they progressed: for Intermediate items vs. Advanced items, Cohen’s d = 1.14, and for the Advanced items vs. Superior items, d = .92. Thus the intended proficiency level of the items had a strong effect on the empirical item difficulty level.

**Russian Study**

The Russian study was designed for students of Russian as a foreign language. The subjects were 120 undergraduate students from Brigham Young University. The test had 124 items that were divided into the three proficiency levels: Intermediate, Advanced, and Superior. The test was administered in eight forms of 60 items each with 18 shared anchor items across the forms.

The Russian test had a Rasch IRT person reliability of 0.93, indicating a high level of internal consistency. This level of reliability confirmed that the examinees could be reliably divided into three or four statistically distinct groups (Linacre & Wright, 2009). The item reliability of 0.94 was very high and indicated that the items functioned distinctly at more than four levels of difficulty (Linacre & Wright, 2009). The means of each of the intended difficulty levels progressed monotonically (see Table 7).

| TABLE 6

| Descriptive Statistics of Arabic Item Difficulty by Intended Proficiency Level |
|-------------------------------|----------------|----------------|
| # of items | Mean logit value | SD  |
| Intermediate | 36  | −0.92 | 0.94 |
| Advanced  | 36   | 0.16  | 0.95 |
| Superior   | 28   | 1.12  | 1.12 |
| Total items | 100 |     |    |
Levene’s test for homogeneity of variances (\(F = 0.10, p = 0.91, df = 2,121\)) indicated that equal variances could be assumed, so an ANOVA with its subsequent posthoc Bonferroni contrast was an appropriate test. Comparisons using the Bonferroni contrasts found statistical differences between the Intermediate and Advanced items (mean difference = –2.05 logits, a 95% CI between –2.63 and –1.47, and \(p < 0.001\)). However, the difference between the Advanced and Superior items (mean difference = –0.25 logits, a 95% CI between –0.98 and –0.46) was much smaller than that in any of the previous studies. Figure 13 shows the differences between the levels. The effect size for Intermediate items vs. Advanced items had Cohen’s \(d = 1.14\), indicating a strong effect of intended level on mean logit difficulty. The effect size for the Advanced items vs. Superior items, \(d = 0.22\), was quite small, indicating a small to medium effect on item difficulty.

This finding of a relatively small effect size between Superior and Advanced items appeared anomalous and was initially puzzling until further investigation. Stable IRT parameter estimates within \(\pm 1\) logit require

<p>| Table 7 |
|-----------------|-----------------|-----------|
| <strong>Descriptive Statistics of Russian Item Difficulty by Intended Proficiency Level</strong> |</p>
<table>
<thead>
<tr>
<th># of items</th>
<th>Mean logit value</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intermediate</td>
<td>51</td>
<td>–1.26</td>
</tr>
<tr>
<td>Advanced</td>
<td>49</td>
<td>0.79</td>
</tr>
<tr>
<td>Superior</td>
<td>24</td>
<td>1.05</td>
</tr>
<tr>
<td>Total items</td>
<td>124</td>
<td></td>
</tr>
</tbody>
</table>
a minimum of 30 examinees, although, as with nearly all statistical procedures, more subjects will result in greater certainty. Dividing the items among eight test forms and having only 120 examinees, of which only a small subset were advanced language learners, resulted in 30 of the 72 (e.g., 41%) Advanced and Superior items having 29 or fewer examinee responses. The paucity of data for these items compared to the items in the other languages could explain the divergence from the pattern that was emerging.

Summary of Results
This research explored the extent to which test items ascend in a hierarchy of difficulty levels when both the passage and the question are based on the ACTFL listening proficiency guidelines. In all but one study (Russian), the intended proficiency levels of the items had a strong effect size on the empirical item difficulties.

It is noteworthy that no items were excluded from the analysis. In every test development process, items that do not function due to poor discrimination or malfunctioning distractors are excluded from the final tests. However, because this study was focused on the items and their intended alignment, poorly functioning items were not excluded. In Figure 14, it is evident from the boxplot distributions that some Intermediate items were more difficult than intended, while some Superior items were easier.

The use of multiple-choice items occasionally produced spurious results. A review of those items indicated that in some instances, distractors were too attractive for the proficiency level for which the question was intended, which led to the item being more difficult than the proficiency scale warranted. In other instances, distractors were so unrealistic that the examinees were drawn to the correct response even though the question and passage were beyond the students’ ability. Finally, some item types were less suited for upper-level proficiency testing. Rupp, Ferne, and Choi (2006) found that when language input is more difficult than the examinee’s level of reading comprehension, multiple-choice questions no longer measure comprehension and measure
problem-solving skills instead. While any single item type does not automatically prevent it from testing at a specific proficiency level, the rigor needed to create items in which the difficulty levels align with proficiency level is increased with multiple-choice questions. In the instances of these language tests, had the malfunctioning items been removed from the analyses, as would have routinely been done in typical test

Note: Because the different languages share no common items or examinees, the IRT calibrations make it inappropriate to compare languages to each other. Each language went through its own calibration process to calculate its logit scale. So while each language logit scale has the characteristics of interval data and maintains its probabilistic meaning, comparing scales may lead to erroneous interpretations. It would be physically analogous to measuring temperature in physics in which Fahrenheit, Celsius, and Kelvin are all used.
development projects, the differences between the empirical difficulty levels would have been even greater. For instance, if one were to eliminate 25 percent of the misaligned items, the remaining data would have even greater separation between the levels with very little possibility of the items from the intended levels overlapping with each other (see Figure 15). A related observation is that given sufficient training and time, item writers can indeed write level-specific items that match their targeted proficiency levels and empirically align with the hierarchical progression of proficiency scales, but training is required.

Discussion
As presented initially, one way to validate a construct is through the process of defining the construct, writing items based on the construct, scoring those items, and then converting those scores to measures. In the results, it is evident that there was a strong effect between the empirical item difficulty measures and the intended difficulties. Through carefully defining the construct of listening and ensuring that all the aspects co-occurred in item creation, the items posited to represent the construct did indeed align with levels of the proficiency scale. The results of this study support the difficulty hierarchy posited by the ACTFL, ILR, and related listening proficiency guidelines.

Implications for Testers and Instructors
Testers wishing to assess curriculum-independent, real-world listening proficiency according to the ACTFL Proficiency Guidelines (2012) should recognize that blended task and text combinations drawn from different proficiency levels provide insufficient information to assign criterion-referenced proficiency ratings. Therefore, they should treat each major level as a separate set of assessment criteria and ensure that those criteria are aligned with the speaker purpose, text characteristics, and listener tasks described for that level. They should also use noncompensatory, criterion-referenced scoring criteria when assigning ratings. Perhaps the most difficult challenge is structuring multiple-choice questions so that their difficulty level is aligned with the difficulty level of the associated listener task.

The implications for instructors are different than those for testers. Whereas blended task and text combinations drawn from different proficiency levels are inappropriate for the assigning of criterion-referenced proficiency ratings, those same blended combinations may provide useful ramps and scaffolding that can help students make progress toward higher proficiency levels. Therefore, instructors desiring to raise the proficiency level of their students should begin by establishing the students’ base level of sustained ability and then use scaffolding techniques to incrementally introduce features from the next higher major proficiency level.

“Scaffolding” is the label that is often applied to the commonsense teaching practice of helping learners to understand increasingly complex concepts and master more difficult skills by selectively teaching the components of the targeted new ability one piece at a time. Then, after the supporting pieces are in place, learners are asked to integrate those skills into a cohesive ability. Application of such a step-by-step process makes the learning tasks more manageable and is often the most effective way to move learners from what they currently know or are able to do toward tasks for which they do not have complete knowledge and skills—that is, toward what they yet need to know or do.

Perhaps a real-world example can illustrate. Shortly after Nelson Mandela’s funeral, while the researchers were preparing the data in this article for a research presentation, there was a news segment featuring a panel discussing the controversy of the ill-prepared sign language interpreter at the memorial service. At first, it appeared that...
the story was simply a factual report of the events at the funeral; however, a more subtle subtext appeared in which the real point of the discussion was presidential security. If an unknown interpreter could have close access to the U.S. president and other world leaders, what would stop terrorist organizations from planting an assassin in such a position? An Intermediate listener could probably identify some key words, phrases
and ideas—Mandela, president, funeral—but would be unable to provide any details. An Advanced listener, on the other hand, would likely be able to provide the details of the story yet fail to make the connection that the narrative was in fact used to support a structured argument about the need for increased vigilance in security protocols. A Superior listener would have access to all of these levels of meaning.

Drawing on this example, an instructor could bring in this type of authentic passage and scaffold it for the class. While it would be inappropriate to use such a passage with Novice listeners, Intermediate listeners could be prepared so as to be successful in comprehending an excerpt from the passage. For example, the portion of the passage that related the facts about what happened at the funeral could be selected and study material prepared, including a lexicon about the deaf community, interpreting, funerals, etc., as well as grammatical features of past narration. Students could then listen, and relisten, to the passage in order to gain comprehension of this paragraph-length narrative as well as overall competence in dealing with this type of text. For a group of Advanced listeners, on the other hand, the whole passage could be used, but the instructor could prepare genre-specific schema to help the students know what to listen for. Because the crossfire format of news shows often have stock characters (the liberal, the neoconservative, the moderator) who use a variety of tactics to argue with each other, a review of character roles, logical fallacies frequently used, and types of arguments could be taught prior to listening. The student could then be allowed to repeatedly listen to the passage in order to make this full, Superior-level text accessible to the Advanced learner. Thus, as instructors manipulate and assist students in reaching beyond their current level, they provide students with the necessary opportunities to practice and develop the skills that are required to reach that next level of proficiency.

As long as instructors are cognizant of the difference between performance and proficiency, these types of activities are very valuable in assisting students to move along the proficiency continuum. The only risk is that instructors with a cursory understanding of the scale and the relationship between performance and proficiency may overinterpret successful performance on carefully scaffolded tasks as evidence of a particular proficiency level. In order to assist language educators in their efforts to provide students with realistic feedback, the ACTFL developed the ACTFL 2012 Performance Descriptors for Language Learners, which explicitly differentiate between proficiency and performance. As learners establish islands of performance mastery in different content areas, those islands of performance can subsequently merge to form “a continent of proficiency” that indicates sustained ability to function at a particular level (Shekhtman, 2003). In summary, then, when the difficulty level of a text is beyond the level of the students, there are only three options:

1. Allow the student to fail.
2. Bring the level of the text down to the level of the student by revising it.
3. Bring the level of the student up to the level of the text by providing direct instruction not only in lexical and syntactical domains but also in helping the learner to explore social and cultural norms and expectations that are salient to understanding the text’s deeper meaning.

Only the last option is likely to improve student abilities.

**Future Research**

The authors are proceeding to apply the lessons learned while validating the ACTFL listening proficiency guidelines to multi-stage computer-adaptive tests with floor and ceiling ratings. Inherent in that test development process is the need to blend
theoretical research related to the assigning of sublevel proficiency scores with practical research into the optimal design of computer-adaptive tests of listening proficiency. One component of this research will be explorations into characteristics of item types that best align with the functions that are required of listeners at each level of proficiency. For example, it will be important to discern to what extent the language in which the question is posed (native or target) has an impact on listeners’ comprehension at the different levels and how item formats other than multiple choice affect the operationalization of the construct. The practical applications of using scales in assessments can help ensure the valid operationalization of the level descriptions and the interpretation of findings by others using the framework.

Conclusion
One classic scene from children’s literature is the conversation at a crossroads between Alice and the Cheshire Cat in Lewis Carroll’s tale Alice in Wonderland. When Alice asks which way she should go, the cat replies, “That depends a good deal on where you want to get to.” It was noted earlier that the proficiency guidelines have served as a practical map of functional language ability, but they are not suitable for every situation. However, this study on the validation of the listening scale demonstrates that if the guidelines are used and operationalized through the conjoint alignment of all of the features of the different levels, the resulting map provides useful information to help administrators, instructors, students, and other stakeholders know what their level of listening proficiency is and for what purposes they can most successfully use such skills across a range of personal and professional tasks and contexts.

References


Submitted April 29, 2014

Accepted June 6, 2014