Foreign Language Annals
Published by the American Council on the Teaching of Foreign Languages

Empirical Validation of Reading Proficiency Guidelines

Ray Clifford
Brigham Young University

Troy L. Cox
Brigham Young University

Abstract: *The validation of ability scales describing multidimensional skills is always challenging, but not impossible. This study applies a multistage, criterion-referenced approach that uses a framework of aligned texts and reading tasks to explore the validity of the ACTFL and related reading proficiency guidelines. Rasch measurement and statistical analyses of data generated in three separate language studies confirm a significant difference in reading difficulty between the proficiency levels tested.*

Key words: *ACTFL Proficiency Guidelines, multistage assessment, proficient reading, scale validation, testing reading*

Introduction

Language proficiency scales grew out of the pragmatic need to functionally describe what language learners can do in a real-world context of language use. While many academics acknowledge that the productive skills such as speaking and writing can fall into a hierarchical order of difficulty, there has been less consensus with the receptive skills. Occasionally, real-life examples support the idea that various texts can be ordered by text complexity and difficulty. For instance, a junk mail postcard received by one of the authors provided a real-world example of texts that were written for different purposes and, commensurate with those purposes, used observably different writing styles. On the front of the card was printed in a large, bold font, the message: "No cost. No obligation. You have definitely won." On the other side of the card, printed in a miniscule font was the clarification, "Should a legal adjudication be made that valid consideration was transferred or expended by offeree pursuant to the specific terms of this written offer, then in such event restitutory indemnification shall be accomplished hereunder upon written request."

Ray Clifford (PhD, University of Minnesota) is Associate Dean, College of Humanities, and Director, Center for Language Studies, Brigham Young University, Provo, Utah.

Troy L. Cox (PhD, Brigham Young University) is Technology and Assessment Coordinator at the English Language Center, Brigham Young University, Provo, Utah.

Foreign Language Annals, Vol. 46, Iss. 1, pp. 45–61.
DOI: 10.1111/flan.12033

Although some researchers have asserted that there is no hierarchical order of difficulty for reading proficiency (Alderson, 2000; Brown & Hudson, 2002; Lee & Musumeci, 1988), scales describing levels of reading proficiency have functioned well in both government and academia. There are several possible explanations for the paucity of empirical evidence. First, the construct of reading proficiency may have been inadequately defined so that the data produced were not related to the level descriptions. Second, the item writers may not have been trained to select reading passages that align with the different levels of the scale and, equally important, did not write questions that aligned with the passages and the scale task descriptions. Finally, the underlying measurement theory used in assessing reading ability did not define a priori what unidimensional trait was being measured, and the research did not generate interval-level data that would be appropriate for use with parametric statistics.

To investigate the validity of the scales, we asked the following research question:

> To what extent do test items ascend in a hierarchy of difficulty levels when both the passage and question are based on the ACTFL Reading Proficiency Guidelines? The associated null hypothesis would be that there is no difference in mean difficulty between or among test items written to assess the Intermediate, Advanced, and Superior levels of reading proficiency.

Background

The following report on this research is divided into a number of steps. First, we review the reading proficiency guidelines and what they entail. Second, we review previous research. Third, we discuss the importance of choosing a measurement theory that aligns with the reading proficiency construct. Finally, we discuss how the guidelines can be operationalized for item writers through (1) carefully defining the construct of reading proficiency and (2) ensuring that for each item there is alignment between the targeted level, the selected passage, and reader task.

Reading Proficiency Guidelines

As with the speaking guidelines, the ACTFL Reading Proficiency Guidelines grew out of the Interagency Language Roundtable (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 scale originally encompassed language proficiency as a unitary construct, it was later divided into the four skill areas of reading, writing, listening, and speaking. 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 (STANAG 6001, 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 defined for the Advanced level. The introduction to the new guidelines states:

> The *ACTFL Proficiency Guidelines 2012—Reading* 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. … By describing the tasks that readers can perform with different types of texts and under

> different types of circumstances, the Reading Proficiency Guidelines describe how [well] readers understand written texts. (ACTFL, 2012, p. 20)

Previous Research on the Guidelines

Perhaps because the productive skills of speaking and writing have been more observable, less attention has been given to the testing of second language receptive skills than to the testing of second language speaking proficiency. While the ACTFL Speaking Guidelines have been operationally validated (Surface & Dierdorff, 2003), that is not yet the case with the ACTFL Reading Guidelines. In fact, previous studies and academic reviews have questioned the validity of both the ACTFL Reading Proficiency Guidelines and the ILR Language Proficiency Guidelines (2012) from which the ACTFL guidelines were extrapolated. Despite the success that federal agencies have had in testing reading proficiency using the ILR proficiency guidelines, researchers have not statistically validated the difficulty hierarchy posited by those related sets of proficiency guidelines.

Specifically, Alderson (2000) concluded that there is "no empirical evidence to validate the *a priori* definitions of levels" (p. 279; italics in original). Brown and Hudson (2002) stated that although the ACTFL guidelines have been frequently described as criterion-referenced, there has been circularity in the descriptions of the texts and the reader's ability. Bernhardt (2011) added the insight that the assessment of reading comprehension has been complicated by readers' use of compensatory processing strategies. In tests of the reading abilities of Italian learners, Lee and Musumeci (1988) concluded that the text types described by Child (1987) did not form a hierarchy of increasing difficulty and that their students' performance was not consistent with the hierarchy of reading skills described in the ACTFL proficiency guidelines. Despite these criticisms, the ILR language proficiency descriptions continue to be used in high-stakes testing within the federal government, as well as the related STANAG 6001 proficiency descriptions used by NATO.

In the attempt to reconcile the apparent contradiction between real-world practices and research results, three studies were particularly valuable. First, Lee (2001) pointed out that much of the debate about the feasibility of classifying texts into distinguishable categories may be attributed to the imprecision of the categorical definitions used. In a massive analysis of the types of language texts found in the British National Corpus, he found that professional linguists did not always agree on the definitions of text classification terms such as genre, text type, style, and register. He also noted that differences in interpretation were of little consequence, because when real-world texts were categorized, the various classification categories were interrelated and were naturally co-selected. Genres are often recognized by their typical text types, and texts selected based on text type are often predominantly from a particular genre.

Second, it was important to recognize that the challenge of dealing with multidimensional traits is not unique to language assessment. The field of cognitive science must also cope with the assessment of complex, multidimensional mental skills such as those encountered when attempting to accurately measure reading ability—an ability that appears to be composed of a constellation of skills that are employed differently by different readers as they read texts of various types, on various topics, for various purposes. Luecht (2003), a cognitive specialist, provided an example of how the assessment of language proficiency may be accomplished through the careful alignment of theoretical constructs, proficiency classifications, and assessment methods. His recommendation was to see if one can measure major steps in skill development instead of attempting to assess all the developmental profiles that one may encounter.

Third, test method has long been recognized as an important and often

outcome-determining variable that can confound research results (Clifford, 1981). In a more recent study, Rupp, Ferne, and Choi (2006) found that using multiple-choice questions to assess reading comprehension could alter the construct being tested. Especially when the difficulty of the passage to be read exceeded the readers' ability, the readers ceased reading and shifted to problem-solving tactics as they attempted to deduce the most likely answer from the options presented.

Clearly, multidimensionality of language proficiency, the compensatory nature of reading skills, and the complexity of testing procedures have made the assessment of reading proficiency a daunting challenge in both first and second language research. Smith (2007) noted in the foreword to the sixth edition of *Understanding Reading* that contradictory models and explanations continue to flourish in the area of reading research.

Choice of Measurement Theory

The psychologist Kurt Lewin stated, "There is nothing more practical than a good theory" (1952, p. 169), yet many social scientists have not considered the theoretical basis of what they have measured and how they have measured it. First, we discuss the importance of defining the dimensionality of the trait. Then we discuss the importance of interval data when using parametric statistics to assess or describe a human trait.

Reading as a Unidimensional Construct

Most constructs in the social sciences are multidimensional, yet the instruments used to measure those constructs assume that the trait being investigated is unidimensional. For a trait to be unidimensional, the underlying assumption is that the ability one is measuring lies on a single axis like weight or length. In the case of length, it is possible to measure a child's height as he or she progresses from being an infant to reaching adulthood and use a predetermined standard of measure (e.g., centimeters or inches) to know when any growth occurred. In the context of measuring reading, no "ruler" exists, so researchers must either create or choose a test to act as the standard of measure. Despite the fact that most test theories in use are based on the assumption of unidimensionality (Ip, 2010), many researchers create instruments without an a priori definition of how they are treating the trait in a unidimensional way and are unaware of the measurement implication of that omission.

For example, when discussing the size of a shirt, it is possible to discuss it multidimensionally, bidimensionally, and unidimensionally. Treating a shirt as a multidimensional object requires measuring neck size, arm length, torso width, and length. These measurements allow for a shirt to be made with a fit that is ideal for one individual, but that size or constellation of measurements is so personalized that it is usually not useful in generalizing to the broader population. Often, dress shirt sizes are treated bidimensionally by measuring two dimensions: neck size and arm length. Using the two measurements in tandem can still result in shirts that fit an individual fairly well, but there are well over 20 different neck-sleeve combinations. These two measures reduce the complexity to a degree that is beneficial to the customer, but the sheer number of combinations can still be too complex for describing a population of shirt customers. Therefore, shirt size can be treated as a unidimensional object with an a priori definition of what the size construct is. As neck size increases, typically arm length and torso length increase as well, so shirt sizes have been operationalized with the labels "small," "medium," and "large." The likelihood of a shirt fitting each individual well decreases when treated unidimensionally, yet the practicality of being able to generalize to broader populations can make this an attractive alternative. This shirt analogy lends itself nicely to the discussion of the proficiency scales as they reflect an amalgamation of traits

that co-occur at levels of increasing difficulty.

Reading requires many skills that interact, yet the theory upon which most research has been conducted has not addressed that complexity and has treated all combinations of questions, question types, and passages as tapping a unidimensional construct. However, if one is to treat the reading skill as unidimensional, it is necessary to carefully define the construct in a way that the contributing skills co-occur in specific alignments as the trait increases in complexity. As with the shirt analogy, in which a shirt does not qualify as a size medium unless it has the minimum neck, sleeve, chest, and torso dimensions expected of a medium, a person does not qualify as an Advanced-level reader unless he or she consistently meets all of the criteria described for that level. Thus to test someone at the Advanced proficiency level requires the use of test items in which the skills needed to complete the Advanced reading tasks co-occur. For example, we could discuss individual components such as reading rate or knowledge of sight words, text characteristics, or reader task; however, the proficiency guidelines contain a unidimensional construct that is an amalgamation of the criteria that co-occur at each specific level of the reading proficiency scales.

Measuring Unidimensional Constructs

In classical test theory, an examinee's ability is estimated by the total number of test questions answered correctly, and an item's difficulty is calculated by dividing the number of examinees who answer the item correctly by the total number of examinees. With this classical approach, both the ability score and the item difficulty index measures are dependent on which examinees take the test and on the difficulty of the items included in the test. A further concern is that for the scores to be considered interval data, the distance between any two scores must be equidistant, and any increase in total score should represent the same increase in ability regardless of where it falls within the range of the observed difficulties. Thus, if one examinee took a pretest and had a score of 10, and later took a posttest and had a score of 15, it would be assumed that the examinee gained in skill by five points. To be interval data, that ability increase of five should have the same meaning whether it is from two to seven or from 18 to 23, yet most educators would agree that the amount of growth between those scores is different. Most social scientists acknowledge that the data from their test scores are not truly interval data and are in fact ordinal, but some ignore this requirement and still use parametric statistics in answering their research questions. However, because each of the easiest and each of the most difficult items contribute the same value to the examinee's total score, it is unlikely that the resulting scores have the properties of interval data.

A better approach is to use Rasch measurement. When a Rasch model of item response theory measurement is used, the raw scores are converted to interval-level data and the parameter estimates for both person ability and item difficulty have the quality of measurement invariance (Engelhard, 2008). That is, when using Rasch statistical analyses to measure a unitary construct, person ability estimates are the same regardless of the difficulty of the items that are presented to the examinees, and the item difficulty estimates are the same regardless of the ability levels of the examinees who respond to them. These properties are highly valuable when trying to validate a criterion-based scale; they also address a major criticism of research in the human sciences that the data have been treated as interval data when they are more likely ordinal (Bond & Fox, 2007).

When Rasch calculations transform person ability and item difficulty estimates into interval data, the resulting measurement units are called logits. Logits (or log odds ratios) are the natural logarithm of odds ratios of success. By being transformed to a log odds ratio, the measures are now

interval data, have additive properties, and can then be transformed back into probabilities (Linacre, 1991). Georg Rasch, the Danish mathematician who developed the measurement model, stated:

> In simple terms, the principle is that a person having a greater ability than another person should have the greater probability of solving any item of the type in question, and similarly, one item being more difficult than another means that for any person the probability of solving the second item is the greater one. (Rasch, 1960, p. 117)

With Rasch measurement, person ability and item difficulty are measured conjointly so that an examinee with a person ability estimate of a given logit value will have a 0.50 probability of correctly answering an item with a difficulty parameter of that same value. For instance, if an examinee has a person ability estimate of 1.00 logits and a question has an item difficulty parameter of 1.00 logits, the probability that that person will respond to that prompt correctly is 1 to 1 for 50–50 odds of answering correctly. The result of this transformation is that if an examinee has a pretest with a logit of −0.50 and a posttest of 0.25, the trait ability has increased by 0.75 logits. As logits are interval, that 0.75 difference would indicate the same difference in ability, be it from −3.00 to −2.25 or from 1.75 to 2.50.

Finally, Rasch measurement provides additional tools to use in determining the reliability of test scores. Reliability is the ratio of the true variance to the observed variance. Unlike classical test theory that only reports a single reliability value across all of the items and examinees in a test setting (e.g., Cronbach's alpha or Kuder-Richardson 20), Rasch reliability reports the relative reproducibility of the test's results. Furthermore, Rasch reliability provides those estimates for every facet that is being measured. When the reliability is close to 1.0, it indicates that the observed variance of what is being measured (person or item) is close to or nearly equivalent to the immeasurable true variance. Thus, if person reliability is close to 1.0, it means the differences in examinee scores are due to differences in examinee ability. If the item separation reliability is close to 1.0, it is an indication that differences in the item difficulty parameters are due to differences in item difficulty.

Operationalizing the Guidelines for Item Writers

In order to create tests that could validate the scales for this study, items needed to be written that were both based on the guidelines and aligned with the construct's unidimensional definition. In order to accomplish this, there needed to be a definition of proficient reading that spanned the entire scale, and item writers also needed training on how to select texts and write items that aligned with the scale.

Proficient Reading

Proficient reading was defined as the active, automatic, far-transfer process of using one's internalized language and culture expectancy system to efficiently comprehend an authentic text for the purpose for which it was written (see Figure 1). This definition of proficient reading can be applied to each ACTFL proficiency level and allows the definition of multistage manifestations of the unidimensional trait that conform to developmental theory upon which the guidelines are based.

FIGURE 1

Definition of Proficient Reading

Proficient reading: The active, automatic, far-transfer process of using one's internalized language and culture expectancy system to efficiently comprehend an authentic text for the purpose for which it was written.

Aligning Task, Condition, and Accuracy to the Scale

That definition of proficient reading was then applied to each level described in the proficiency guidelines to ascertain the tasks and conditions that would be assessed at each level. 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 was defined as a separate performance standard with its own combination of aligned communication expectations. These specifications are summarized in Figure 2, where the author's purpose, based on Child's (1987) text modes, is aligned with the types of texts typically used to accomplish those tasks and where the reader's task is aligned with the author's purpose for creating the text. The Novice level is not included in the figure, as it is characterized by the reader's inability to sustain the Intermediate level. The Distinguished level is included to provide a ceiling for the Superior level, but at this point there have been no requirements to test at that level.

As with the other skill modalities described in the ACTFL Proficiency Guidelines, the reading proficiency ratings are noncompensatory. To qualify for a given

FIGURE 2

Overview of Purpose, Text, and Task Alignment by Level

Level	Author Purpose	Text Type	Reader Task
Intermediate 1	**Orient** by communicating main ideas.	Simple, short sentences with simple vocabulary. Often, sentences may be resequenced without changing the meaning of the text. Text organization is loose without much cohesion, but follows societal norms.	Orient oneself by identifying the main topics, ideas, or facts.
Advanced 2	**Instruct** or inform by communicating organized factual information.	Connected factual discourse with compound and complex sentences dealing with factual information. Sentences are sequenced within cohesive paragraphs, but some paragraphs may be resequenced without changing the meaning of the text. The identity of the author is not important.	Understand not only the central facts, but also the supporting details such as temporal and causative relationships.
Superior 3	**Evaluate** situations, concepts, and conflicting ideas; present and support arguments and/or hypotheses with both factual and abstract reasoning, using language that is often accompanied by the appropriate use of wit, sarcasm, irony, or emotionally laden lexical choices.	A multiple-paragraph block of discourse on a variety of unfamiliar or abstract subjects such as might be found in editorials, official correspondence, and professional writing. References may be made to previous paragraphs, external events, common cultural values, etc. The "voice" of the author is evident.	Learn by relating ideas and conceptual arguments. Comprehend the text's literal and figurative meanings by reading both "the lines" and "between the lines" to recognize the author's tone and infer the author's intent.
Distinguished 4	**Project** lines of thought beyond the expected, connect previously unrelated ideas or concepts, and present complex ideas with nuanced precision and virtuosity—often with the goal of propelling the reader into the author's world of thought.	Extended discourse that is tailored for the message being sent and for the intended audience. To achieve the desired tone and precision of expression, the author will often demonstrate the skillful use of low-frequency vocabulary, cultural and historical insights, and an understanding of the audience's shared experiences and values.	Read "beyond the lines" to understand the author's sociolinguistic and cultural references, follow innovative turns of thought, and interpret the text in view of its wider cultural, societal, and political setting.

rating, the individual must consistently meet all of the construct criteria for that level (Swender & Vicars, 2012, p. 5). In assessing reading proficiency, this requirement means that the reader must be able to consistently comprehend texts of the specified type for the purpose for which they were created. Stated another way, the reader must successfully accomplish those comprehension tasks that are aligned with the author's purpose. Blended or nonaligned combinations of reader and author 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 where consistent performance across the aligned factors for each level is expected. Thus, testing whether the reader can perform lower-level reading tasks on higher-level text passages provides insufficient information to assign a proficiency rating. For example, understanding the main idea (an Intermediate-level task) of a Superior-level text would not qualify the reader for a Superior or even an Advanced proficiency rating. Again, the expectation of the reading guidelines is that the author's purpose and the reader's task are congruent. If the author's purpose is to narrate a story, then the reader's 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 author's purpose distinguishes assessment practices from teaching practices, because proficiency testing places the focus on *whether* readers can comprehend texts rather than on the instructional focus of *how* readers comprehend texts at each of the tested levels.

Background Summary

The reading proficiency scales have been in use for a number of years as a practical way to describe overall language ability level, yet little research has been found to validate their use. The research that has been conducted seems to contradict the practical findings of those who use the guidelines as the basis for their tests. The reasons for the failure to validate operational experience could come from a number of factors, including a failure to understand the scale, a failure to define the construct unidimensionally, applying an inappropriate measurement theory, and not having a clear definition of proficient reading that spans all levels of the scale. In order to research the extent to which test items (where both the passage and the question are based on the ACTFL Reading Proficiency Guidelines) ascend in a hierarchy of difficulty levels, those factors would need to be controlled.

Method

To investigate the validity of the reading scales and to answer the research question about the extent to which test items carefully based on the ACTFL Reading Proficiency Guidelines ascend in a hierarchy of difficulty levels, reading test items targeted at the major levels of Intermediate and higher were created in three different languages (English, Chinese, and Spanish).

The item-writing process included training item writers to create items that aligned with the scale. Once developed, the test development team reviewed the items for alignment with the targeted proficiency level and trialed with a small representative sample of examinees. The final step was empirical testing of the items to determine whether their statistical difficulties clustered by level and were arrayed in the hypothesized order. The process was conducted first in English for reading tests requested by NATO and was later replicated in Chinese and Spanish. A description of each instrument, the number of questions targeting each major proficiency level, and the subjects who took the test are presented in the Results section below. 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 could be developed that aligned with the ACTFL/ILR/NATO reading

proficiency guidelines, the following procedure was followed in the three different languages. First, item writers were taught about the construct of reading 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 texts for each level, to write questions that aligned the reader's task with the author's purpose, and to create plausible response distractors that aligned with each level's expectations.

Internal Validation

Before the test items were administered, they went through another review process where the item writers were asked to ensure that the questions being asked were aligned with the author's communicative purpose, that the reading passage was aligned with the typical characteristics of texts used 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:

- At-level difficulty
- 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 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 subsequently administered to learners of varying ability levels.

Data Analysis Procedures

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 measurement 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 requires that three conditions be met:

- The items designed to measure specific proficiency levels should cluster at their intended difficulty levels.
- The mean difficulty 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.

After each test was analyzed for reliability through person and item separation statistics with Rasch measurement, the following procedures were used to test for these conditions. For the English and Chinese studies, the statistical test used to determine if the intended item difficulty logits were different 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% confidence interval (CI), and the p value for each comparison. Because the specifications for the Spanish test included only two proficiency levels, an independent samples t test was used with those data.

The null hypothesis was that there would be no difference in the mean difficulty between the groups. 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.

Language Study Results

Items were created for tests in three languages: English, Chinese, and Spanish. The tests were created for different projects

and thus 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.

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 NATO version of the test was finished in 2009, and in 2010 it was administered to 182 NATO personnel from 12 different nations. The test was later administered to an even greater number of students in university settings. The test for this study consisted of three ACTFL proficiency levels: Intermediate with 24 items, Advanced with 21 items, and Superior with 19 items. Items were drawn at random, and each examinee took 20 of the 24 items at the Intermediate level, 20 of the 21 items at the Advanced level, and all[1]19 items at the Superior level. The test was administered to a total of 581 examinees from two distinct populations: personnel associated with NATO and students enrolled in U.S. universities.

The English test had a Rasch IRT person reliability of 0.87, indicating a relatively 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 0.99 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 levels progressed monotonically (see Table 1).

Comparisons using Bonferroni's 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 3 shows the differences between the levels. Although some of the items between the levels had difficulty logits that did overlap, as a whole the items by level were statistically different. The null hypothesis that the difference in the means between the groups was zero could be rejected for all comparisons. The effect size among the three levels was very strong between the levels as they progressed: for Intermediate items vs. Advanced items, Cohen's $d = 2.26$, 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.

Chinese Study

This test was initially designed for students enrolled in U.S. universities studying Chinese as a foreign language. The subjects were 211 students from 10 different schools. The test had 70 items that were divided into

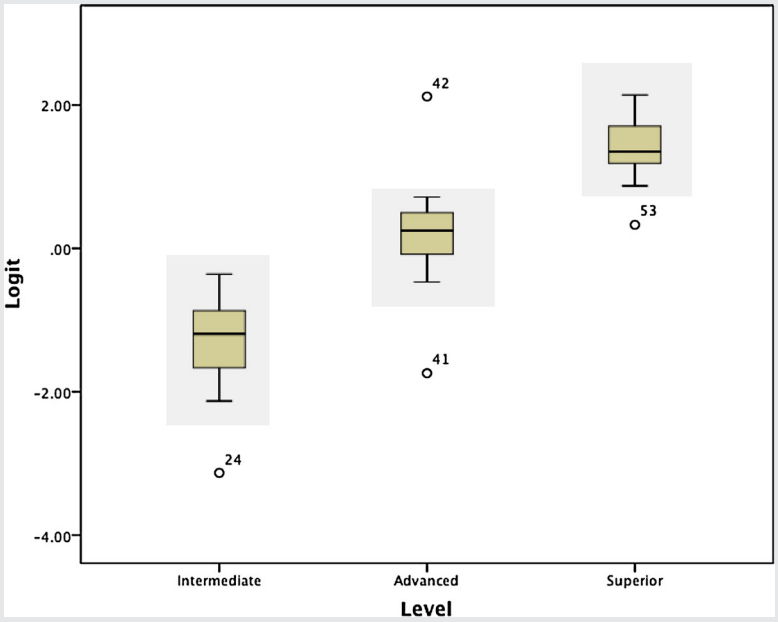
TABLE 1

Descriptive Statistics of English Item Difficulty by Intended Proficiency Level

	# of Items	Mean Logit Value	*SD*
Intermediate	24	−1.28	0.63
Advanced	21	0.19	0.69
Superior	19	1.42	0.46
Total items	64		

FIGURE 3

Boxplot Distribution of English Item Difficulty by Intended Proficiency Level



the three proficiency levels: Intermediate, Advanced, and Superior. The items were administered adaptively in five-item testlets, and a sufficient number of students answered questions at all levels to calculate comparative statistics. There were six testlets (a total of 30 items) for both the Intermediate level and the Advanced level, and two testlets (a total of 10 items) for the Superior level. Students needed to provide evidence of sustained performance at one level before they were presented testlets at a higher level. Students only answered a subset of the testlets based on their performance at each level, and the greatest number of items that any student encountered was 55. The minimum number of items any student would encounter was 15, and in that instance, all of the items were at the Intermediate level.

The test had a fairly high internal consistency with a Rasch person reliability of 0.83, indicating that the examinees could be reliably divided into two or three statistically distinct groups (Linacre & Wright, 2009). The item reliability of 0.92 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).

Comparisons using Bonferroni's contrasts found statistically significant differences between the Intermediate and Advanced items (mean difference = −1.44, with the 95% CI between −2.35 and −0.53 and $p < 0.01$), and the Advanced and Superior items had a mean difference of −2.14, with a 95% CI between −3.42 and −0.86 and $p < 0.001$. Figure 4 shows the differences between the levels. Although Intermediate-level items had a slightly skewed distribution, it was not so much that parametric statistics were inappropriate. As with the English test, some of the items between the levels had difficulty logits that overlapped, but as a whole, the items by level were statistically different. The null hypothesis that the difference in the means between each of the groups was zero could

TABLE 2

Descriptive Statistics of Chinese Item Difficulty by Intended Proficiency Level

	# of items	Mean logit value	*SD*
Intermediate	30	−1.22	1.93
Advanced	30	0.22	0.83
Superior	10	2.35	0.86
Total	70		

be rejected. The effect size among the three levels was very strong: for Intermediate items vs. Advanced items, $d = 0.98$, and for the Advanced items vs. Superior items, $d = 2.50$. Thus the intended proficiency level of the items had a strong effect on the empirical item difficulty level.

Spanish Study

This test was initially designed for students enrolled in U.S. universities studying Spanish as a foreign language and was designed to measure only through the Advanced level. The subjects were 550 students from four different schools. The test had 57 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. Each form shared five Intermediate items and five Advanced items for a total of 10 anchor items on both forms.

FIGURE 4

Boxplot Distribution of Chinese Item Difficulty by Intended Proficiency Level

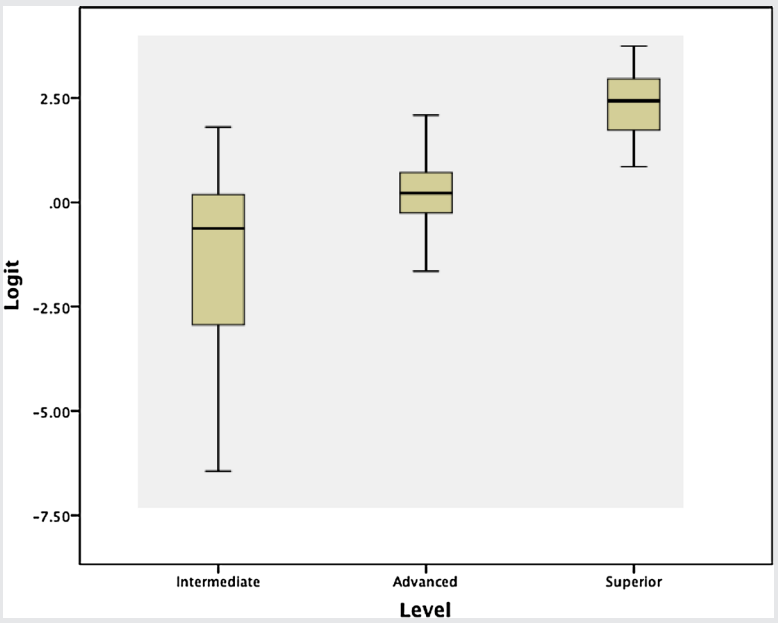


TABLE 3

Descriptive Statistics of Spanish Item Difficulty by Intended Proficiency Level

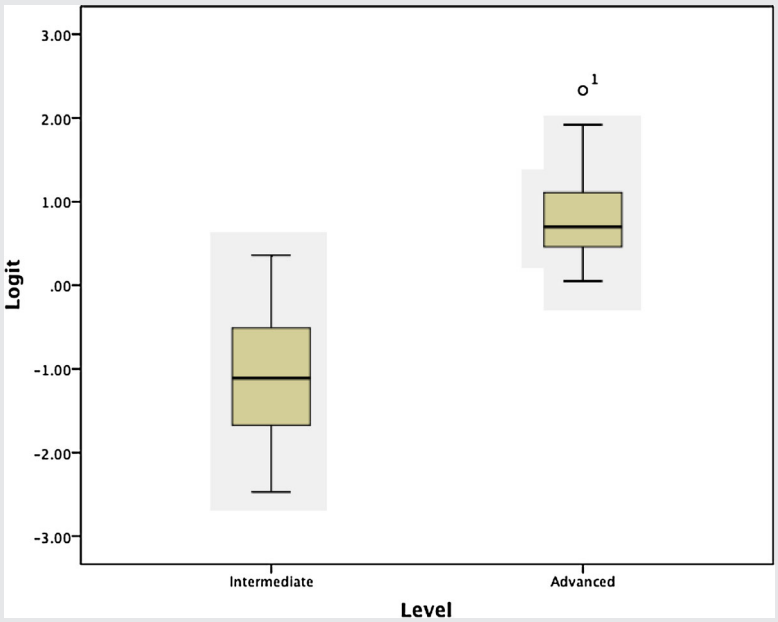
	N	Mean	*SD*
Intermediate	24	−1.07	0.77
Advanced	33	0.77	0.54
Total	57		

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 0.98 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).

An independent samples *t* test between the Intermediate and Advanced items was conducted to determine if the two groups of items differed in item difficulty. An examination of the data indicated that while the Intermediate items were normally distributed, the Advanced items were slightly skewed (see Figure 5); thus Welch's procedure was used. The 95% CI for the difference in the means was between −2.20 and −1.50 ($t = -9.99$, $p < 0.001$, $df = 38.9$). The null hypothesis that the difference in the means was zero could be rejected, and the items' intended proficiency level had a strong effect ($d = 2.77$) on the empirical item difficulty.

FIGURE 5

Boxplot Distribution of Spanish Item Difficulty by Intended Proficiency Level



Language Studies Summary

This research explored the extent to which test items ascend in a hierarchy of difficulty levels when both the passage and question are based on the ACTFL Reading Proficiency Guidelines. In each of the three studies, the intended proficiency levels of the items resulted in statistically significant different item difficulty levels. This means that we can reject the null hypothesis that there is no difference between the levels of the guidelines and can support the argument that the guidelines ascend in a hierarchy of difficulty.

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, as this study was focused on the items and their intended alignment, poorly functioning items were not excluded. As would be expected when dealing with multiple-choice questions, some of the items in these studies did not function as intended. 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 much easier than the intended proficiency that the examinees were drawn to the correct response even though the question and passage were beyond the students' ability. This led to some items being easier than expected. To avoid any perception of bias, these items were not eliminated from our analyses. Had the malfunctioning items been removed from the analyses—as would have been routinely been done in typical test development projects—the differences between the empirical difficulty levels would have been even greater.

For instance, with the Chinese test, if we were to eliminate 10 items from the Intermediate and 10 items from the Advanced item pools, the remaining data would have had even greater separation between the levels (see Table 4) with very little possibility of the items from the intended levels overlapping with each other (see Figure 6). 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.

Discussion

The results of this study support the difficulty hierarchy posited by the ACTFL, ILR, and related reading proficiency guidelines. It is not clear why previous analyses did not obtain these results, but any of the following factors may have obscured the relationships found in this study:

- Failure to align the task and conditions as described at each skill level
- Failure to apply noncompensatory, criterion-referenced scoring
- Failure to include subjects with a full range of abilities

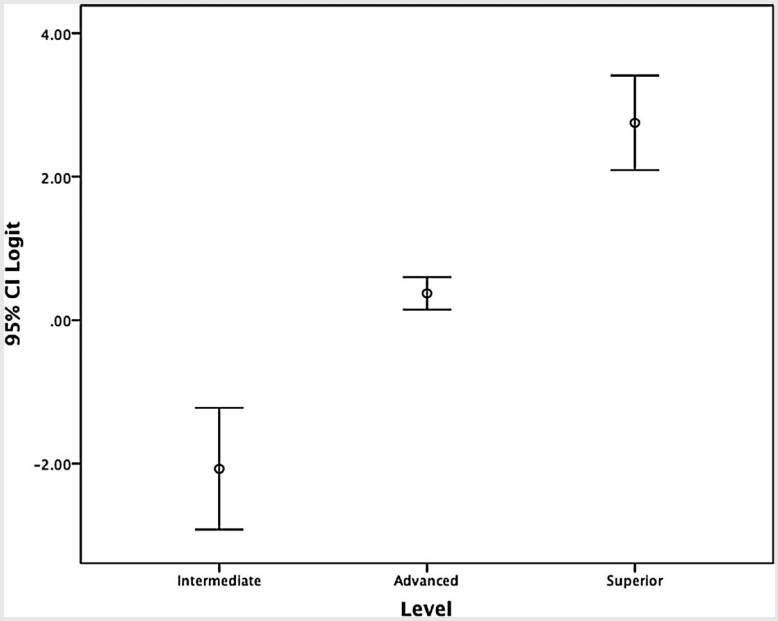
TABLE 4

Descriptive Statistics of Chinese Item Difficulty by Intended Proficiency Level

	N	Mean	*SD*
Intermediate	20	−2.07	1.81
Advanced	20	0.37	0.48
Superior	10	2.75	0.92
Total	70		

FIGURE 6

Chinese Item Difficulty Means With 95% Confidence Interval by Intended Proficiency Level



- Failure to convert the results to interval data before conducting comparative analyses

For instance, it appears that in the research conducted by Lee and Musumeci (1988), questions of varying task levels were associated with each text sample and that level-by-level criterion scoring was not applied (p. 175). In addition, their Figure 6 (p. 179) indicated that while the researchers hypothesized that the students would improve a full proficiency level with every semester of language study, the percentage of correct responses by level would suggest that the only level where the students demonstrated sustained performance was the Intermediate level, or Level 1. If the higher-level questions were beyond the students' sustained abilities, their scores on those items might have represented the problem-solving skills observed by Rupp et al. (2006)—and not the relative difficulty of the texts and/or the reading tasks presented.

Implications for Testers and Instructors

Testers wishing to assess curriculum-independent, real-world reading proficiency according to the ACTFL Proficiency Guidelines 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 both that those criteria are aligned with the author purpose, text characteristics, and reader tasks described for that level and that non-compensatory, criterion-referenced scoring criteria are used when assigning ratings.

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 provide useful ramps

and scaffolding that can help students 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 and master difficult or complex concepts or skills by selectively teaching the components of the targeted new ability one piece at a time. Then, after the pieces are in place, learners are asked to integrate those skills to carry out a more complex and challenging task. 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 the known (what they currently know or are able to do) to the unknown (what they yet need to know or do).

For example, a group of students may be able to understand the main idea of sentence-length questions and answers about topics in their immediate surroundings, but they are not yet able to understand the details of real-world communications such as newspaper articles. To be able to understand authentic news reports, those students will have to acquire the lexicon needed to describe real-world events, learn the verb tenses needed to place events into correct temporal relationships, and develop the morphological sensitivities needed to comprehend the relationships described. Instructors will often teach each of these contributing skills separately and give achievement tests to assess students' progress in each area. This teaching approach has clear pedagogical advantages, and the tests of the isolated skills can have diagnostic value. However, the ability to pass a test on the vocabulary used in a report of a particular current event does not prove that the student can read articles on that topic: Even the mastery of all the contributing skills does not guarantee that the learner has developed and integrated all of the abilities needed to efficiently read and comprehend newspaper articles. While each of these multiple contributing skills is necessary, they remain enabling skills rather than proof of general reading proficiency. Thus, there is a role for both classroom-based progress tests where the elements being tested are not fully aligned with the targeted proficiency level and for culminating general proficiency tests where author purpose, text type used, and reader task are all aligned.

From a proficiency perspective, the fact that a reader can get the main idea (an Intermediate-level task) of a text generated for an Advanced communication purpose does not indicate Advanced reading ability. Therefore, instructors should periodically administer proficiency tests to evaluate the effectiveness of the instructional scaffolding techniques in promoting an increase in general reading proficiency. When the reader can accomplish Advanced comprehension tasks at the Advanced level, there is evidence of Advanced reading proficiency.

Future Research

The authors are proceeding to apply the lessons learned while validating the ACTFL Reading Proficiency Guidelines to create tests of reading proficiency in multiple languages. Inherent in that development process is the need to conduct theoretical research related to the assigning of sublevel proficiency scores, as well as practical research into the optimal design of computer-adaptive tests of reading proficiency.

Note

1. There were originally 20 Superior items; however, one of the passages was eliminated because of changes in global politics that made it culturally inappropriate.

References

ACTFL. (2012). *ACTFL Proficiency Guidelines 2012*. White Plains, NY: Author.

Alderson, J. C. (2000). *Assessing reading*. Cambridge, UK: Cambridge University Press.

Bernhardt, E. (2011). *Understanding advanced second-language reading*. New York: Routledge.

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

Brown J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge, UK: Cambridge University Press.

Child, J. (1987). Language proficiency levels and the typology of texts. In H. Byrnes & M. Canale (Eds.), *Defining and developing proficiency: Guidelines, implementations and concepts* (pp. 97–106). Lincolnwood, IL: National Textbook.

Clifford, R. (1981). Convergent and discriminant validation of integrated and unitary language skills: The need for a research model. In A. Palmer, P. Groot, & G. A. Trosper (Eds.), *The construct validation of tests of communicative competence* (pp. 62–70). Washington, DC: Teachers of English to Speakers of Other Languages.

Engelhard, G. Jr (2008) Historical perspectives on invariant measurement: Guttman, Rasch, and Mokken. *Measurement*, *6*, 155–189.

Herzog, M. (n.d.). How did the language proficiency scale get started? Retrieved December 15, 2012, from http://www.govtilr.org/Skills/IRL%20Scale%20History.htm

Interagency Language Roundtable (ILR). (2012). *ILR language proficiency skill level descriptions*. Retrieved April 2, 2013, from http://www.govtilr.org/skills/ilrscle1.htm

Ip, E. H. (2010). Empirically indistinguishable multidimensional IRT and locally dependent unidimensional item response models. *The British Journal of Mathematical and Statistical Psychology*, *63*, 395–416.

Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. New York: Routledge.

Lee F. L., & Musumeci, D. (1988). On hierarchies of reading skills and text types. *Modern Language Journal*, *72*, 173–185.

Lee, Y. W. D. (2001). Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology*, *5*, 37–72.

Lewin, K. (1952). *Field theory in social science: Selected theoretical papers by Kurt Lewin*. London: Tavistock.

Linacre, J. M. (1991). Log-odds in Sherwood Forest. *Rasch Measurement Transactions*, *5*, 162–163 (Electronic version). Retrieved April 8, 2013, from http://www.rasch.org/rmt/rmt53d.htm

Linacre, J. M. (2012). *Winsteps®* (Version 3.75.0) [Computer Software]. Beaverton, Oregon: Winsteps.com. Retrieved January 1, 2012. Available from http://www.winsteps.com/

Linacre J. M., & Wright, B. D. (2009). *A user's guide to WINSTEPS*. Chicago: WINSTEPS.

Luecht, R. (2003). Multistage complexity in language proficiency assessment: A framework for aligning theoretical perspectives, test development, and psychometrics. *Foreign Language Annals*, *36*, 527–535.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedogogiske Institut.

Rupp, A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing*, *23*, 441–474.

Smith, G. (2007). *Understanding reading: A psycholinguistic analysis of reading and learning to read* (6th ed.). Mahwah, NJ: Lawrence Erlbaum.

STANAG 6001. (2010). *NTG—language proficiency levels* (4th ed.). Retrieved April 8, 2013, from http://www.ncia.nato.int/Opportunities/BizOppRefDoc/Stanag_6001_vers_4.pdf

Surface E., & Dierdorff, E. (2003). Reliability and the ACTFL oral proficiency interview: Reporting indices of interrater consistency and agreement for 19 languages. *Foreign Language Annals*, *36*, 507–519.

Swender E., & Vicars, R. (2012). *ACTFL oral proficiency interview tester training manual*. White Plains, NY: ACTFL.

Submitted December 26, 2012

Accepted January 18, 2013