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

A Rationale for Criterion-Referenced Proficiency Testing

Ray Clifford
Brigham Young University

Abstract: *This article summarizes some of the technical issues that add to the complexity of language testing. It focuses in particular on the criterion-referenced nature of the ACTFL Proficiency Guidelines—Speaking; and it proposes a criterion-referenced interpretation of the ACTFL guidelines for reading and listening. It then demonstrates how using criterion-referenced testing and scoring enhances the accuracy of reading and listening proficiency ratings while also providing informative feedback to learners.*

Key words: *assessment, criterion-referenced, listening, proficiency, reading*

Introduction

In the field of second language learning, there is growing interest in assessing not only speaking and writing proficiency but also learners' proficiency in reading and listening comprehension. As the profession begins to more rigorously assess reading and listening proficiency levels, it will be crucial that appropriate test design and test scoring procedures are applied.

A major reason for testing is to make classification decisions. Cronbach and Glaser made this point in the introduction to their now-classic 1957 textbook; yet the prevalence of classical statistical procedures based on average results, curves, and standard deviations has obscured the fact that there is a second approach to testing referred to as Criterion-Referenced testing. This second approach to making test-based decisions contrasts with the historically predominant purpose for testing—and therefore with the procedures for test design, construction, scoring, and score interpretation that are most commonly used today. It is, therefore, useful to begin by comparing these two distinct philosophical approaches to making test-based decisions and to highlight how those differences in purpose are reflected in the design, construction, scoring, and interpretation of scores for each type of test.

Norm-Referenced and Criterion-Referenced Tests

In classroom settings, the most frequently used type of test compares test takers against each other and arrays those test takers by their scores from highest to lowest

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

Foreign Language Annals, Vol. 49, Iss. 2, pp. 224–234. © 2016 by American Council on the Teaching of Foreign Languages.
DOI: 10.1111/flan.12201

or vice versa. This approach is particularly applicable when the intent is to grade the test takers on a curve, as is often done for curriculum-based, classroom tests. Because the test takers are compared against others, and their relative standing is dependent on the general, or normal, ability level of their comparison group, this type of test is called a norm-referenced (NR) test. NR tests compare students against other students, and because the purpose of NR tests is to compare people against other people, the items that are selected for those tests are primarily chosen based on their ability to distinguish between test takers of varying abilities. Items may be selected to cover specific topical areas, but if everyone answers an item correctly, that item is discarded because it does not discriminate among learners; that is, it does not separate the "best" from the "rest."

In contrast, criterion-referenced (CR) tests do not compare test takers against each other. Rather, CR tests compare test takers against a set of clearly stated expectations or criteria. To provide enough information to assess whether someone has met a specific criterion, the expectations, or criterion statement, must describe three elements: (1) the task to be completed; (2) the conditions or contexts in which the task is to be performed; and (3) the performance standard, or level of success or accuracy, that is required. Because test takers' scores are not compared with or referenced to others' level of mastery, a CR assessment may reveal that all students, some students, or even no students actually meet the stated expectations and thus pass or fail a CR test.

Both NR tests, designed to compare test takers against each other, and CR tests, designed to measure test takers' knowledge and skills against a predetermined performance standard, can be used in instructional settings: NR tests are more commonly used to assess course-specific learning and to assign course grades, while CR tests can be used to assess mastery of specific learning outcomes as well as curriculum-independent skills and higher-order, program-level instructional results. Because of their independence from curriculum and instruction as delivered in a particular teaching/learning context, CR tests can be used to compare the abilities of students from different classes as well as students with different learning experiences against a common set of external ability expectations (Brown & Hudson, 2002; Shrock & Coscarelli, 2007; Smith & Stone, 2009).

The difference in the testing purposes of CR and NR tests drives differences in test design and in scoring procedures. Table 1 summarizes some of the major differences between a typical NR classroom test and a CR proficiency assessment. The ACTFL Oral Proficiency Interview (OPI) was selected as an example of a CR test because of its wide use in educational settings. As Table 1 indicates, the main distinguishing features of CR tests are as follows:

- The purpose of a CR test is to compare each test taker's ability against stated criteria for language use in real-world and curriculum-independent situations, rather than against the ability level of other test takers.
- Success requires a sustained, spontaneous ability to perform the communicative tasks that are specified, in the contexts that are described, and with the degree of accuracy that is expected.
- Language educators assign ratings by assessing the test taker's ability at each criterion level and then identifying the highest level of sustained ability.
- The level-by-level results are not totaled or averaged across the levels covered by the test.

The OPI is specifically designed to be "a criterion-referenced, rather than a norm-referenced, assessment" (Swender & Vicar, 2012, p. 1). The summary chart on page 14 of that manual shows how each of the major levels in the ACTFL (2012) speaking proficiency guidelines describes a unique

TABLE 1

Comparison of Typical NR and CR Test Characteristics

	Example	Testing Purpose	Domains Covered	Type of Learning Tested	Scoring Procedures
NR Tests	Chapter test in a textbook	Compare a student's ability with the ability scores of other students	Curriculum-based content	Direct application of rote learning and rehearsed material	Generate a single compensatory total score or an average test score across several sections or skills
CR Tests	OPI	Determine whether the test taker has specified real-world abilities	Curriculum-independent, real-world content	Far-transfer learning that enables spontaneous and creative responses	Generate a noncompensatory, floor and ceiling proficiency rating that treats each skill or skill level separately

combination of the three essential elements of communication—task, conditions, and accuracy—while the fourth column, text type, further defines the conditions associated with the communication tasks. That chart is reproduced in Table 2.

The note in Table 2 establishes the need for a rating process, referred to as "floor" and "ceiling" rating. To understand how the concepts of floor and ceiling are applied in the testing of language proficiency, it helps to understand that the proficiency levels represent a hierarchy of ability levels, where each succeeding level is more demanding than the level below it. In that hierarchical array of ascending difficulty levels, the highest level at which the test taker shows mastery of all the criteria for that level is called that person's floor level. Everyone has their own floor level, which can change as one's language skills improve (or degenerate). A test taker's ceiling level is the next-higher level above that person's floor level—i.e., the level where the test taker no longer meets all of the criteria for the level.

For example, a test taker might show sustained ability at the Advanced level but not meet one or more of the criteria for a Superior rating. That test taker's floor level would be Advanced, and his or her ceiling level would be Superior. In applying floor and ceiling ratings, it is important to note that CR ratings are noncompensatory. Test takers can be considered successful only if their language sample meets the full expectations—the criteria for task, conditions, and accuracy—for a given proficiency level. For this reason, the scoring, or rating, of CR tests is noncompensatory; that is, meeting the expectations for even two of the three components, e.g., task and accuracy but not conditions, at any given level of performance is not sufficient to earn that proficiency rating. What is more, even if test takers exceed expectations in one of the areas, such as accuracy, that higher-than-required performance is not sufficient to compensate for failure in another area. For that reason, rating principle number 5 in the OPI manual states, "In evaluating a speech sample, testers

TABLE 2

Assessment Criteria—Speaking

Proficiency Level	Functions or Global Tasks	Context/Content	Accuracy/Comprehensibility	Text Type
Superior	Discuss familiar and unfamiliar topics concrete and abstractly, support opinions, hypothesize.	Most informal and formal settings/*Wide ranged public-interest topics and some special fields of interests and expertise*	No pattern of error in basic structures. Errors virtually never interfere with communication or distract the interlocutor from the message.	Extended discourse
Advanced	Narrate and describe in all major time frames and deal effectively with an unanticipated complication in a routine situation or transaction.	Most informal and some formal settings/*Topics of personal and general current interest*	Can be understood without difficulty by speakers unaccustomed to dealing with non-native speakers.	Oral paragraphs/ Connected discourse
Intermediate	Create with language, ask and answer simple questions, and handle a simple situation or transaction.	Some informal settings and a limited number of transactional situations/ *Predictable, familiar topics related to daily activities and personal environment*	Can be understood, with some repetition, by speakers accustomed to dealing with non-native speaker.	Discrete sentences and strings of sentences
Novice	Communicate minimally with formulaic and rote utterances and produce words, phrases, and lists.	Most common informal settings/*Most common aspects of daily life*	May be difficult to understand, even for speakers accustomed to dealing with non-native speakers.	Individual words, phrases, and lists

Note: A rating at any major level is determined by **sustained performance** of the functions of the level, within the contexts and content areas for that level with the degree of accuracy described for the level, and in the text type for the level. The performance must be sustained across **all of the criteria** for the level, **all of the time**, in order to be rated at that level.

Note: Table reproduced per Swender & Vicar, 2012, p. 14.

and raters look for patterns of strength (the 'floor' performance) and patterns of weakness (the 'ceiling' performance)" (Swender & Vicar, 2012, p. 15).

What has not been widely recognized is that the ACTFL proficiency guidelines for each of the skill modalities of speaking, writing, reading, and listening follow the same pattern that is found in the speaking guidelines. In each instance, the proficiency guidelines provide a level-by-level set of tasks, conditions, and accuracy expectations that are to be assessed using CR test designs and scoring procedures. Since it is easier to recognize the CR components of the guidelines when they are presented in a chart format, a summary of the ACTFL proficiency guidelines for interpretive reading is presented in Table 3.

It is useful to note that in Tables 2 and 3, only the major proficiency levels are shown. That is because only the major proficiency levels represent an independent set of criteria, where each level is defined by its own unique constellation of task, conditions, and accuracy expectations.

The reality is that language learners may make progress concurrently in multiple levels of the proficiency guidelines. As observed by both instructors and testers, language learners do not completely master the communication tasks and topical domains that are associated with one proficiency level before they begin learning the skills that are required to qualify for the next higher rating. In fact, learners commonly develop conceptual control or even partial control over the skills that are required at the next higher proficiency level by the time they have attained sustained, consistent control over the lower level. This aspect of language learning is evident in the production skills of speaking and writing; what is more, it is also evident in the receptive skills of reading and listening comprehension. Figure 1 depicts this reality: To qualify for any specific level of proficiency, the learner must ascend to the next-higher level of proficiency; the learner must also demonstrate sustained control over the tasks, conditions, and accuracy expectations associated with the lower level(s) of proficiency.

NR and CR Scoring

Unlike CR assessments of speaking and writing, tests of reading and listening comprehension have generally not used CR, or floor and ceiling, approaches to scoring. Instead they have followed the design and scoring procedures associated with NR testing and have reported a single ability score. To then interpret that score, researchers and test developers have applied various statistical procedures to determine the cut scores that would define the ranges that would best equate to each proficiency level being tested (Cizek, 2008; Cizek & Bunch, 2007). This NR scoring model has hindered the validation of the ACTFL proficiency guidelines in the receptive skills (i.e., reading and listening comprehension) for several decades.

However, the fact that language learners develop abilities that extend beyond the level at which they have demonstrated full control is a key reason why floor and ceiling levels have been used in rating speaking and writing proficiency. These procedures have been shown to result in high levels of interrater reliability when testing speaking proficiency (Surface & Dierdorff, 2003) and should in fact be used when assessing listening and writing. The importance of using such procedures is clearly noted in the guidelines themselves: The final column in Table 1 and the note at the bottom of Table 2 focus on the most often overlooked feature of true CR tests—the application of CR scoring procedures. In addition, recent research has demonstrated that CR testing principles (as depicted in Table 3), including level-by-level test design and floor and ceiling scoring, can be applied to the receptive skills of reading and listening comprehension (Clifford & Cox, 2013; Cox & Clifford, 2014). As noted by Luecht (2003), assessing multidimensional traits, such as language proficiency, requires

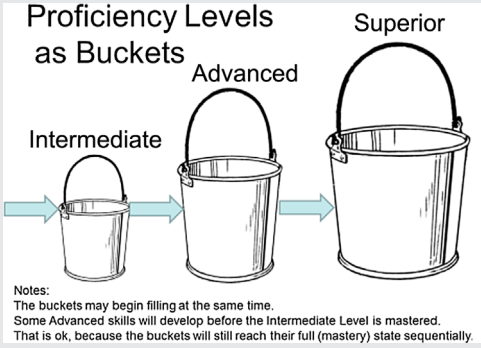
TABLE 3

Assessment Criteria—Reading

Level	Reader Task	Conditions: Author Purpose	Conditions: Text Type	Conditions: Content	Accuracy Expectations
Superior	Understand literal and figurative meanings by reading both "the lines" and "between the lines." Recognize the author's tone and unstated positions. Evaluate the adequacy of arguments made.	Evaluate situations, concepts, and conflicting ideas. Present and support arguments and/or hypotheses with both factual and abstract reasoning.	Multiple-paragraph prose on a variety of professional or abstract subjects such as found in editorials, formal papers, and professional writing.	Multiple, well-organized abstract concepts interlaced with attitudes and feelings. Social/cultural/political issues, with abstract aspects and supporting facts presented as well. Most allusions and references are explained by their context.	Understands the facts; the details; and the author's opinion, tone, and attitude.
Advanced	Understand the facts and supporting details including any causal temporal and spatial relationships.	Convey structured, factual information, supporting details, and factual relationships in extended narratives and descriptions.	News reports, magazine articles, short stories, human interest features, and instructional and descriptive materials.	Concrete information about real-world phenomenon with supporting details, as well as interrelated facts about world, local, and personal events.	Grasps both the main ideas and the supporting details.
Intermediate	Understand the main idea, orient oneself by identifying the topic or main idea.	Orient by communicating one or more general ideas.	Very simple announcements; ads, personal notes.	Information about places, times, people, etc. that are associated with everyday events, personal invitations, or general information.	Recognizes the main idea and some broad, categorical distinctions.
Novice	Recognize some random items in a list or short text.	List, enumerate.	Lists, simple tables.	Sparse or random; format or external context may reveal internal relationships.	Correctly recognizes some words.

FIGURE 1

An Ascending Hierarchy of Proficiency Levels.



three elements: a theoretical construct model, a conforming test development model, and an aligned scoring model.

Following Luecht's guidance, CR tests of language proficiency should not use NR scoring procedures. Shrock and Coscarelli (2007) agreed and made the point that if a CR test is to be legally defensible, then its scoring model must be aligned with the construct to be tested and with the test specifications that define that construct. To apply these CR requirements to the testing of language proficiency, separate criterion specifications must be defined for the proficiency construct defined at each major proficiency level—and each major level of the test must be scored independently of the other levels.

In contrast, the ACTFL sublevels are not independent levels. Rather, they are used to report typical stages of progress within a major floor level. Because each major level is tested separately, an individual's progress at the ceiling level can also be described in incremental stages that culminate in the demonstration of sustained ability—at which point that level becomes a new floor ability level. The following ability categories, represented by the mnemonic REDS, can prove useful in communicating learners' progress toward the ceiling level:

- R: Random ability—a response pattern that is sporadic or the equivalent of random guessing. (In a multiple-choice test, the typical upper boundary of the random category is about 25% correct.)
- E: Emerging ability—a response pattern that is better than chance, but where the accuracy of correct responses is less than 50%. (In a multiple-choice test, the typical upper boundary of the emerging category is about 50% correct.)
- D: Developing ability—a response pattern that demonstrates the ability to correctly answer at least half of the items at a given proficiency level, but the demonstrated accuracy is not high enough to be considered as sustained, consistently accurate performance. (In a multiple-choice test, the typical upper boundary of the developing category is the lower boundary of the sustained category.)
- S: Sustained ability—a response pattern that shows sustained or consistently accurate performance at a given level. When sustained ability is demonstrated at a level that was previously considered a ceiling level, that ceiling level is recognized as a new floor level. (In a multiple-choice test, the typical lower boundary of the sustained category is about 75% correct.)

Of course, the term *sustained ability* requires an operational definition, and that definition may vary from test to test depending on the statistical properties of

the items in that test. However, whether sustained performance levels have been established based on expert judgments that take into account the reality that neither humans nor their levels of performance are completely consistent (Glass, 1978) or by more sophisticated statistical procedures (Janssen, Tuerlinckx, Meulders, & De Boeck, 2000), the results are quite consistent. For well-designed and well-constructed tests, where all of the items are aligned with CR specifications, sustained ability is typically set in the 70 to 80% range.

The following hypothetical example illustrates how failing to use CR scoring procedures can invalidate the results of a test that was designed to provide a CR assessment. Imagine that three candidates for a teaching position (Alice, Beth, and Carl) took a three-part reading proficiency test with items that were specifically designed to cover the Intermediate, Advanced, and Superior proficiency levels on the ACTFL scale. Each part of the test (Intermediate, Advanced, and Superior) had a maximum score of 100 points, and the exam as a whole totaled 300 points. The hypothetical candidates' scores on each part of the test are shown in Table 4. As shown in Table 4, using NR procedures, this multilevel test generated a single, overall score of 190 (63%) for all three candidates. However, using a CR scoring approach—that is, applying floor and ceiling rating procedures to the level-specific scores—reveals that despite having the same total score, Alice, Beth, and Carl actually exhibited different proficiency profiles. Table 5 applies typical values for the REDS categories described above to each person's score on each part of the test.

Just as a floor and ceiling rating procedure adds accuracy and explains important differences in ability profiles when rating OPI tests, applying that same scoring process to a reading test reveals information that would have been hidden if traditional NR scoring procedures had been used. As the labels in the cells of Table 5 indicate, if one applies a success criterion of 75%, all of the candidates exhibited sustained ability on the Intermediate-level portion of the test. For Alice and Carl, that level was their highest level of sustained performance (or floor ability level), and their ceiling level was Advanced. However, Beth demonstrated sustained ability (or a floor ability level) at the Advanced level, with a ceiling level of Superior. Using this CR-based information, an employer looking for someone with Advanced reading proficiency could see that only Beth met that requirement—a finding that was entirely masked when using NR scoring procedures that combined the candidates' level-by-level scores into a single total score.

This simple example shows the impact of applying inappropriate scoring procedures to a CR proficiency test. Because total and average scores are always influenced by every item on a test regardless of each item's

TABLE 4

Test Scores for Three Job Candidates

	Intermediate Score	Advanced Score	Superior Score	Total Score
Alice	95	65	30	**190**
Beth	90	80	20	**190**
Carl	85	65	40	**190**

TABLE 5

Proficiency Test Results for Three Job Candidates

	Intermediate Score	Advanced Score	Superior Score	Total Score
Alice	95 Sustained	65 Developing	30 Emerging	190
Beth	90 Sustained	80 Sustained	20 Random	190
Carl	85 Sustained	65 Developing	40 Emerging	190

targeted proficiency level, test takers' results measured by their total scores are not aligned with those test takers' noncompensatory, level-by-level proficiency ratings. This example also demonstrates the reality that trying to unambiguously derive noncompensatory proficiency ratings from compensatory total scores is always difficult, often inaccurate, and sometimes impossible.

Discussion and Conclusion

Classroom tests may be either NR or CR, depending on the instructors' instructional goals and grading requirements. However, when language proficiency using the ACTFL proficiency guidelines is assessed, CR testing procedures that align with the guidelines' level-by-level tasks, conditions, and accuracy expectations are required. Several critical conclusions can thus be drawn.

1. The ACTFL proficiency guidelines contain level-by-level ability constructs that define communication tasks, the conditions under which the tasks are to be accomplished, and the accuracy of communication that is required at that level. To effectively determine individuals' proficiency levels, those construct models must determine both the design of the test used and the scoring model that is applied.
2. Only CR testing and scoring yields a score for each criterion level tested that is not influenced or diluted by test takers' developing abilities at other levels. For example, even though a test taker can successfully complete a detailed, Advanced-level comprehension task on an Intermediate text, this success does not demonstrate that the reader would be able to comprehend the details of more challenging, Advanced-level texts. Similarly, simply because a test taker can comprehend the topic (an Intermediate-level task) of a Superior-level text does not indicate that the test taker can consistently demonstrate Superior-level reading ability; this small success only demonstrates that the test taker has an emerging ability that lies somewhere between the Intermediate and Superior levels. In stark contrast to conventional NR test item development procedures, every test item on a CR test must be fully aligned with the criterion that the item is designed to measure—and when constructing ACTFL proficiency tests, educators must carefully align every item with the task, conditions, and accuracy expectations of the one level that item is designed to assess.
3. The level-by-level scores generated in an ACTFL proficiency test must be evaluated using floor and ceiling rating procedures to arrive at a noncompensatory result, so that higher-than-required successes in one aspect of the assessment (task or conditions or accuracy) do not mask lower-than-expected performance in one or more of the other areas. Such CR scoring makes transparent how each proficiency rating was derived and

exactly what the test taker's sustainable level of proficiency is.

4. CR assessments and scoring procedures are already applied in the design of OPI and Writing Proficiency Test (WPT) testing procedures and are included in the training of ACTFL OPI and WPT raters. As a result, the advantages of CR testing procedures are evident to those certified as OPI and WPT testers and raters. However, this approach may not yet be uniformly applied when testing listening and reading proficiency.
5. Traditional scoring practices that produce a total or average score introduce compensatory factors that reduce the accuracy and interpretability of CR test results.
6. Only when CR testing procedures are fully implemented using both CR test designs and CR scoring models that are fully aligned with the construct models expressed in the ACTFL proficiency guidelines are the test results maximally accurate and defensible.

Adhering as described above to best practices in test development and scoring through the application of CR floor and ceiling rating procedures provides an additional benefit. By combining their ACTFL sublevel ratings at the floor level with the REDS categories at the ceiling level, CR tests provide learners with more reference points to use in tracking their progress. Instead of providing a test result that indicates only low, mid, and high stages of progress for each major proficiency level, such a two-rating system has considerable diagnostic value in that it explicitly states the test takers' sustained, or floor, proficiency level, and also indicates their progress toward the next higher proficiency level. For instance, learners at the Intermediate level may receive ratings such as:

- Intermediate Low (with random ability at the Advanced level).
- Intermediate Mid (with random ability at the Advanced level).
- Intermediate Mid (with emerging ability at the Advanced level).
- Intermediate High (with developing ability at the Advanced level).

All of the above factors argue for applying the same CR testing and scoring principles when testing reading and listening proficiency that are applied when testing speaking and writing proficiency.

References

ACTFL. (2012). *ACTFL proficiency guidelines—speaking*. Retrieved April 1, 2014, from http://www.actfl.org

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

Cizek, G. J. (Ed.). (2008). *Setting performance standards: Concepts, methods, and perspectives*. Boca Raton, FL: Routledge.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

Clifford, R., & Cox, T. (2013). Empirical validation of reading proficiency guidelines. *Foreign Language Annals*, *46*, 45–61.

Cox, T., & Clifford, R. (2014). Empirical validation of listening proficiency guidelines. *Foreign Language Annals*, *47*, 379–403.

Cronbach, L. J., & Glaser, G. C. (1957). *Psychological tests and personnel decisions*. Urbana: University of Illinois Press.

Glass, G. V. (1978). Standards and criteria. *Journal of Educational Measurement*, *15*, 237–261.

Janssen, R., Tuerlinckx, F., Meulders, M., & De Boeck, P. (2000). A hierarchical IRT model for criterion-referenced measurement. *Journal of Educational and Behavioral Statistics*, *25*, 285–306.

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.

Shrock, S. A., & Coscarelli, W. C. (2007). *Criterion-referenced test development: Technical and legal guidelines for corporate training* (3rd ed.). San Francisco: Pfeiffer.

Smith, E. V. Jr., & Stone, G. E. (2009). *Criterion-referenced testing: Practice analysis to score reporting using Rasch measurement models*. Maple Grove, MN: JAM Press.

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

Submitted March 16, 2016

Accepted April 4, 2016