Standardized Testing in Reading

Troy L. Cox

Brigham Young University, troyc@byu.edu

Follow this and additional works at: https://scholarsarchive.byu.edu/facpub

Part of the Arts and Humanities Commons

Original Publication Citation
Cox, T. (2017) Standardized Testing in Reading. TESOL Encyclopedia for English Language Teaching

BYU ScholarsArchive Citation
Cox, Troy L., "Standardized Testing in Reading" (2017). Faculty Publications. 5885.
https://scholarsarchive.byu.edu/facpub/5885

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.
Standardized Testing in Reading

TROY L. COX

Framing the Issue

While some may think that standardized testing is a recent innovation restricted to the United States, it has actually been part of educational landscape for centuries (Min & Xiwen, 2001; Spolsky, 1995; Barnwell, 1996). The first standardized tests were the proposed solution to combat the problems that existed with appointing civil servants based on cronyism instead of merit. An unknown functionary in the Chinese government hypothesized that a small sample of behavior gathered under carefully controlled circumstances would predict future behavior in uncontrolled situations (Wainer, Bradlow, & Wang, 2007). This observation led to the first, standardized, job-screening tests used to select candidates based on standards that evaluated skills. Getting a job became a product of what-you-knew instead of who-you-knew, and in many ways was a revolutionary idea.

Standardized testing came to forefront in the United States in the early 20th century when military service brought together soldiers from a wide variety of educational backgrounds. Seat-time or self-proclamations of ability were so varied that the need arose to objectively assess what skills individuals had (Spolsky, 2000). Standardized testing soon evolved to other purposes ranging from their use as screening instruments in college admissions to the goal of holding states accountable for educating underserved populations.

The standardized assessment of L2 reading has been problematic from the start, often due to the lack of a clear definition of what reading entails. Franz (1939) surveyed 50 different institutions of higher learning and found that each was assessing a different set of constructs. Some required students to translate a passage from the L2 to the L1 verbatim while others allowed paraphrasing. Some allowed the use of dictionaries, but used subjective descriptors on how dictionary use (e.g. without excessive use) would affect the score. Many had no clear definition on how passage difficulty was determined relying instead on subjective impressions (e.g. of average difficulty or of moderate difficulty). These factors among others made replicating studies difficult.
Instead of relying on in-house test development, many institutions outsourced their assessments to testing companies and institutions that had expertise in assessment. While outsourcing assessments to any outside source promotes uniformity in what is being assessed, the one-size-fits-all approach necessitates careful evaluation to ensure the interpretations of the assessments are used validly. Further care is warranted when comparing standardized exams from different test producers who might define the construct and operationalize it differently. Haphazard acceptance of all standardized reading tests as being equal keeps the field in the state it was in the 1930s.

**Making the Case**

The first thing to remember about standardized tests is the word *standard*. They are designed to be administered under standard conditions. While this observation is tautological, to serve the purpose of being a standard measurement instrument (like a tape measure or a scale), the instrument needs to be the same so that individuals from diverse groups and backgrounds can be fairly compared to one another. Thus, standardized exams need to be administered under the same conditions. For example, if the test is to be timed, then proctors in every location need to enforce time limits. If examinees are to be given a five-minute warning prior to the end of the exam, proctors in all locations need to ensure that happens. If examinees are expected to bring two pencils, then all examinees regardless of location must be well informed of that need prior to the exam date. If examinees have questions on the content, proctors in all locations need to be instructed on how to interact with them. Failure to ensure the conditions are standard can impact the reliability of the tests results and thus impact the validity of the score interpretation.

In the creation of standardized tests, developers approach the task with two main paradigms: norm-referenced and criterion-referenced. Much of the research and tradition in this area has been with norm-referenced testing. This paradigm compares examinees to each other and is intended to order them from the person with the most ability to the one with the least. The *norm* in norm-referenced refers to the statistical normal distribution, and thus emphasizes the examinee’s location within that distribution. To create a test, developers delineate the population to be tested (e.g., literate adults, K-5, ESL Generation 1.5, etc.) and define the construct (e.g., reading to learn, scanning for information, etc.). Test specifications are developed with the intent of dispersing the examinee results so there is a spread of scores. To achieve this end, developers attempt to develop test items that 30% to 70% of the examinee population will be able to answer correctly. The results are then equated with previous test administrations and examinees are given a standard score (e.g., a transformed *z*-score) that shows their position relative to others. Since the scores are spread out, it is easier to make fine distinctions when space or resources are limited. For example, if there were a study abroad program that only had space for 20 participants and there were 100 applicants, a norm-referenced test would allow the administrators to determine who the top 20 were. These results
would also be helpful in determining who the lowest performers were. For example, if there were enough remediation funds to assist three students, the results of a norm-referenced test would help administrators identify those students. The strength of norm-referenced testing is that institutions can set their own cut scores when making decisions based on available resources. The weakness, though, is that it can be more difficult to infer what the students can or cannot do based on external standards. In some instances, even the top examinees may not have mastered skills that would be needed to succeed in a certain job. Conversely, even the worse performing examinees might have sufficient skills to succeed.

While not as pervasive in standardized testing as norm-referenced assessments, criterion-referenced testing is becoming more common. This paradigm compares examinee performance to external standards and is intended to report the examinee’s mastery of those standards. The criterion in criterion-referenced refers to external standards, scales or criteria that have been adopted by the institution and could refer to a curriculum or a hypothesized hierarchy of levels. To create a test, developers delineate the population to be tested to ensure the criteria are appropriate and use the criteria as the basis to create test specifications. Test specifications are defined with the intent to fully represent the criteria to be tested regardless of item difficulty. For example, if the external criteria states that *examinees can read a personal introduction and recognize the person’s name*, that criteria would be in the test specifications and test items would be based on it even if 99.9% of examinees would be able to answer the item correctly. On the other end of the spectrum, a criterion might state that *examinees can infer author tone and mood in 19th century literary works*, that criterion would be in the test specifications and test items would be based on it even if 0.1% of examinees would be able to answer the item correctly. The results are then reported in conjunction to the standard. This allows institutions to evaluate how examinees perform relative to stated standards. For example, school districts might use criterion-referenced exams to know how many students have mastered learning outcomes for grade levels or organizations might use criterion-referenced exams to certify examinees have the abilities needed to succeed in a work environment. The strength of criterion-referenced testing is that the scores directly relate to the standards. Since the scores are related to the standards, diagnostic information can be gathered and it is possible to create profiles of strengths and weaknesses. By extension, it is also possible that either *all* or *none* of the examinees have met the standards being tested. This binary distinction of mastery-nonmastery can make it difficult to use the test results as the basis for selecting the highest or lowest scores for limited resources. The usefulness of the test then is closely tied to the criteria upon which it is based. If the criteria are not based on sound principles or do not align with natural language development, the results are not easily interpretable.

Among the more widely known scales used with criterion-referenced testing are the Common European Framework of Reference (CEFR) and the American Council on the Teaching of Foreign Language (ACTFL) proficiency guidelines. The CEFR was established by the Council of Europe to describe achievement in learning outcomes (Verhelst et al., 2009). Since it was initially designed to be curricular in nature, teachers, schools, and institutions have broad latitude in defining how the
guidelines are operationalized in their individual contexts. Thus, when evaluating exams based on the CEFR guidelines, it is vital to understand how they are operationalized. The ACTFL guidelines were derived from the US government’s Interagency Language Roundtable (ILR) guidelines and are functional descriptors of what people can and cannot do in a language (ACTFL, 2012). While ACTFL has maintained control of the descriptors and the interpretation of them, test publishers may claim their exams are based on the criteria. Once again, it is vital to ask how the guidelines are operationalized.

One important note about reading proficiency guidelines is that some have claimed this skill is impossible to define hierarchically as there is a confounding interaction between background knowledge and second language reading ability. The criticism is that the proficiency descriptors that exist are too circular—that is an advanced level text is one that an advanced level reader can read and one that an intermediate reader cannot (Lee & Musumeci, 1988; Alderson, 2000). Clifford & Cox (2013) found the framework can be hierarchal when there is alignment between the author’s purpose, the text type, and the reader task.

Both norm-referenced and criterion-referenced tests are valuable for their own, different purposes, but it can often be difficult to discern which paradigm a test follows. Students who perform well on norm-referenced tests will also typically perform well on criterion-referenced tests. Conversely, students who perform poorly on one will perform poorly on another. This occurrence results in high correlations between the two test types that may give the impression the tests are equivalent. Test marketers and salesmen will often claim equivalency but, unless both tests were designed to measure the same construct and follow the same paradigm, equivalency claims should be suspect. Careful consumers need to critically ask about the table of test specifications to determine how the test was designed in order to justify how the scores are to be interpreted.

**Pedagogical Implications**

With the pervasiveness of standardized assessments, educators should plan on helping their students succeed with both their classroom learning outcomes and standardized test scores. When examinees are familiar with a test format, their scores improve (Gulek, 2003; Mele-McCarthy, 2007). When they can take practice tests under the same conditions of standardized tests, their scores improve even more (Smith, 1991; Guthrie, 2002). Some examinees are more test-wise than others, so a major implication for teachers is to take some time to level the playing field by teaching their students test-taking strategies. The difficult balance is to ensure the amount of time in preparing for assessments does not unduly affect instruction. This impact can be mitigated if standardized assessments are chosen that closely align with the instructional objectives.

The story is told of the drunkard standing under the street lamp looking for his car keys. When asked why he is only looking under the lamp, his reply is, “that’s where the light is.” Unfortunately many standardized tests use street light item
types such as multiple-choice questions because they are easy to score. Items that can easily be scored *en masse* often do not align with objectives that require deeper levels of understanding. While it is important for students to be familiar with the item types they will encounter, it is even more important for educators to seek out tests that align with their course objectives. With developments in natural language processing, rating responses that are productive will soon be possible perhaps allowing question-types that allow examinees to summarize what they have read.

Most educators are required to use the test their school has chosen or, in the case of university entrance exams, the test the students want to take. In these instances, teachers can examine the test and look for the criteria that do align with the learning objectives, emphasize them, and make them transparent to the students. For the areas in which there is misalignment, teachers should help students understand the big picture. Passing a test to be admitted to a university does little good if the student is unable to complete the course reading assignments and will fail after one semester. Through encouraging self-regulated learning, teachers can help students prioritize what needs to done to be successful on the test and master skills that move beyond the test. To modify a well-known movie trope, “The day may come when standardized tests no longer matter, but that day is not now.” Regardless of how teachers feel about standardized tests—whether they have improved things by awarding people on merit, or they have diminished instructional impact— instructors need to proactively help their students understand how to use the tests to their advantage.

**SEE ALSO:** Levels or Stages of Word Knowledge

**References**


**Suggested Readings**

