Testing and Standards

A Brief Encyclopedia

Sandra Wilde

HEINEMANN
Portsmouth, NH
Contents

Acknowledgments v
Introduction vii

Accountability (Consumers, Taxpayers, and Citizens) 1
Authenticity in Assessment 4
Carrot and Stick (Standards, Testing, and Motivation) 7
Criterion- and Norm-Referenced Tests 8
Curriculum Alignment 11
Emotional Aspects of Testing 13
Equity 14
Essential Readings 17
Ethics 23
Gain Scores and Value Added 26
Grade Level 29
Group Versus Individual Assessment 32
High Stakes 35
How Much Do Tests Cost? 37
Internet Resources 39
Knowing Something Versus Answering Test Questions 42
Lake Wobegon Effect 45
The National Assessment of Educational Progress (NAEP) 47
Passing (Cutoff) Scores 49
Percentiles and Stanines 52
| Politics of Testing                  | 54 |
| Professional Organizations          | 56 |
| Reading Comprehension Assessment    | 59 |
| Reliability                         | 64 |
| SAT Tests                           | 67 |
| Spelling Assessment (as an Exemplar of Broader Issues) | 70 |
| Standard Deviation                  | 73 |
| Standardized                        | 76 |
| Standards                           | 79 |
| Test Security                       | 82 |
| Test Specifications                 | 84 |
| Testing Teachers                    | 86 |
| Validity                            | 87 |
| What Should an Educated Person Know? | 90 |
| Where Do We Go from Here?           | 92 |
| References                          | 97 |
| Appendix                            | 105 |
and cheating born. (See Ethics.) And the pressure created by adding a new, more intense level of sanctions, coming from beyond the classroom, pulls everyone involved even further away from focusing on learning for its own sake.

---

**criterion- and norm-referenced tests**

**What it means:** Criterion-referencing compares one to a standard; norm-referencing compares one to others.

**Example:** Criterion-referenced: How fast you run a mile. Norm-referenced: Did you come in first or tenth in a field of twenty?

**What you need to know about it**

The distinction between criterion- and norm-referenced assessments is, for teachers, one of the most important concepts to understand in the field of testing and measurement. Let’s look at examples of each and then explore some relevant issues.

Scores on the SAT test are an example of norm-referenced assessment. Here’s how it works. A score of 500 means, roughly, that you did better than half the group that the test was normed on and worse than the other half. (The norming group is meant to represent a typical group of test-takers.) A score of 200 means that virtually everyone in that group did better than you. A score of 800 means that you did better than almost everyone else. Your score doesn’t in itself say anything about how many answers you got right or wrong. (Thus a score of 800 isn’t, as it’s often referred to, a perfect score meaning that you got every answer right; it merely indicates, rather, that your score falls in the range achieved by the top 0.13 percent of test-takers in the norming group; see Standard Deviation.)

---

2. This discussion has been somewhat simplified in order to give a broad outline of what a norm-referenced test entails.
A percentile score (see Percentiles and Stanines) gives similar information. A 500 SAT score is at the 50th percentile. You’re at the 50th percentile for height if half the people (in a class, in the country, in your gender) are shorter than you and half are taller. Grading on a curve is also norm-referenced; your grade depends not on how many test items you got right, but on how many you got right compared to other students who took the test. This could be all the other students who took the test that day, or it could be all the other students who have taken the test in the past. A norm always implies, indeed requires, a norming group. (I once tutored a first-grader from a wealthy community who was an object of great concern since he had the lowest standardized reading test performance in his class. When I finally saw his scores, I discovered that he was right on the national mean. Compared to an elite group, he appeared to be in bad shape; compared to a more representative group, he was average.)

The number of pool lengths that you can swim without stopping is a clear-cut example of a criterion-referenced assessment. It is what it is, regardless of how any other swimmer did. The raw score of how many items you got right on a test is also a criterion-referenced measurement. A grading scheme that defines 90 to 100 percent correct as an A, 80 to 89 percent correct as a B, and so on is criterion-referenced.

These swimming and testing examples are, however, different in some important respects. Swimming is a direct measurement of swimming ability, since the performance itself is what we’re interested in, but a test on, for instance, geometry, is only an indirect measure of the underlying knowledge that’s presumably our real goal. For instance, the instructor might have inadvertently made the test too hard or too easy. Thus norm-referenced tests and practices like grading on a curve were designed to help compensate for the effects of inadequacies in the measurement tool.

The results of traditional standardized tests (such as the Stanford Achievement Tests and Iowa Tests of Basic Skills) are typically reported as normed references: percentiles, stanines, and grade-level equivalents.
are all ways of reporting where a test-taker stands in relation to others. However, many states currently, as part of their new standards-based approaches to education, have switched to a criterion-referenced test: This is what we want our students to know, and we’re going to be measuring whether they know it. Student scores are determined not by how they did in comparison to others but by how many test items they got right (which may be stated as a raw score, a percentage, or a scaled score that makes comparisons across tests possible), or how their performance in an area like writing stood up against a rubric.

Although this may seem like a fairer form of assessment, it still has major pitfalls. The first is the difficulty of accurately measuring knowledge, particularly on a single, usually multiple-choice measure. Understanding the main idea of a book you’ve read isn’t the same thing as reading a short passage and then picking out the correct main idea statement from four alternatives. (See Reading Comprehension Assessment.) One could argue about the extent to which they’re closely related, but my point here is that they’re not the same and that a test is at best an approximate measure of whether students have met the criterion we have in mind.

Second, since it’s difficult even to measure whether a particular knowledge criterion has been reached, it’s still more difficult to determine what a passing score should be. The public usually gets upset when what is perceived as too few or too many students pass a state test. If 95 percent pass, the bar must have been set too low; if 50 percent fail, it must have been set too high. The pressures to recalibrate are strong ones.

Ultimately, then, if passing scores end up being determined not by an objective judgment of how one knows whether a criterion has been reached but instead by what proportion of students should, for political reasons, be allowed to have passed, a criterion-referenced measure has metamorphosed into a norm-referenced one. (Susan Harmon [2001] suggests that we should refer to these as politically-referenced tests.) Similarly, as Gerald Bracey (2000a, 107) has
pointed out, a norm-referenced test can be inappropriately used as a criterion-referenced one, as when the Chicago schools assigned a particular score on the *Iowa Test of Basic Skills* as a requirement for promotion to the next grade. Why is this inappropriate? Because rather than providing an accurate assessment that students haven’t learned enough to proceed effectively to the next grade, the schools just assume this to be true of those who score lowest on a single test.

**Curriculum Alignment**

**What it means:** Developing curriculum to fit a test.

**Example:** Spending considerable time practicing solving analogy questions (e.g., Pig is to oink as cow is to ??) because they appear on the state test.

**What you need to know about it**

What relationship should there be between the tests that students take and the curriculum they’re taught? In the case of classroom tests, this relationship is usually a simple one. The students are taught, then they’re given a test based on the material they’ve covered. (This is, of course, only one, relatively narrow, example of the forms that classroom assessment might take.) With system-wide (district or state) tests, it’s more complex.

Aligning a curriculum to match an upcoming test is considered unethical (Haladyna, Nolen, and Haas 1991; see *Ethics*). Perhaps even more centrally, curriculum theory is built around notions of what’s valuable for students to learn, not the principle, “Find out what standardized test your students will be taking and develop your program around it.” (Although this practice does of course take place.) But in today’s standards-based climate, the picture is a little murkier.

In a typical standards-based system, a state department of education will define curriculum goals and then develop tests to match them. Particularly if the tests are high-stakes, teaching then tends
Thank you for sampling this resource.

For more information or to purchase, please visit Heinemann by clicking the link below:


Use of this material is solely for individual, noncommercial use and is for informational purposes only.