PEARSON

assessment report

Fundamentals of Standardized Testing

Sasha Zucker

December 2003

Fundamentals of Standardized Testing

The Purpose of Testing

Tests are a familiar part of classroom instruction. Each week, teachers use a wide variety of tests, such as spelling tests, mathematics pop quizzes, and essay tests. Despite this variety, tests generally share some common goals:

- measuring what students know and can do
- improving instruction
- helping students achieve higher standards

Most of the tests taken by students in classrooms are "teacher-made," designed or selected at the teacher's initiative and tailored to unique classroom circumstances and daily instructional needs. These *instructional tests* provide the teacher with immediate feedback about a student's mastery of a subject area or specific skill. However, because classroom circumstances vary, these teacher-made tests differ considerably from teacher to teacher. A spelling test developed by a teacher for a particular third-grade class will not provide any useful information if it is given to a different third-grade class studying a different set of words. Similarly, a pop-quiz given to one algebra class may be unsuitable for a class covering the same material at a different pace or in an alternate sequence. While undeniably useful at the classroom level, results from instructional tests are unlikely to be comparable across classrooms or schools districts.

The New Era of Testing

Due to the *standards-based accountability* required by the *No Child Left Behind Act of 2001* (NCLB), tests results will be used to hold school districts accountable for raising student academic achievement and identifying schools in need of improvement. NCLB requires each state to establish standards for reading, mathematics, and science. States must develop or select tests aligned to their academic standards and set *performance levels* to identify each student as having reached *basic*, *proficient*, or *advanced* achievement in a subject. Students must demonstrate proficiency in reading and mathematics once per year in grades 3 through 8 and once in high school (grades 10 through 12). Additionally, beginning with the 2007-2008 school year, students must be tested three times in science (once in grades 3 through 5; once in grades 6 through 9; and once in

grades 10 through 12). The results of these tests must be reported widely to the public and used by states to demonstrate whether students are making adequate yearly progress (AYP). State school systems will be held accountable if increasing numbers of students do not obtain proficiency in each subject area (reading, mathematics, and science).

The state is also responsible for demonstrating the achievement of the widest possible range of student populations. Specifically, tests must report the level of student achievement disaggregated by gender, ethnicity, disability, economic disadvantage, English proficiency, and migrant status. The general student population, as well as students in these specified groups, must make progress in achievement each year. The U.S. Department of Education and state educational agencies will use these reports of adequate yearly progress to determine which school systems require assistance or, in the case of consistently underperforming schools, intervention.

Qualities of an Effective Test

The requirements of NCLB pose a significant challenge to state educational systems: *All* students must have the same chance to be successful at showing what they know and can do in periodic, high-stakes assessments. Consequently, states must select or design high-quality tests that can be used by the general student population while meeting the special requirements of certain groups and even the needs of individual students. Moreover, the high stakes involved compel states to be certain that the tests accurately measure student achievement. For a test to solve this combination of challenges effectively, it must be proven to be:

- **Reliable** – The test must produce consistent results.
- **Valid** – The test must be shown to measure what it is intended to measure.
- **Unbiased** – The test should not place students at a disadvantage because of gender, ethnicity, language, or disability.

Standardized Tests

Standardized tests provide a clear solution to the challenges posed by NCLB. A standardized test, such as the *Stanford Achievement Test Series*, *Tenth Edition*, is carefully designed for consistency of format, content, and administration procedure. The reliability of a standardized test is verified by statistical evidence gathered by the test publisher during national studies in which representative groups of students take the test under standardized conditions. By aligning a standardized test with the instructional standards that it is intended to measure, a

test publisher can ascertain one facet of the test's validity. Finally, analysis of the student population enables a test publisher to design a fair test that accounts for the population's diversity and the special needs of individual students. The highest quality standardized tests are produced and used in compliance with accepted guidelines found in the *Standards for Educational and Psychological Testing* (published jointly by the American Research Association, the American Psychological Association, and the National Council on Measurement in Education).

Despite the substantial amount of development effort required, a well-designed standardized test offers a relatively affordable and efficient way of measuring the achievement of a large number of students. When a high-stakes test must be selected to inform decisions that affect the future of a single student or an entire school district, standardized tests that are proven to be reliable, valid, and fair offer the best option for measuring levels of student achievement.

Test Question Formats

While there is no set format for all questions on standardized tests, the most common standardized test question formats include Multiple-choice Questions and Short-answer Questions.

Multiple-choice Questions

One of the most familiar types of test question is the "multiple-choice" format. Also called *selected-response*, this format presents two or more possible answers from which the student chooses. Typically, there is only one correct answer while the other possible answers represent common errors. The ability of multiple-choice questions to efficiently produce highly reliable test results is well established. However, multiple-choice questions are clearly limited in the kinds of achievement that they can measure; they are not suited for determining a student's ability to apply critical thinking skills and carry out complex tasks, such as in performing a scientific experiment. Despite these drawbacks, the selected-response format offers testing agencies a large degree of control in designing reliable, valid, and fair standardized tests.

Short-answer Questions

The short-answer question format, also known as the *open-ended* or *constructed-response* format, presents the test-taker with a question that is answered by a fill-in-the-blank or short written response. Answers to constructed-response questions are hand-scored using a rubric that allows for a range of acceptable and partially correct answers. Questions and answers in this format provide a more sophisticated evaluation of student performance than selected-response questions.

However, the reliability of scores obtained using constructed-response questions depends more heavily on the scoring method. Carefully designed constructed-response questions with a clear scoring rubric can provide important information about student performance and knowledge that cannot be as effectively demonstrated by the selected-response format.

Standardized Testing Frameworks

In addition to designing to account for concerns of reliability, validity, and fairness, test publishers design a standardized test according to how its results will be reported and used. The number of correctly answered questions on a test, the student's *raw score*, only has meaning in the context of the test's *interpretive framework*. Types of interpretive frameworks include Norm-referenced Testing (NRT) Criterion-referenced Testing (CRT), and Standards-based Testing.

Norm-referenced Testing (NRT)

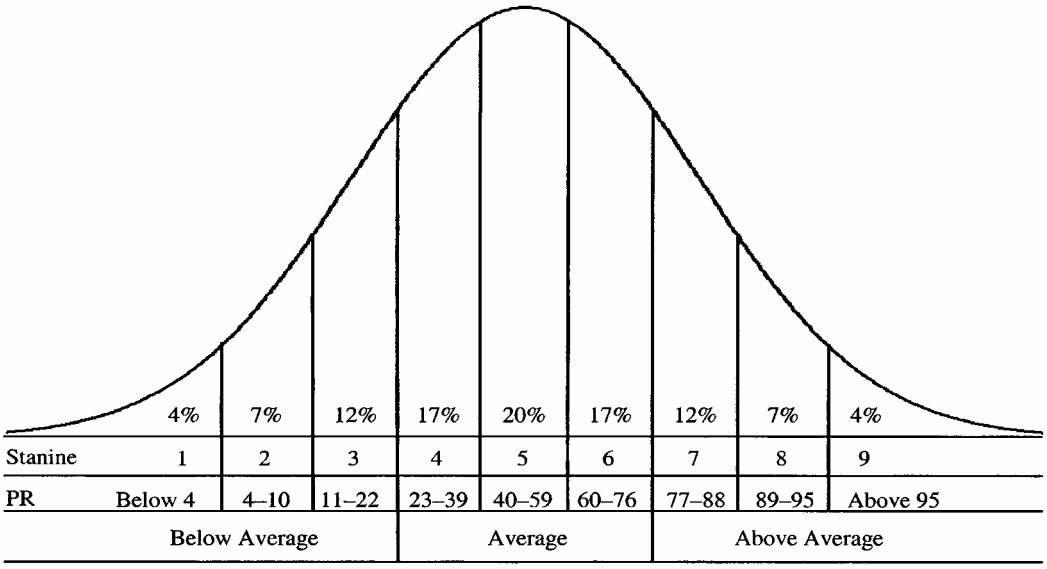
A standardized test designed in the NRT interpretive framework can be used to compare a test-taker's results to the results of a *reference group* that has taken the same test. To norm a test so that results can be compared, a test publisher gathers *normative data* through field trials of the test with a representative, national sample of students. To compare groups as large as entire school systems, norm-referenced tests are typically designed to cover a broad range of what test-takers are expected to know and be able to do within a subject area.

When reporting the results of a norm-referenced test, the test-taker's raw score can be used to make a comparison to the reference group in various ways. Two common methods for making this comparison are to report the test's result as a *percentile rank* or as a *stanine*.

A *percentile rank (PR)* reports the percentage of test-takers whose results are above or below a certain score. For example, a test-taker with a PR of 80 on a test performed better than 80% of the corresponding reference group. The highest possible PR is 99, meaning that the test-taker scored higher than 99% of the reference group, while the lowest PR is 1, and a PR of 50 is the average.

A *stanine* indicates the relative standing of a test-taker's score in comparison to the reference group with a low of one, a high of nine, and five as the average. Stanines 1, 2, and 3 are considered "below average"; stanines 4, 5, and 6 are considered "average"; and stanines 7, 8, and 9 are considered "above average." Each stanine represents an approximately equal unit of achievement. Therefore, the difference between stanines 2 and 4 represents about the same difference in achievement as between stanines 5 and 7. The percentage of scores in the reference group that are classified in each stanine is 4, 7, 12, 17, 20, 17, 12, 7, and

4 respectively. Stanines may correspond to certain ranges of percentile ranks and are typically presented as a curve, such as in Figure 1 below.



	4%	7%	12%	17%	20%	17%	12%	7%	4%
Stanine	1	2	3	4	5	6	7	8	9
PR	Below 4	4–10	11–22	23–39	40–59	60–76	77–88	89–95	Above 95
	Below Average			Average			Above Average		

Figure 1. Stanines and corresponding percentile ranks presented as a curve.

Beyond these two methods, there are many other ways in which the results of well-designed norm-referenced tests can be used to compare what test-takers know and can do. The most appropriate method for reporting the results depends on the kind of comparison required and the manner in which the comparison will be used.

Criterion-referenced Testing (CRT)

Rather than compare a student's test result with the results of a reference group, criterion-referenced tests are intended to measure a level of mastery according to a specific set of performance standards. Hence, the content of a criterion-referenced test often includes more focused subject matter than a norm-referenced test. The test-taker's score corresponds to a *performance level*, such as basic, proficient, or advanced. NCLB requires each state to design or select an assessment yielding results that can be used to classify students into performance levels for the corresponding academic subject.

Standards-based Testing

Standards-based testing allows states to accomplish both objectives (NRT and CRT) at once by incorporating elements of norm-referenced and criterion-referenced testing. A standards-based test is both normed to a reference group and aligned to a set of performance standards. This framework, also called the augmented NRT model, enables states to report standards-based information

(content standards scores), performance levels (cut-scores), and percentile rank information for every student. For example, a test publisher can use a state's academic standards to augment an existing norm-referenced test so that the test-taker's results can be used for both comparisons to a reference group and assigning performance levels.

Typically, statewide results from the first year that a standards-based test is administered are used to establish the test's reference group. Careful design by the test publisher ensures that the test is valid for measuring student mastery of the academic standards. Because NCLB requires states to report student performance levels while also comparing the results of specified student populations to the results of previous years, properly designed standards-based tests are especially suited to meet NCLB requirements.

Conclusion

Standardized tests being developed today provide increasingly useful and sophisticated information with applications potentially beyond NCLB accountability. Because of the precise degree to which standardized tests can be customized to the needs of a state educational agency, teachers will find increased value in using standardized test results to guide instruction. Traditional teacher-made tests will continue to play an important role in the classroom, while standardized tests will play a vital role in measuring students' progress, improving instruction, and helping students achieve higher standards.

Available Related Topics in This Series

Case, B. (2003). *Universal design*. San Antonio, TX: Pearson Inc.

Hicks-Herr, S., & Hoffmann, J. (2003). *Augmentation: An implementation strategy for the* No Child Left Behind Act *of 2001*. San Antonio, TX: Pearson Inc.

Jorgensen, M., & Hoffmann, J. (2003). *History of the* No Child Left Behind Act *of 2001 (NCLB)*. San Antonio, TX: Pearson Inc.

Jorgensen, M., & McBee, M. (2003). *The new NRT model*. San Antonio, TX: Pearson Inc.

Massad, C. E. (2003). *Maximizing equity and access in test construction*. San Antonio, TX: Pearson Inc.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

Association of American Publishers. (2001). *Education assessment: A primer*. Washington, DC: Author.

No Child Left Behind Act of 2001, Pub. L. No. 107–110, 115 Stat. 1425 (2002).

Traxler, A. E. (1961). Fundamentals of testing for parents, school boards, and teachers. *Test Service Notebook, 27*.

U.S. Department of Education. (2003). *No Child Left Behind, accountability and AYP*. Presentation at the Student Achievement and School Accountability Conference, October 2002. Retrieved November 11, 2003, from http://www.ed.gov/admins/lead/account/ayp/edlite-index.html

U.S. Department of Education. (2003). *No child left behind: A parent's guide*. Washington, DC: Author.

U.S. Department of Education. (2003). *Standards and assessments*. Presentation at the Title I Director's Conference, February 2003. Retrieved November 11, 2003, from http://www.ed.gov/admins/lead/account/standassess03/edlite-index.html

Wall, J. E., & Walz, G. R. (Eds.). (2004). *Measuring up: Assessment issues for teachers, counselors, and administrators.* Greensboro, NC: CAPS Press.

The White House. (2001). *No Child Left Behind*. Washington, DC: The White House.

Young, M. (2001). Using *assessments for multiple purposes: Norm-referenced, criterion-referenced, and standards-referenced interpretive frameworks.* Unpublished manuscript.

Additional copies of this and related documents are available from:
Pearson Inc.
19500 Bulverde Rd.
San Antonio, TX 78259
1-800-211-8378
1-877-576-1816 (fax)
http://www.hemweb.com/library/researchreports/index.htm