

BURNED AT THE HIGH STAKES

*A Somewhat Pseudo
Self-Test About Testing*



ILLUSTRATED BY BLAIR THORNEY

BY W. JAMES POPHAM

TODAY, A BLAZE of test-talk is sweeping the country. To illustrate, most newspapers now routinely rank local schools on the basis of students' standardized test scores. Such rankings, whether high or low, invariably trigger test-talk on the part of parents, teachers, and members of school boards. Some newspaper editors, in recognition of the competitive virtues of test-based comparisons, have even considered publish-

*W. James Popham, a UCLA emeritus professor, has written extensively about educational assessment including two recently published books, *Modern Educational Measurement: Practical Guidelines for Educational Leaders* and *Testing! Testing! What Every Parent Should Know About School Tests*, both published in 1999 by Allyn and Bacon.*

ing school-by-school rankings in their sports section. Although people love to applaud a winner, they find it truly gratifying to look down on a loser.

Yet, though test-talk is common these days, not many people speak the language of assessment with unbridled confidence. This is true even for teachers, few of whom actually took a formal course in testing. And, although some might have endured a brush with testing during a course in educational psychology or instructional methods, most have been forced to pick up their insights about assessment as a consequence of on-the-job experience. That's why the following somewhat pseudo self-test is included in this issue of the *American Educator*. You can quickly find out how muddle-free your own test-talk really is. Or you can even try it out on colleagues or unsuspecting friends.

THE SELF-TEST

Directions: Below you will find 17 items containing assessment-related words or phrases presented in boldface type, each of which will be followed by two definitions. Your task is to decide which of the two definitions accurately represents the item's boldfaced word or phrase. At the end of the self-test, an italicized answer key has been provided. The correct answers, incidentally, actually are correct.

1. Affective Assessment

A. A test of someone's *affection*, that is, the degree of amorousness one person feels toward another person. These tests, usually self-report inventories, have recently been employed, with notable economic success, by computer-based singles agencies.

B. Measurement devices intended to assess the attitudes, interests, and values of individuals. In school settings, because of the imprecision of affective assessment instruments, educators should only employ self-report affective inventories to arrive at inferences about the status of a group of students. Self-report affective inventories should not be used to make an inference about the affective status of an individual student.

2. Authentic Assessment

A. A classroom test that has, as a consequence of a teacher's officially notarized affidavit, been formally designated as a genuine classroom test instead of a lesson plan, grocery list, or situation-comedy TV script.

B. Any form of assessment calling for students to supply responses to tasks that are more "real-world" than "academic" in nature. This label for more reality-rooted assessment has fallen out of favor in recent years because its converse, namely, *inauthentic* assessment, seems inane.

3. Cognitive Assessment

A. A measure of a person's intellectual capabilities or potentials. Cognitive *achievement* tests assess a student's knowledge and/or skills. Cognitive *aptitude* tests predict a student's future behavior, for example, in a subsequent academic setting.

B. A totally mental form of testing in which, without any spoken words whatsoever, teachers "think" the questions they wish to ask a class while students "think" their responses to those questions. The Sierra Club has applauded this form of paper-free assessment because of its environmental sensitivity.

4. Criterion-Referenced Test

A. Based on the Greek term *krinein* which means "to separate," a criterion-referenced test is given to married couples who, contemplating a possible dissolution of their marriage, can determine from numerical test scores, complete with decimals, whether a marital breakup appears to be warranted.

B. An assessment of a student's status with respect to defined criterion behaviors such as a body of knowledge or a skill. The student's test performance is *referenced* back to the criterion behaviors when the teacher says, for example, "Your score, Sally, indicates you have mastered 85 percent of 500 Kurdish vocabulary terms represented by yesterday's 20-item vocabulary quiz."



5. Embedded Assessment

- A. A teacher-made test that, either because of security precautions or the test's emotional significance to the teacher who created it, is typically taken to bed by the teacher each night.
- B. A classroom assessment administered during regular instruction so that students regard the test as little more than part of the teacher's routine instructional activities.

6. Grade-Equivalent Score

- A. Expressed in terms of grade levels and months of the school year, one way of describing a student's performance on a standardized achievement test. A grade-equivalent score of 6.4 indicates that the student's test score is approximately equal to how a sixth-grader would typically score in the fourth month of the school year. Because of questionable assessment assumptions, grade-equivalents are used less frequently these days.
- B. A special kind of award given to students by teachers on the basis of students' extraordinary in-class effort. This award, used much like a "Get-Out-of-Jail-Free" card in Monopoly, can be employed at any time by a student as a grade-equivalent for an assignment or test chosen by the student.

7. Norm-Referenced Test

- A. An approach to assessment based on the carefully documented test performances of Norman Nutley, a Midwest student whom university anthropologists were able to observe and assess for 13 years in grades kindergarten through 12 during the 1930s. Students' scores on tests taken today are referenced back to Norm's original performances on a comparable test.
- B. A comparative approach to educational assessment in which a student's performance on a test is compared with the performances of other students who have already taken the same test. The most common norm-referenced assessment occurs when students' scores on nationally normed standardized tests are reported as percentiles, that is, are interpreted relatively according to the performance of the norm group.

8. Normal Curve Equivalent (NCE)

- A. An indication of where a student's test score would have been located on the baseline of a test-score distribution if the test scores had been arrayed, at least approximately, in the form of a normal curve.
- B. A distribution of test scores posing as a normal curve but actually arrayed in a decisively non-normal fashion, for example, skewed or leptokurtic. A normal curve equivalent (NCE) is sometimes referred to as an "ersatz normal curve," hence can also be designated as an ENC.

9. Performance Test

- A. An assessment of students' skills or knowledge requiring responses to tasks designed to elicit fairly substantial products or behaviors. Examples of performance-test tasks would include a directive that a student write an original essay or a requirement for a student to deliver a 10-minute extemporaneous speech.
- B. Tests that are submitted for the annual "performance awards" of the National Council on Measurement in Education. Teachers who submit each year's award-winning classroom tests receive perma-plaques of their test, suitably framed for wall-mounting.

10. Portfolio Assessment

- A. The systematic appraisal of a student's collected work samples. A chief purpose of this form of assessment is to nurture, via portfolio-based conferences between a student and a teacher, students' skills in evaluating their own performances.
- B. In certain charter schools, portfolio-assessed students are allowed to place their worst test performances in a portfolio, and those tests are not used in grading the student. At the end of the school year, all portfolios are burned in a symbolic celebration, often referred to as a school's Assessment-Free Festival.



11. Psychomotor Assessment

- A. A type of psychological measurement calling for students to repair an absolutely unrepairable electric motor. The nature of the student's response to this impossible task allows counselors to classify a student as a neurotic, a psychotic, or a potential educational administrator.
- B. The measurement of a student's ability to perform such small-muscle skills as handwriting, keyboarding, and weaving or such large-muscle skills as pole vaulting, long-distance running, and gymnastics.

12. Quartile

- A. A mathematics performance test in which the student is given one quart of plastic tiles, then must arrange those tiles in pre-specified geometric patterns during exceedingly brief time periods. Often used with children who are being considered for talented and gifted programs, quartile tests have also been employed in recent years to predict whether school-site administrators can arrange class schedules without major duplications or omissions.
- B. A point in a set of test scores that divides the scores into 25-percent segments. The first quartile is equivalent to the 25th percentile; the second quartile is equivalent to the 50th percentile; and the third quartile is equivalent to the 75th percentile. This term is frequently misused by educators who incorrectly regard a quartile as a *quarter*, that is, one-fourth of a set of scores. It would be incorrect to say, "John scored in the first quartile." But it would be correct to say, "John scored in the lowest quarter of the score-distribution."

13. Reliability

- A. A test's inherent dependability, that is, the degree to which the test's results will support teachers if their grade assignments are formally challenged by parents or administrators. Though not widely known, the U.S. Marine Corps motto, *semper fidelis*, was first applied to marine drill sergeants' boot-camp tests thought to be so remarkably reliable that they were regarded as "always faithful."
- B. The consistency with which an educational test measures whatever it is measuring. There are, however, three related but *different* kinds of test consistency, namely, (1) *stability* reliability, that is, the consistency of results between two time-separated testing occurrences, (2) *alternate-form* reliability, that is, consistency of results on two different forms of the same test, and (3) *internal consistency* reliability, that is, the degree to which a test's items are functioning in a consistent manner.

14. Rubric

- A. A scoring guide containing the evaluative criteria by which the quality of students' constructed responses to tests can be judged. *If* properly conceptualized, a rubric can be a potent instructional tool to help teachers and students alike.
- B. A geometric, hand-held cubic puzzle designed to assess a student's spatial-visualization skills. Remarkably popular among laypersons, "Rubric's Cube" has also been used by many teachers for class-management purposes. Because it is a relatively insoluble task, the puzzle has proven useful in keeping hyperactive students occupied.

15. Stanine

- A. A distinctive variety of sea anemone, typically used as a wrong-answer option when biology teachers create multiple-choice exams for their students (often called *urchins*).
- B. A nine-category reporting system, developed by U.S. military measurement specialists, intended to describe examinees' test performances rather generally. In this system, a stanine number nine represents the best student performance while a stanine number one represents the worst student performance. The grossness of this score-reporting approach accurately reflects the inherent imprecision of most educational measurement.



16. Test Bias

A. This occurs when an assessment instrument contains items that offend or unfairly penalize a student on the basis of the student's personal characteristics such as gender, ethnicity, geographic locale, or socioeconomic status.

B. A projective self-report test that presents incomplete fictitious stories to examinees, then uses the responses to those stories to classify any biased examinees into such discrete categories as Male-Chauvinist Bigot, Flaming-Feminist Fanatic, or Klan-Caliber Racist.

17. Validity

A. The extent to which a rationale for the use of a test is fundamentally well founded. If, for example, a principal informs teachers that they must administer a standardized test or lose their jobs, the test is said to possess sufficient validity.

B. The degree of accuracy reflected in the score-based inferences that educators make about students. Although tests are often described as being valid or invalid, validity technically refers to the accuracy of a score-based inference rather than to the test itself.

SELF-TEST ANSWER KEY, PERFORMANCE LEVELS, AND ACTION IMPLICATIONS

Correct Answers

1: B; 2: B; 3: A; 4: B; 5: B; 6: A; 7: B; 8: A; 9: A; 10: A; 11: B; 12: B; 13: B; 14: A; 15: B; 16: A; 17: B.

Performance Levels and Action Implications

Advanced Assessment Literacy = 17 Correct

You should consider taking up after-school moonlighting as a measurement consultant, or you might begin writing assessment-related articles such as this (unless you guessed at most items).

Superior Assessment Literacy = 16-13 Correct

You should identify the terms whose correct definitions you failed to identify, write those definitions on a 3x5-inch index card, and commit such definitions to memory via sub-vocal practice during faculty meetings.

Barely Adequate Assessment Literacy = 12-9 Correct

You should immediately acquire a copy of an enthralling 1999 book, *Classroom Assessment: What Teachers Need to Know*, written by an amiable UCLA emeritus professor and published by Allyn and Bacon, a firm that regularly pays royalties to its authors (www.abacon.com/education). Read it nightly!

Borderline Assessment Illiteracy = 8-5 Correct

You should get a copy of the book cited above, but get a friend to read it aloud to you, very slowly.

Full-Blown Assessment Illiteracy = 4-0 Correct

There is the likelihood that your illiteracy extends well beyond the realm of assessment. Keep these results out of the hands of colleagues and superiors. And for purposes of your own self-esteem, avoid taking self-tests such as this in the future.

