

A SOMEWHAT PSEUDO SELF-TEST ABOUT TESTING

By W. James Popham

Today's teachers are buffeted by a torrent of talk about testing. 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 publishing 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, most teachers do not speak the language of assessment with unbridled confidence. During pre-service teacher education, few teachers actually took a formal course in testing. And, although some teachers might have endured a one-week brush with testing during a course in educational psychology or instructional methods, most of the nation's teachers have been forced to pick up their insights about assessment as a consequence of on-the-job experience. Some of those assessment insights, unfortunately, aren't especially insightful. As a consequence, much of the test-talk currently employed by teachers is mildly muddled.

Accordingly, it's time for teachers to take stock of their own familiarity with testing terminology, and 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 testing talk really is. In the interest of ego-preservation, you may wish to complete the self-test alone—in a locked room.

THE SELF-TEST

Directions: Below you will find 20 items containing assessment-related words or phrases presented in bold-face type, each of which will be followed by three definitions. One of these definitions will be correct; the other two won't. Your task is to decide which of the three definitions accurately represents the item's bold-faced word or phrase. At the end of the self-test, an italicized answer key has been provided, along with suggested courses of action based on your performance. The correct answers, incidentally, actually are correct.

Please consider each item's word or phrase, presented in alphabetical order, then decide whether *A*, *B* or *C* is the correct definition. If you retain a record of your per-item choices, a subsequent visit to the self-test's answer key is apt to be more meaningful. After consulting the answer key, you should first add together the number of correctly answered items. Then, depending on the performance level you attained, consider the course of action suggested. If your total score of correctly answered items was greater than 20, contact the author for a self-test on addition.

*A version of this self-test appeared in the Winter 2000-2001 *American Educator*, Vol. 24, No. 4, pp. 35-39.

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, are typically employed only when students have reached secondary schools and are clearly postpubescent. Such affective inventories have recently been employed, with notable economic success, by computer-based singles agencies.

B A rating form suitable for use in faculty meetings to assess an educator's *affectation* level. These rating-forms are especially helpful when used to evaluate newly appointed principals who sometimes display thinly veiled condescension toward their former teacher-colleague now regarded as the "classroom infantry."

C 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.

C A technique for surreptitiously observing students as they engage in routine, authentic activities. For example, disguised as cafeteria workers during lunch-time, English teachers might covertly eavesdrop to see if students are employing suitable patterns of subject-verb agreements in the heat of authentic disagreements.

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 special test intended to measure a student's spatial-visualization aptitude. The student is presented with items containing pictorial representations of intersecting gears,

then asked to identify which gear's cog is missing. Because some of these missing cogs are so small as to be *nitpicky*, the tests became known as *cognitive* assessments.

C 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 Because kindergarten children, if assessed with standardized tests, often cry while emitting copious tears, the expression to describe this type of premature testing first became known as a “cry-teary-one” test, then was re-spelled in a more technically acceptable manner.

C 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, that “Your score, Sally, indicates you have mastered 85 percent of 500 Kurdish vocabulary terms represented by yesterday's 20-item vocabulary quiz.”

5. *Distractors*

A The mechanisms employed to make a student's examination experience more challenging. Most distractors are patterned after home-court basketball games in which visiting players who are about to shoot a free-throw must be do so in spite of raucous fans who wave balloons or pennants as the free-throw is taken. In classrooms, the most common kinds of distractors used during exams are (1) loud recordings of symphonic music or (2) a teacher's oral reading, with feeling, of selected Shakespearean sonnets.

B The incorrect answer-options that, along with the correct answer, represent the alternatives from which students must choose when responding to a multiple-choice test item. Thus, in a multiple-choice test item with five answer-options, there would be one correct answer and four distractors. In earlier times, these wrong-answer options were sometimes referred to as *foils*.

C A multiple-choice answer option often used chiefly in rural schools, especially in agriculture classes dealing with farm equipment. Whenever “tractor” is the incorrect option, the prefix “dis” negated its correctness, hence created a “distractor” option.

6. ***Embedded Assessment***

A A teacher-made test that, because of either security precautions or the test's emotional significance to the teacher who created it, is typically taken to bed by the teacher each evening.

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.

C A simulation type of assessment found only in high school sex education courses. Students are permitted to take part in these assessments only after receiving their parents' consent.

7. ***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 numerical indicator that is fundamentally equivalent to a grade. Thus, rather than assigning to students a readily comprehensible A, B, C, D, or F, teachers can employ obscure quantitative equivalents, for example, 70, 60, 50, 40, and 30. Whenever possible, the use of decimals to augment numerical grade equivalents is recommended. As a consequence, teachers can inform students and parents that a child earned, for example, a geography grade of precisely 40.72.

C 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.

8. ***None-of-the-above***

A A final answer-option used in multiple-choice tests to render an item more difficult by removing the certainty that one of the item's foregoing answer-options *must* be correct. A "none-of-the-above" option can, in certain instances, strengthen a multiple-choice item. However, an "all-of-the-above" option can be misleading to students, so should never be employed.

B A last-chance option used in multiple-choice tests to give students a "prayer" of answering correctly. The popularity of this answer-option was increased some years ago by Sally Field's television program, *The Flying Nun*. During that era, many students

misinterpreted the answer-option to be “None of the above,” and thus regarded it as some sort of a heaven-directed plea for inspiration. In a number of states, the constitutional separation between church and state has precluded use of this answer-option in classroom tests.

C This is a response made by teachers to an administrative memorandum with which they disagree. The memorandum is returned to the sender with, “None-of-the-above” boldly marked on it, thereby, indicating a succinct but thoroughgoing rejection of the memorandum’s contents.

9. ***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 K through 12 during the 1930s. Students’ scores on tests taken today are referenced back to Norm’s original performances on a comparable test.

B Because early teacher training in the U.S. was carried out almost exclusively in normal schools, this kind of a traditional test (typically of the true-false variety) is referenced back to the assessment approaches advocated by instructors in the nation’s first normal schools.

C A comparative approach to educational assessment in which a student’s performance on a test is contrasted 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.

10. ***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 This phrase refers to parts of a teacher’s anatomy that, either as a consequence of constraining undergarments or serious cosmetic surgery, have been made to resemble a normal physiological curve rather than a body-part that, in reality, is not so normal at all.

C 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.

11. ***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 Only seen in recent years with computer-based assessments, this kind of test is judged not only on psychometric grounds, but also by the quality of the music, graphics, or cartoons accompanying it. Tests that display a truly bravura performance can be nominated for each year's Electronic Socrates Awards.

C 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-plaqued copies of their test, suitably framed for wall-mounting.

12. ***Portfolio Assessment***

A Originated during World War II by the U. S. Navy, a form of shipboard directional testing focused on such locations as port, starboard, and aft. In the early versions of these tests, the quality of *left*-handed sailor's directional skills was formally recorded in a folio, hence a *portfolio*. *Starboard*folio assessment, intended for *right*-handed sailors, was never refined by the Navy's assessment specialists.

B 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.

C 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.

13. ***Psychomotor Assessment***

A A form of performance testing restricted to classes in automechanics wherein the student is obliged to determine if a malfunctioning internal-combustion engine is dominantly experiencing a mechanical or a psycho-emotional problem.

B 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.

C 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.

14. **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 test of preschool children's physical skill, namely, their ability to walk down a curved 10-foot aisle formed by two rows of quart-size milk cartons. Without touching any carton, the children must march down this "quart-aisle."

C 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."

15. **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.

C The degree to which school administrators can rely on teachers to administer standardized achievement tests according to prescribed procedures. Teachers who give their students twice the allowable testing time are, therefore, regarded as relatively unreliable.

16. **Rubric**

A A scoring guide containing the evaluative criteria by which the quality of student's constructed responses to tests can be judged. *If* properly conceptualized, a rubric can be a potent instructional tool to help both teachers and students.

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.

C A code-word used by teachers instead of the more readily understood "scoring guide." Because the descriptor "scoring guide" was far too clear and readily understandable, while "rubric" was truly incomprehensible, "rubric" has been warmly embraced by most measurement specialists.

17. **Scale Score**

A A home economics performance test employed in home-school settings, typically used at the approach of dinner-time. Children are required by a parent to remove the external covering of one or more fish designated for that evening's meal. A parent-determined score is given to the child based on the number of scales that fail to be removed from the evening's entrée.

B A test employed in introductory music classes to determine how rapidly a beginning piano student can play a one-handed musical scale. The student's performance is evaluated according to the speed and accuracy with which a scale is played.

C A way of describing a student's raw score (number correct) on a test according to a new, arbitrarily chosen scale such as one that ranges from 200 to 600. For example, when a student earns a score of 550 on the SAT, this is a scale score. The numerical values of the scales used for these score-reporting systems can vary substantially, for example, from zero to 50 or from 500 to 1,000.

18. **Stanine**

A A test-related motivational ploy used by teachers of nine-year-old students (typically third-graders). The ploy takes the form of the following teacher-issued threat: "If you children don't score high enough on the statewide standardized tests, you'll never grow any older—in other words, you will *stay nine!*"

B 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*).

C 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.

19. ***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.

C This descriptor is used to characterize teachers who are more inclined to test than to teach. Any teacher who spends more actual classroom minutes assessing students instead of instructing them is said to possess a verifiable "test bias."

20. ***Validity***

A A descriptive quality of educational tests drawn from the legal field's traditional conception of *valid* arguments, that is, *strong* courtroom arguments. Because experience suggests that longer courtroom arguments are generally stronger ones, lengthier educational tests are also deemed to be more valid than shorter ones.

B 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.

C 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 the test itself.

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

Correct Answers

1: B; 2: B; 3: A; 4: C; 5: B; 6: B; 7: A; 8: A; 9: C; 10: A; 11: A; 12: B; 13: C; 14: B; 15: B;
16: A; 17: C; 18: C; 19: A; 20: C.

Performance Levels and Action Implications

Advanced Assessment Literacy = 20 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 = 19-16 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 = 15-13 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 = 12-8 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 = 7-0 Correct

There is the likelihood that the nature of your illiteracy extends well beyond the realm of assessment. For purposes of your own self-esteem, you should avoid taking self-tests such as this. Indeed, the number of things you should avoid is almost infinite.

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 this year by Allyn and Bacon.