DOCUMENT RESUME

ED 078 020 TM 002 835

AUTHOR	Popham, W. James; And Others
TITLE	Of Measurement and Mistakes.
PUB DATE	29 Mar 73
NOTE	6p.; Testimony before the General Subcommittee on Education, Committee on Education and Labor, U.S. House of Representatives, Washington, D.C., March 29, 1973
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Comparative Analysis; *Criterion Referenced Tests; Evaluation Techniques; *Measurement Goals; Measurement Techniques; *Norm Referenced Tests; Program Evaluation; Research Methodology; Speeches; Student Evaluation

ABSTRACT

Because of misconceptions regarding appropriate measurement strategies, it is necessary to draw distinctions between two major measurement methodologies, norm-referenced and criterion-referenced measurement, as they relate to determining basic academic capabilities. Norm-referenced measures are used to ascertain an individual's performance in relationship to the performance of other individuals on the same measuring device. Criterion-referenced measures are used to ascertain an individual's status with respect to some criterion, that is, an explicitly described type of learner competence. Because of the wide use of norm-referenced standardized achievement tests, many assume that they are the only instruments that should be used to find out how well a school is working or a pupil is learning. But typical standardized tests are unsuitable for these purposes because of problems with their interpretability and their psychometric properties. Criterion-referenced tests remedy some of these weaknesses because they can: (1) be more accurately interpretable; (2) detect the effects of good instruction; and (3) allow us to make more accurate diagnoses of individual learners' capabilities. If sufficient care is taken to support the development of high quality criterion-referenced measures, legislation to distribute federal funds on the basis of educational deficiencies rather than census determiners appears to be sound. (Author/KM)

OF MEASUREMENT AND MISTAKES*

W. James Popham
University of California, Los Angeles
and
The Instructional Objectives Exchange

You can't measure mileage with a tablespoon. But everyone knows that, so no one tries to. After all, tablespoons were designed to serve a clearly identifiable measurement function, thus they are never employed for assessing such things as distance, sound and heat. Significant problems arise, however, when the mission of a measuring instrument is not so patently obvious, hence it can be mistakenly used in situations whereby it yields apparently respectable but misleading data.

For there are seductive dangers associated with the possession of data. We live in an increasingly evidence-conscious society, and the person who can trot forth a sufficiently impressive array of data often becomes the winner in policy disputes. After all, our data-devotee will claim that he has the facts and the other side operates only on intuition. But, quite obviously, the quality of a data-based argument or decision depends on the quality of the data. Injudicious selection of measuring instruments is likely to yield indefensible data. Unfortunately, in the field of education we are currently suffering from the afflictions of a markedly misapplied measurement tradition.

Not only with respect to the particular bill currently under consideration by this Committee, but because misperceptions regarding appropriate measurement strategies may impinge upon one's appraisal of comparable legislation, it is necessary to draw distinctions between two major measurement methodologies as they relate to determining the basic academic capabilities of the nation's youth. More specifically, differences will be identified between a norm-referenced measurement approach and a criterion-referenced measurement approach. The purposes of these two assessment strategies will be examined along with illustrations of how, if the wrong type of approach is utilized, misleading data will result.

The Basic Distinction

Norm-referenced measures are used to ascertain an individual's performance in relationship to the performance of other individuals on the same measuring device. The meaningfulness of an individual score emerges from the comparison. It is because the individual is compared with some normative group that such measures are described as norm-referenced. Most standardized tests of achievement or

*Invitational testimony before the General Subcommittee on Education, Committee on Education and Labor, U.S. House of Representatives, Washington, D.C., March 29, 1973

intellectual ability used in this country can be classified as norm-referenced measures. Such tests are designed to yield a series of relative performance descriptions, that is, relative to the norm group. It is expected that we will be able to distinguish between Mary who scores at the 65th percentile (of the norm group) and Harry who scores at the 48th percentile (of the norm group).

Criterion-referenced measures are used to ascertain an individual's status with respect to some criterion, that is, an explicitly described type of learner competence. It is because the individual's performance is compared with an established criterion, rather than the performance of other individuals, that these measures are described as criterion-referenced. The meaningfulness of an individual score is not dependent on comparisons with other individuals who took the test. We want to know what an individual can do, not how he stands in comparison to others. For example, the dog owner who wants to keep his dog in the back yard may give the dog a fence-jumping test. The owner wants to find out how high the dog can jump so that the owner can build a fence high enough to keep the dog in the yard. How the dog compares with other dogs is irrelevant. Another example of a criterion-referenced test would be the Red Cross Senior Lifesaving Test, where an individual must display certain swimming skills to pass the examination irrespective of how well others perform on the test. Merely because a group of weak swimmers sign up to take the lifesaving test on a given occasion would not mean that the best performance of that group would necessarily be high enough to pass the test.

Since norm-referenced measures are devised to facilitate comparisons among individuals, it is not surprising that their primary purpose is to make decisions about individuals. Which pupils should be counseled to pursue higher education? Which pupils should be advised to attain vocational skills? These are the kinds of questions one seeks to answer through the use of norm-referenced measures, for many decisions regarding an individual can best be made by knowing more about the "competition," that is, by knowing how other, comparable individuals perform.

Although criterion-referenced tests are also used to make decisions about individuals, there is usually a difference in the context in which such decisions are made. Generally, a norm-referenced measure is employed where a degree of selectivity is required by the situation. For example, when there are only limited openings in a company's executive training program, the company is anxious to identify the best potential trainees. It is critical in such situations, therefore, that the measure permit relative comparisons among individuals. On the other hand, in situations where one is only interested in whether an individual possesses a particular competence, and there are no constraints regarding how many individuals can possess that skill, criterion-referenced measures are preferable. In this sense, criterion-referenced measures may be considered absolute indicators.*

*For a more detailed treatment of the distinctions between norm-referenced and criterion-referenced measurement approaches, see Popham, W.J. (Ed.) Criterion-Referenced Measurement: An Introduction, Educational Technology Publications, Englewood Cliffs, N.J., 1971.

The Misapplied Measurement Tradition

For many years in our nation we have relied heavily on the use of norm-referenced measures. Almost without exception, the many standardized achievements tests used throughout the land fit the classic norm-referenced measurement model. When these devices were used in a fashion consistent with their chief mission, that is, to permit comparisons among individual pupils, then appropriate data were produced. But when these tests were used for other purposes, such as to secure a clear picture of what reading skills a particular child possessed, then the resulting data may have typically been more misleading than helpful.

Yet, because these tests have been widely used for so many years, and because they are produced by reputable commercial publishers (who distribute them with a host of sophisticated measurement trappings such as technical reliability and validity reports), many educators and most citizens assume that standardized achievement tests are the only respectable instruments one should use when attempting to find out how well our schools are working, or more specifically, just how well an individual pupil is learning.

For purposes such as these, the use of a norm-referenced test will often produce spurious data. And the tragedy is that such data may be influential in arriving at far-reaching decisions regarding our nation's educational enterprise. For example, several recent reports have focused on extensive analyses of the relative contribution of numerous factors to the quality of education. The results appear to be disappointing. Teachers don't seem to make much of a difference. Financial expenditures don't seem to make much of a difference. Indeed, schools themselves don't seem to make much of a difference. But much of a difference with respect to what? Invariably the index of pupil achievement used in these large scale analyses has been performance on norm-referenced tests. And, as we shall see, there are characteristics of these measures which render them sufficiently inappropriate for such analyses that the resulting data and subsequent conclusions should be viewed with great suspicion if not complete disdain.

Deficiencies in Norm-Referenced Tests

There are two main problems with typical standardized tests, which render them unsuitable for widescale use in assessing the status of our children's educational attainments. These deficits are associated with the interpretability and the psychometric properties of norm-referenced tests.

Interpretability. Most standardized tests are developed by commercial test publishers who must design the instruments so that they can effectively service an entire nation. Practical economics preclude test publishers from developing a separate test for New York and another version for North Dakota, even though the instructional emphases of these two states may vary considerably. The way that test publishers get out of this bind is to develop a very

general test which, while it may not be perfectly congruent with a given school district's curricular preferences, will at least cover some of them. But to the extent that a particular district is emphasizing content and skills other than those included in the very broad standardized test, a misleading impression of the district's effectiveness or an individual child's capabilities may be created by the use of such tests.

Indeed, it is to the advantage of the commercial test publishers to keep achievement tests at very general levels, for then educators throughout the nation can derive the characteristic Rorschach dividend; they can usually see what they want to in an ink blot. Thus, when certain tests yield subscale scores such as "reading comprehension," it is inordinately difficult to get a precise fix on what is meant by that score. Only by dissecting the test itself can the user secure a defensible idea of what the instrument is measuring. For purposes such as accurately locating our nation's educationally disadvantaged youngsters, we need more crisp interpretaions than are afforded by the bulk of norm-referenced tests.

Just imagine that by employing a standardized achievement test we had located a child who scored below the tenth percentile on a mathematics achievement test. We know, of course, that we have a child who needs help in math. But what kind of help? The typical scores on a standardized math achievement test are often given in phrases as general as "basic operations" or "geometric relationships." With such imprecise descriptors it is next to impossible to really identify what the learner's weaknesses are, much less to correct them.

Psychometric Properties. As we have seen, the chief purpose of norm-referenced tests is to permit comparisons among individuals. Because of this, such tests must produce variant scores. In fact, the more that pupil scores can be spread out, the better. Test items which are answered correctly by most students, since they contribute little to total score variance, must be deleted or modified. To contribute to total score variance an ideal item is one which is answered correctly by half the people taking the test (preferably those who scored highest on the total test) and incorrectly by the other half (preferably those who scored lowest on the total test). Most standardized tests which have been revised several times contain a great many such items since, for purposes of spreading out those taking the test, these items function effectively. But, in general, such test items are most highly correlated with native intellectual ability. In other words, as standardized achievement tests are revised and refined through the years in order to maximize the variability of pupil scores, they more and more closely resemble a classic intelligence test. Thus, norm-referenced tests are often quite insensitive to detecting the effects of even high quality instruction.

To illustrate, suppose a teacher attempts to teach an important concept and, prior to instruction, administers a test item which almost everyone misses. Yet, after a really fine instructional job, the same test item is answered correctly by everyone. But,

because it produces no score variance among students, this kind of item would have to be excluded from a standardized achievement test. This not only leads to insensitive tests but creates the further problem that oft-revised standardized tests many times do not contain the very test items which deal with the central concepts of a field.

Counteractions by Criterion-Referenced Tests

Largely in an effort to remedy some of the weaknesses of norm-referenced measures, criterion-referenced tests are designed in such a way as to (1) be more accurately interpretable, (2) detect the effects of good instruction, and (3) allow us to make more accurate diagnoses of individual learners' capabilities.

Defined Pupil Competencies. One of the important ingredients of a well devised criterion-referenced test is an explicitly defined criterion. Putting it another way, since the whole conception of this measurement strategy is based on referencing scores to a criterion set of learner behaviors, then the behaviors must be described without ambiguity. Most current criterion-referenced measurement specialists are advocating that a domain of learner behaviors be delineated in such a way that from the domain description (often called an item form) an almost unlimited number of test items could be generated. It must be noted that "test item" should be conceived of as representing a wide range of measurement techniques, not merely paper and pencil tests. Because of the characteristic accuracy of the criterion descriptions, we have a far better idea of what it is that the student can or can't do. This becomes particularly important when, upon assessing the students, we discover serious educational deficiencies. With a typical norm-referenced test we would have only a global idea of the general sort of student weakness; with a criterion-referenced test the deficits can be pinpointed and thus more readily ameliorated.

Sensitivity to Instruction. Because criterion-referenced tests need not produce considerable score variance, they can consist even of items which, after instruction, most learners answer correctly. They can retain items which are based on the primary curricular emphasis. As a consequence, such tests are characteristically more sensitive than norm-referenced tests for purposes of detecting instructional effects.

Accurate Diagnoses. Because they are more carefully explicated, criterion-referenced tests typically provide us with a more fine-grained analysis of exactly what the pupil can and can't do. The differential skills we hope learners will acquire can be more accurately portrayed via a well described criterion-referenced test in contrast to its often amorphous norm-referenced counterpart. And for promoting instructional improvement, accurate diagnosis is an indispensable first step.

What About Teaching to the Test?

Discussions such as these often lead to the assertion that precisely explicated tests will encourage instructors to teach to the test, and that such a practice is somehow reprehensible. Contrary to the wide-spread belief that teaching to the test is an instructional sin, we must recognize that *if the test is truly defensible*, then we should applaud those who can teach pupils to master it. The kind of test which will be defensible is not a particular set of items, however, but a *sample* from an almost infinite number of items that could be generated from our well described criterion. In other words, we should not be teaching to a given set of 10 double-digit multiplication problems, but instead to *any* set of 10 double-digit multiplication problems randomly selected from a well defined item pool. Thus the learner acquires mastery of a *class* of skills, not a limited number of items reflected by a particular test. This approach is central to proper use of criterion-referenced testing.

Spending Money and Measuring Skills

The general thrust of the legislation currently under consideration involves the distribution of federal educational funds on the basis of measured educational deficiencies rather than census determiners. Further, there appears to be a recognition of the importance of employing appropriate measurement methodology when identifying eductionally disadvantaged youngsters. Assuming that sufficient care can be taken to support the development of high quality criterion-referenced measures for this purpose, the general scheme for targeting federal dollars appears to be sound. For when we are attempting to identify those young people who truly need educational assistance, then using out-dated census figures as the determiner may be worse than measuring mileage with a tablespoon. It's more like measuring baking soda with a speedometer.