DOCUMENT RESUME

ED 038 670 CG 005 214

AUTHOR Ebel, Robert L.
TITLE Some Limitations of Criterion-Referenced Measurement.
INSTITUTION American Educational Research Association, Washington, D.C.; Michigan State Univ., East Lansing.
PUB DATE 2 Mar 70
NOTE 9p.; Paper presented at American Educational Research Association Convention, Minneapolis, Minnesota, March 2-6, 1970

EDRS PRICE EDRS Price MF-$0.25 HC-$0.55
DESCRIPTORS *Achievement Rating, Achievement Tests, Changing Attitudes, Criteria, *Critical Incidents Method, *Evaluation Methods, Measurement, *Measurement Techniques, Teaching Techniques, *Test Construction, Testing, Tests

ABSTRACT

Contrary to the impression which exists in some quarters, criterion-referenced measurements are not a recent development that modern technology has made possible and that effective education requires. The use of criterion-referenced measurements can not be expected to improve significantly our evaluations of educational achievement. The major limitations of criterion-referenced measurements are: (1) they do not tell us all we need to know about achievement; (2) they are difficult to obtain on any sound basis; and (3) they are necessary for only a small fraction of important educational achievements. It is true that norm-referenced measurements of educational achievement need to have content meaning as well as relative meaning. We need to understand not just that a student excells or is deficient, but what it is that he does well or poorly. However, these meanings and understandings are seldom wholly absent when norm-referenced measures are used. They can be made more obviously present and useful it we choose to do so. (KJ)

Some Limitations of Criterion-Referenced Measurement*

Robert L. Ebel
Michigan State University

Every mental test is intended to indicate how much of some particular characteristic an individual can demonstrate. To determine and express "how much" one needs a quantitative scale. Even those tests used primarily for categorical pass-fail decisions almost always involve a quantitative scale on which a critical "passing" score has been defined. Because the human characteristics that mental tests seek to measure are often complex and hard to define, appropriate quantitative scales are not easy to establish. Some of the most difficult problems of mental measurements arise in the process of getting a useful scale.

The essential difference between norm-referenced and criterion-referenced measurements is in the quantitative scales used to express how much the individual can do. In norm-referenced measurement the scale is usually anchored in the middle on some average level of performance for a particular group of individuals. The units on the scale are usually a function of the distribution of performances above and below the average level. In criterion-referenced measurement the scale is usually anchored at the extremities, a score at the top of the scale indicating complete or perfect mastery of some defined abilities, one at the bottom indicating complete absence of those abilities. The scale units consist of subdivisions of this total scale range.

*Prepared for AERA Symposium, "Criterion-Referenced Measurement: Emerging Issues" Minneapolis, Minnesota, March 1970.

It is interesting to note that the percent grades which were used almost universally in schools and colleges in this country up to about 1920 represent one type of criterion-referenced measurement. True, the extremities of the scales used for percent grades in most courses were very loosely anchored in very poorly defined specifications of what would constitute perfect mastery. But this lack was more a consequence of the great difficulty in developing such definitions than of failure to appreciate their importance. Little has happened to the subject matter of education since 1920 that would make the task of defining complete mastery any easier. If anything, as the scope of our educational content and objectives has broadened, the task has probably become more difficult.

Thus the replacement of norm-referenced measures by criterion-referenced measures in education is not likely to be easy. If it were to happen in the next decade, as some seem to advocate, educational measurement would have come full circle. Those who accept the half-truth that there is nothing new under the sun would have another example to cite. More importantly, the difficulties and limitations of criterion-referenced measures, which half a century ago led to their virtual abandonment, would once again become apparent and would, in all probability start the pendulum to swinging back toward norm-referenced measurements.

This is not to say or to imply that there is no value in criterion-referenced measurements, or no possibility of using them effectively. They have a kind of meaning, a very useful kind, that norm-referenced measurements lack. In some instances good criterion referenced measures can be obtained.[1]

But it is to say that the idea of criterion-referenced measurement is not new, that recent emphasis on norm-referenced measurements has not been misplaced, and that good criterion-referenced measures may be practically unobtainable in many important areas of educational achievement.

Criterion-referenced measures of educational achievement, when valid ones can be obtained, tell us in meaningful terms what a man knows or can do. They do not tell us how good or how poor his level of knowledge or ability may be. Excellence or deficiency are necessarily relative concepts. They can not be defined in absolute terms. The four-minute mile represents excellence in distance running not in terms of any absolute standards for human speed, but because so few are able to run that fast for that long.

Now in many areas of education we do pursue excellence. In many areas we are concerned with deficiency. For these purposes we need norm-referenced measures. To say that such measures leave us in the dark about <u>what</u> the student is good at doing or poor at doing is seldom a reasonable approximation to the true situation. Usually our knowledge of typical test or course content gives us at least a rough idea of amount of knowledge or degree of ability.

One limitation of criterion-referenced measures, then, is that they do not tell us all, or even the most important part, of what we need to know about educational achievement. Another is, as we have already suggested, that good criterion-referenced measures are often difficult to obtain. They require a degree of detail in the specification of objectives or outcomes that is quite unrealistic to expect and impractical to use, except at the most elementary levels of education.

The argument that effective teaching begins with a specification of objectives seems logical enough. If we will settle for statements of general objectives, unencumbered with the details of what is to be taught, how it is to be taught, or what elements of knowledge or ability are to be tested, it is practically useful. But general objectives will not suffice as a basis for criterion-referenced tests. And the formulation of specific objectives which would suffice costs more in time and effort than they are worth in most cases. Further, if they are really used, they are more likely to suppress than to stimulate effective teaching.

The good teacher knows and is able to do thousands of things that he hopes to help his students to know and become able to do. Some of them are recorded in the readings he assigns or in the lecture notes he uses. Others are stored in his memory bank for ready recall when the occasion arises. Why should he labor to translate all these detailed elements of achievement into statements of objectives? If he should do so how could he actually keep such a detailed array of statements in mind while teaching? And if he were to manage such a tour de force, how formal, rigid and dull his teaching would become.

There is obvious logic in the argument that teachers need to think hard about their objectives in teaching. But when the argument is extended to call for specific statements of objectives, written before the teaching begins, it involves assumptions and implications that are open to question. One is that instructional efforts are guided more effectively by explicit statements of objectives than by implicit perceptions of those objectives. Another is that the effectiveness of a teacher's efforts depends more on the explicitness than

or the quality of his objectives, or that explicitness means quality where objectives are concerned. The implication is that programmed teaching which has been carefully planned in detail is likely to be better than more flexible, opportunistic teaching.

Have you ever seen a statement of objectives for educational achievement (not just an outline of learning tasks to be performed) which did justice to all the instructor actually taught in the course and which therefore provided a solid foundation for criterion-referenced measurements of achievement in the course? If you have, did you not find that these objectives substantially duplicated the instructional materials used in the course?

Criterion referenced measurement may be practical in those few areas of achievement which focus on cultivation of a high degree of skill in the exercise of a limited number of abilities. In areas where the emphasis is on knowledge and understanding the effective use of criterion-referenced measurements seem much less likely. For knowledge and understanding consist of a complex fabric which owes its strength and beauty to an infinity of tiny fibers of relationship. Knowledge does not come in discrete chunks that can be defined and identified separately.

Another difficulty in the way of establishing meaningful criteria of achievement is that to be generally meaningful they must not be idiosyncratic. They must not represent the interests, values and standards of just one teacher. This calls for committees, meetings and long struggles to reach at least a verbal concensus, which in some cases serves only to conceal the unresolved disagreements in perceptions, values and standards. These processes involve so much time and trouble that most criterion-referenced type measurements are

idiosyncratic. Is this not what was mainly responsible for the great disagreements Starch and Elliott[2] found in their classic studies of the grading of examination papers? To the extent that criteria of achievement are idiosyncratic they lack validity and useful meaning.

So a second limitation of criterion-referenced measurement is the difficulty of basing such measurement soundly on adequate criteria of achievement. The third and final limitation to be discussed here is less a limitation of the method of measurement itself than of one of the principal justifications that has been offered for its use. This justification argues that when the goal of teaching and learning is mastery, criterion-referenced measurements are essential, since only they are capable of indicating whether or not the mastery has been attained.

Given the assumption of mastery as a goal, this justification is logically unassailable. But should mastery be the goal? At first glance it is most attractive. Partial learning cannot possibly be as good as complete learning. Only a goal that is fully attained can be fully satisfying.

More than forty years ago Professor H. C. Morrison[3] at the University of Chicago developed and popularized a method of teaching based on the mastery of "adaptations" of understanding, appreciation or ability. These, unlike skills, seemed to Professor Morrison not to be matters of degree. "...the pupil has either attained it or he has not." To achieve such an adaptation the instructor should organize his materials into units, each focused on a particular adaptation. He should then follow a systematic teaching routine: teach, test, reteach, retest, to the point of actual mastery.

For a time Morrison's ideas were popular and influential. Around 1930, the Education Index listed 14 articles per year on applications of the system he had advocated. By 1950 the rate had fallen to about 5 articles per year. The Education Index volume for the 1967-68 academic year lists not a single article on this subject.

Recently the concept of mastery has been re-introduced into educational discussions as a corollary of various systems of individually prescribed instruction, and as a solution to the problem of individual differences in learning ability. Several authoritites[4-8] have pointed out, quite correctly that these differences can be expressed either in terms of how much a student can learn in a set time, or in terms of how long it takes him to learn a set amount. Why, they ask, should we not let time be the variable instead of amount learned?

Their arguments have great force when applied to basic intellectual skills that everyone needs to exercise almost flawlessly in order to live effectively in modern society. But these basic skills make up only a small fraction of what the schools teach and of what various people are interested in learning. Look about you at the various talents and interests that different people have developed. See how these differences complement each other in getting done the diverse jobs that need doing in our society. Then ask why we should expect or require a student of a subject to achieve the same level of mastery as every other student of that subject.

Ernest E. Bayles[9] made this point in his criticism of the Morrison method. He made another to which we have already alluded. Abilities, understandings and appreciations are, in the experience of almost everyone, not all-or-none adaptations. They are matters of degree. None but the simplest

of them can ever be mastered completely by anyone. Hence any criterion of mastery is likely to be quite imperfect and arbitrary. To the extent that it is, our criterion-referenced measurements will also be imperfect and arbitrary as were the percent grades that norm-referenced measurements replaced fifty years ago.

To summarize, the major limitations of criterion-referenced measurements are these:

1. They do not tell us all we need to know about achievement.
2. They are difficult to obtain on any sound basis.
3. They are necessary for only a small fraction of important educational achievements.

Contrary to the impression that exists in some quarters, criterion-referenced measurements are not a recent development that modern technology has made possible and that effective education requires. The use of criterion-referenced measurements cannot be expected to improve significantly our evaluations of educational achievement.

It is true of course that norm-referenced measurements of educational achievement need to have content meaning as well as relative meaning. We need to understand not just that a student excells or is deficient, but what it is that he does well or poorly. But these meanings and understandings are seldom wholly absent when norm-referenced measures are used. They can be made more obviously present and useful if we choose to do so.

References

[1]Ebel, Robert L. "Content Standard Test Scores" Educational and Psychological Measurement.

[2]Starch, Daniel and Elliott, E. C. "Reliability of Grading High School Work in English," School Review 20:442-57, 1912.

[3]Morrison, Henry C. The Practice of Teaching in the Secondary School, University of Chicago Press, 1926.

[4]Adkins, Dorothy C., Measurement in relation to the educational process. Educational and Psychological Measurement, 1958, 18, 221-240.

[5]Glaser, R. Instructional technology and the measurement of learning outcomes. American Psychologist, 1963, 18, 519-521.

[6]Carroll, John. A model for school learning. Teachers College Record, 1963, 64, 723-733.

[7]Gagne, R. The conditions of learning. New York: Holt, Rinehart, & Winston, 1965.

[8]Bloom, B. S. Learning for mastery. UCLA, CSEIP Evaluation comment, May 1968, 1 (2).

[9]Bayles, Ernest E. "Limitations of the Morrison Unit" Science Education 18:203-7 December 1934.