ED 093 014 CS 500 749

AUTHOR Buley, Jerry L.
TITLE Criterion Referenced Measurement in Speech-Communication Classrooms: Panacea for Mediocrity. Research Report.
INSTITUTION Arizona State Univ., Tempe. Communication Research Center.
PUB DATE Apr 74
NOTE 16p.; Paper presented at the Annual Meeting of the Central States Speech Communication Association (Milwaukee, Wisconsin, April 4-6, 1974); Some pages may reproduce poorly

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS *Communication (Thought Transfer); *Criterion Referenced Tests; *Educational Research; Higher Education; Measurement Instruments; *Measurement Techniques; Norm Referenced Tests; *Speech Instruction

ABSTRACT

The philosophical underpinnings of the typical testing practices of speech communication teachers in regard to norm-referenced measurement contain several assumptions which teachers may find untenable on closer inspection. Some of the consequences of these assumptions are a waste of human potential, inefficient use of instructional expertise, development of negative attitudes toward school and self, and creation of mental health problems in a significant number of students. Criterion-referenced measurement was developed in response to the weaknesses of norm-referenced measurement, and the assumptions of both types of measurement receive critical attention. (Author/RB)

COMMUNICATION
RESEARCH
CENTER

RESEARCH REPORT

CRITERION REFERENCED MEASUREMENT IN SPEECH-COMMUNICATION CLASSROOMS: PANACEA FOR MEDIOCRITY

Jerry L. Buley

Central States Speech Communication Association
Milwaukee, Wisconsin
April 4 - 6, 1974

Department of Speech and Theatre, Arizona State University, Tempe AZ 85281

CRITERION-REFERENCED MEASUREMENT IN SPEECH COMMUNICATION CLASSROOMS: PANACEA FOR MEDIOCRITY

by Jerry L. Buley

Abstract

The philosophical underpinnings of the typical testing practices (i.e., norm-referenced measurement) of Speech Communication teachers contain several assumptions which the same teachers may find untenable upon closer inspection. Some of the consequences of these assumptions are a waste of human potential, inefficient use of instructional expertise, development of negative attitudes toward school and self, and creation of mental health problems in a significant number of students.

Criterion-referenced measurement was developed in response to the weaknesses of norm-referenced measurement, and used correctly should alleviate many of the negative consequences of the latter.

This paper investigates the assumptions of both types of measurement and discusses their implications for instruction in the speech communication classroom.

CRITERION-REFERENCED MEASUREMENT IN SPEECH COMMUNICATION CLASSROOMS: PANACEA FOR MEDIOCRITY

by Jerry L. Buley
Arizona State University

Introduction

As in everything man does, a compromise must be made between the best job he can possibly do and what is feasible. It is no different in the design and administration of tests. The methodologists scream at us from one side that, because the results of tests are used to make critical decisions, the tests must be designed carefully and that we must be sure they are valid and reliable. On the other side we have all the responsibilities of teaching which eat into our time. The result is that we do the best we can do with what we have.

We can make sure that we are using the best teaching methodology and measurement techniques which are consistent with the time we have available to employ them. The purpose of this paper is (1) to describe the measurement practice, (2) present and criticize the assumptions underlying that practice, (3) to present an alternative measurement practice which would seem to meet these criticisms, and finally (4) discuss methods for evaluating tests in which the alternative measurement practice has been used.

Current Measurement Practice

Usually what that means.is that we spend several hours developing the best test items we can. Then we administer the items to our students. If we have time and the appropriate facilities, we may perform an item analysis to find the relative discrimination and level of difficulty for each item. Discrimination, as you may know, is a measure of how well the question differentiates between those who perform well on the test and those who do not perform well. The difficulty of an item is simply the proportion of students getting the item wrong.

Usually we use results of an item analysis to select items for the same test when it is administered the next time to another group of students.

After the students take the test we assign grades by looking at the distribution of scores on the test. Scores at the top of the distribution become A's, those between the top and the middle become B's, those at the middle become C's, and so on.

What I have described is norm-referenced measurement. That is, each student's score is compared to all scores to determine how well he has performed relative to all other students who took the same test; or, the norm.

Assumptions and Criticisms of Current Measurement Practice

Assumption I: Test should have a broad distribution of scores

Instructors who use norm-referenced measurement are primarily interested in increasing the variability in the scores on their tests. For example, they may throw out questions which do not discriminate between those who perform poorly and those who perform quite well. Also, they will include questions which range in difficulty from quite easy to quite difficult to make sure some people will get a few correct and few will get them all correct. An argument frequently used to defend this practice is that human behavior tends to be normally distributed along a single dimension. Therefore, it is logical to expect that, since test scores are examples of human behavior, they will be normally distributed.

A criticism of this argument is that very few teacher-made tests are unidimensional, nor usually are they intended to be. Usually, a test covers a whole gamut of concepts and ideas even when only over a single unit of instruction. Some students may perform better than others on some tasks but worse than the others on other tasks. The result of this could be to decrease the variability of scores or produce a very skewed distribution. Thus, there is no necessary reason to believe that the distribution of scores will conform to a normal curve.

A second criticism of this argument is that it assumes that scores, even following instruction, must be normally distributed. If we were training students how to talk very sexy, we might tell them they must learn to talk slowly, with low volume, lazy enunciation with a lower pitch, and with a lot of breathiness. We might coach them over a period of time and then put them on an audio tape and grade them. We would expect that nearly all of the students will perform most of the behaviors correctly. In other words, the distribution of scores would be very negatively skewed after instruction. In fact, the better adapted our coaching is to the individual student's weaknesses, the less variability there will be in the results of measurements. Thus, to assume a normal distribution is to assume that instruction has been designed to greatly benefit only a minority of the students.

A second argument used by instructors to defend their emphasis on variability in scores is that current statistical techniques for obtaining estimates of reliability and validity of tests require that the results contain some variability. If variability is low; i.e., there is little difference between high and low scores, we are unable to accurately estimate the reliability of our test.

We can often hear teachers saying that because everyone received a high score on a particular test the test must have been bad, or at the least that it was too easy. If the test cannot discriminate between the good and poor students and if a reliability coefficient cannot be derived, then it must be a bad test.

It is possible that a test may be absolutely reliable and valid in every sense of both terms but receive a spuriously low reliability coefficient because there was little variability in the scores. You and I both would agree that to throw out a perfectly good test because of an artifact of the statistical technique used to evaluate it, would be absurd.

Assumption II: Tests only measure student accomplishment

Many teachers who use norm-referenced measurement believe, or act as though they believe, that their tests are only measures of student accomplishment. For example, if a class of students has an average number of correct responses of 30 on a 60-item test, such an instructor is very likely to say that this was a poor class. As anyone who has had a taste of measurement can tell us, any test score contains at least three components:

A. Student achievement

B. Teacher performance

C. Measurement error.

We have methods for estimating the amount of measurement error in a test (Dick & Hagerty, 1971). We do not have much to help us separate out the component associated with teacher performance and/or the component associated with student achievement. However, it is much more probable that one person's behavior is at deviance than it is that the collective behavior of 30 people is at deviance. Thus, the laws of probability would suggest that, when the average score on a test is far below the possible score, the teacher's performance may have been at fault. Usually, this information is lost in the assignment of grades because the highest score gets an A despite the fact that it represents only 40 correct responses on a 60-item test. Should a 40 get an A one semester on the same test that it takes a 58 to get an A in another semester because the teacher taught better?

Assumption III: Competition produces motivation for higher grades

For hundreds of years teachers have assumed that competition increases motivation for grades. Norm-referenced measurement, because it compares students to other students, is assumed to produce competition between the students for grades. Benjamin Bloom (1971) in an article entitled, "Affective Consequences of School Achievement," has taken norm-referenced measurement to task on this very point.

Bloom (1971, pp. 13-15) suggests that our schools actually have two curricula. One is an explicit curriculum which is the formal content the student is expected to learn. The other is an implicit curriculum which teaches the student who he is in relation to others. While he may learn the latter slowly, its effect is cumulative over a 7- to 12-year period in school. Thus, it is not something he will easily forget.

Throughout his progression through the levels of public school and college, the student is constantly compared to the other students. Nowhere else in his life is he judged so frequently by others and in such precise terms as he is in school. The majority of workers, for example, are expected to meet some minimal standard of work--usually quite low.

Bloom presents evidence (1971, pp. 15-26) which shows that two-thirds of our students acquire a non-positive or even a very negative attitude toward schools, learning, and academics in general (e.g., Russell, 1969; Michael, et al., 1964; and Kahn, 1969).[1]

Bloom also cites evidence which shows that about two-thirds of our students acquire a negative self-concept as a direct result of always having someone performing better than they are performing (Torshen, 1969). Finally, Bloom presents other evidence to the effect that there is a relationship between the use of present measurement practices and mental health. This is probably more true for the bottom third of the classes than for the other two thirds (Glidewell, 1967; Torshen, 1969; and Bowers, 1962).

The result of the use of norm-referenced measurement to produce motivation to learn is that it creates an enormous waste of human potential. The system is geared to produce low self-concepts, negative attitudes toward academia, and even may be linked to mental dysfunctions.

At the same time, the use of norm-referenced measurement creates an inefficient use of instructional expertise. Because some students enter an instructional unit without the knowledge or skills necessary to learn the content of that unit, the teacher must try to bring the weak students up to the required level. This is a waste of time for the students who already have the required skills or knowledge. It also

[1] Although Bloom did not broach the subject, this may be a significant reason why school bond elections have a more difficult time passing every year.

prevents the instructor from presenting the instructional unit in the most efficient manner, resulting in a waste of money and teacher manpower.

In summary, many if not all of the assumptions of norm-referenced measurement may be untenable upon closer inspection. In addition, there are affective, mental health, and economic disadvantages to this current measurement practice.

Criterion-Referenced Measurement: An Alternative

Criterion-referenced measurement is not really a new form of measurement. It bears some resemblence to the use of percentages to evaluate student performance. That is, 90% correct is an A, 80 to 89% is a B, and so on. The point is that the student is no longer compared to other students. Instead, the student is compared to a standard or criterion chosen by the instructor before the instructional unit begins.

The instructor who employs criterion-referenced measurement is primarily interested in testing as a feedback system than as a method for differentiating among students. It is a dichotomous system. Either the student has met the criterion, or he has not. When the measurement procedure is used in conjunction with mastery learning (Block, 1971), the instructor uses the information from testing to locate the areas in which the student is weak and then focuses on them to bring the student up to the criterion level of performance.

It is possible that all students in a class will get all items correct on a criterion-referenced test. In fact, the goal of the instructor who employs criterion-referenced measurement is that the majority (or perhaps even all) of the students achieve a particular criterion level of performance (typically 80%).

As you can see, the procedure places the responsibility for grades on the teacher rather than on the student. When used in conjunction with some form of mastery learning, the student can experience success after success in the learning environment. This should alleviate many, if not all, of the affective consequences of norm-referenced measurement.

If all students meet the criterion in one instructional unit, then they should have the same entry level of skills and/or knowledge for the next instructional unit. The instructor in the next instructional unit would not have to spend time bringing some of the students up to the required level of performance. Therefore, there should be more efficient use of instructional expertise.

Criterion-referenced measurement meets many, if not all, of the criticisms raised against norm-referenced measurement. However, there is a problem with criterion-referenced measurement. This is the problem of obtaining some measure of reliability.

Since criterion-referenced measurement can--and probably will--have very little variability in the scores, it is not possible to use present statistical techniques to derive an estimate of reliability. I have developed a technique which may get around this that I would like to present.

A Method for Evaluating Criterion-Referenced Measurement

Reliability coefficients are based on an estimate of measurement error which is a simple concept in psychometrics.[2] It is the expression of the difference between actual reality and the image of reality produced by our measurement (Dick & Hagerty, 1971, p. 10; Ferguson, 1971, p. 362).

Since it is impossible to know what reality is except by some measure of reality, psychometricians have traditionally taken two measures of the same reality and then compared them. To the extent that the two measures provide the same result, the measure is said to be reliable. The extent to which they do not provide the same result is called the unreliability or measurement error in the measure.

The psychometrician usually does not even obtain two measures of reality and compare them. What he does is to obtain one measure from each student and then find the mean of all these. Since the difference between the mean and the mean of all error in a test is assumed to be zero,[3] the average difference between the students' scores and the mean for the test may be assumed to be the basis for an estimate of measurement error and ultimately reliability.

[2] The following discussion is based in large part on a paper I presented to the Purdue Doctoral Honors Seminar, 1973.

[3] That is, error is randomly distributed and thus would have a mean that is neither above or below the mean of the scores.

This exercise in tangential thinking has heretofore been the primary method for obtaining a measure of reliability of a test. If there is no variability in the test scores, then it is next to worthless. I suggest that we need to go back to the original conceptualization of error in measurement and obtain two measures of reality.

A way to get these two estimates of reality is to ask the student to evaluate the accuracy of his own response and then compare that with the instructor's evaluation of the student's response. The deviation between the instructor's evaluation and the student's evaluation is thus another estimate of error.

Any given response from a student can be evaluated by the instructor as either right or wrong. The student can perceive his response to be either right or wrong. Thus we have the following two by two matrix:

		Instructor's Evaluation of Student's Response	
		RIGHT	WRONG
Student's Evaluation of Own Response	WRONG	1	2
	RIGHT	3	4

When the student and the instructor agree on the correctness of the student's response (Cells 2 and 3), we can say there is no error either in the measurement device or in the instructional unit. However, when the instructor and the student disagree as to the correctness of the response (Cells 1 and 4), we know there is error somewhere. Either the measuring instrument has led the student astray, or the instructor has taught something other than what he thought he had taught.

In order to obtain this data, all that need be done is simply to ask the student to respond twice to each item on a test. The first response is his answer to the question. The second response is his evaluation of his own response.

There are several interesting things which might be done with this data. First the instructor could sum the frequency of occurrence of Cells 1 and 4 across all subjects for each item. This would tell him which items contain or measure the most error.

Also, the instructor could sum the frequency of occurrence of Cells 1 and 4 across all items and all students and divide by the number of students times the number of items. This would produce an estimate of the total amount of measurement error in the test.

Another interesting use might be to sum the frequency of Cells 1 and 4 across all items for each student to find the students who are the most in error (i.e., those who deviate most frequently from the instructor's evaluation of the correctness of a response).

Finally, the instructor might administer the test at the beginning and the end of the instructional unit and look at the changes in the frequencies in the cells of the matrix. One would expect that the frequencies in Cells 1 and 4 would go down (error) and the frequencies in Cells 2 and 3 would go up (non error).

If the instructor finds, for example, that the frequency of Cell 4 increases between the two administrations, he has definite evidence that he was teaching the wrong behavior as being correct.

While the ideas I have presented above seem conceptually sound, I have not conducted any definitive research to verify that they are sound. I do use the technique in all of my own testing, however; and I have found it to be very valuable.

Summary

The assumptions of the traditionally used norm-referenced measurement were found to be untenable after examination. Further, the use of norm-referenced measurement in Speech Communication classrooms may be associated with mediocrity or worse.

Criterion-referenced measurement promises a panacea, especially if it is used in conjunction with mastery learning. The major weakness of criterion-referenced measurement lies in the difficulty of obtaining an estimate of reliability. An innovative method for accomplishing this is to compare student and instructor evaluations of the correctness of the student's response.

References

Block, J. R. (ed.) Mastery Learning: Theory and Practice. New York: Holt, Rinehart and Winston, 1971.

Bloom, B. S. Affective Consequences of School Achievement. In Block, J. R. (ed.) Mastery Learning: Theory and Practice. New York: Holt, Rinehart and Winston, 1971.

Bower, E. M. Mental Health in Education. Review of Educational Research, 1962, 32, 441-454.

Buley, J. L. Measurement Error in Criterion-Referenced Measurement. Paper presented at the Purdue Doctoral Honors Seminar on Trends and Issues in Communication Education, February, 1973.

Dick, W., & Hagerty, N. Topics in Measurement: Reliability and Validity. New York: McGraw-Hill, 1971.

Ferguson, G. A. Statistical Analysis in Psychology and Education. New York: McGraw-Hill, 1971.

Kahn, S. B. Affective Correlates of Academic Achievement. Journal of Educational Psychology, 1969, 60, 216-221.

Michael, W. B., Baker, C., & Jones, R. A. A Note Concerning the Predictive Validities of Selected Cognitive and Non Cognitive Measures for Freshmen Students in a Liberal Arts College. Educational and Psychological Measurements, 1964, 24, 373-375.

Russell, I. L. Motivation for School Achievement: Measurement and Validation. Journal of Educational Research, 1969, 62, 263-266.

Stringer, L. A., & Glidewell, J. C. Early Detection of Emotional Illnesses in School Children. Final Report. St. Louis, Miss.: St. Louis County Health Department, 1967.

Torshen, K. The Relation of Classroom Evaluation to Students' Self-Concepts and Mental Health. Unpublished Ph.D. Dissertation, University of Chicago, 1969.