DOCUMENT RESUME

ED 104 946 TM 004 404

AUTHOR Athey, Irene
TITLE Models for Measuring Growth and the Measurement of Outcomes.
PUB DATE 31 Mar 75
NOTE 9p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, D.C., March 30-April 3, 1975)

EDRS PRICE MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS Criterion Referenced Tests; Elementary Education; Intelligence Tests; Longitudinal Studies; *Measurement; *Models; Norm Referenced Tests; *Program Effectiveness; *Reading; Reading Achievement; Reading Comprehension; *Reading Tests; Standardized Tests; Testing

ABSTRACT

The need for improved measures is particularly acute in reading because, in spite of the magnitude of time and effort which continues to be invested in reading, there is no insurance that the outcome is indeed proportionate to the effort involved. How much of the educational system's performance relative to its own goals is measured by a standardized test is unknown. Yet this is the kind of information which must be available to a school system if it is to make sound decisions on the effectiveness of its programs. The longitudinal evaluation study using criterion-referenced measures of important reading-related skills which is briefly described in the report is seen as offering a new model for tests development which allows for some user involvement in the construction process, as well as contributing significantly to the solution of problems raised in the context of this report. (Author/RC)

Models for Measuring Growth and the Measurement of Outcomes

Irene Athey

University of Rochester

The national concern for greater productivity in education which has characterized the 1970s has focussed attention on the need for developing not only better instructional methods but also better measures of the outcome of such instruction. The need for improved measures is particularly acute in reading because, in spite of the magnitude of time and effort which continues to be invested in reading, there is no assurance that the outcome is indeed proportionate to the effort involved. To the contrary, studies of school productivity frequently result in the policy implication that little can be done to improve program effectiveness (Jensen, 1969).

The alternative implication that it may be the measures of outcome employed are inadequate to the task is rarely drawn. In recent years, however, Jencks (1972) and Bormuth (1970) have seriously questioned the relevance and utility of norm-referenced measures of achievement typically used in studies of school productivity.

An important question with respect to standardized tests is their relevance to the objectives of an educational system, or to the comparison of programs and units within a system. Since the typical measure is referenced to relative comparisons among persons in the standardization sample, it does not refer specifically to the educational system's intended or possible achievement. Nor does a particular student's obtained score on such a test

Paper presented at the annual conference of the American Educational Research Association, Washington, D. C., March31, 1975.

reflect the level of reading he has attained relative to the levels and domains of reading materials in his educational system. How much of the educational system's performance relative to its own goals is measured by a standardized test is unknown. Yet this is the kind of information which must be available to a school system if it is to make sound decisions on the effectiveness of its programs.

The appropriate version of a norm-referenced test administered annually to the same subjects will show whether a particular student has improved his standing relative to the rest of the group. It will not show in absolute terms the amount of growth which has taken place in a single student or in the group as a whole. Still, test data might be used to yield some estimate of absolute growth if the content of the tests tapped the same type of skills at successive age levels. In his analysis of reading tests, however, Singer (1973) points out that they subsume different types of skills, and that the relationship between these types shifts with increasing age. Broadly speaking, in the first four grades techniques of decoding, efficient oculomotor skills, and other mechanical aspects of reading are emphasized. These skills, according to Singer, reach an effective level of mastery by the fourth grade. Thereafter, verbal and reasoning processes , which never reach a ceiling but continue to develop indefinitely, assume a major role in reading performance. Singer sees norm-referenced tests as confounding these two basic types of skills on which the traditional distinction between "learning to read" and "reading to learn" has been based. His solution is to use norm-referenced tests in two ways in the first four grades. The first is the one typically used in schools where the student takes the level of a test appropriate for his grade or age and receives a percentile score. The second is to administer an equivalent form of the first-grade test each year to assess absolute growth in the skills

involved in "learning to read." On the same measure used in these two ways, the same student may be seen over a period of four years to improve substantially toward mastery level on the learning-to-read skills, but to improve little relative to his peers on the ability to read for information.

Singer's solution to the problem of effecti measurement in reading raises a number of interesting issues. In the first place it seems to suggest that all, or at least most , of the important decoding skills are assessed on a first-grade test, although it seems more likely that many of these skills are not introduced until second grade or even later. If this is so, it cannot be true that most children reach complete mastery of these skills by fourth grade, and indeed, research and experience suggests that it is not. A possible solution to this problem might be to administer the second- and third-grade tests repeatedly, in addition to the first-grade test. However, such a solution raises a related question about the content of norm-referenced tests, especially vis-à-vis their relation to intelligence tests. On the decoding skills, Singer (1973b) points out, the correlation with intelligence decreases from first to fourth grade, as more and more students approach mastery. To the extent that the correlation fails to reach zero, we must conclude that either some of the subjects have not attained the level of mastery assumed by Singer to be universal by fourth grade, or that the test is also measuring ~~the presence of~~ verbal and reasoning factors, or possibly both. Under these contingencies these conditions, the use of norm-referenced tests for diagnostic purposes in the first four grades, as suggested by Singer, would seem to have some attendant problems.

After fourth grade, by contrast, the correlation with intelligence continues to be high, especially when the reading task is difficult enough to challenge even the brighter students. Singer concludes that beyond fourth grade reading tests are systematically biased toward the kind of questions that appear on intelligence tests, and that this bias increases with grade level. Perhaps findings such as those of Coleman et al. (1966), which show

the effect of schooling becoming less and less are related to this sytematic bias.

As a result of the high relationship between reading and intelligence tests, a student's previous knowledge and experience tend to be a determining factor in his performance on standardized reading tests. In fact, Simon (1970) has shown that high levels of performance may be attained when the testee receives only the questions, without the passages from which they are derived. Clearly, under these conditions, the test has ceased to be a test of reading comprehension.

The concept of information gain has recently been offered as an alternative which avoids the contamination of pretest knowledge in reading scores (Bormuth, 1971). Information gain is the amount of information an individual gains, as measured by asking questions before and after reading the selection. By adjusting the final score on the basis of pretest knowledge, an estimate of how well the passage has been understood and processed is obtained. Even when this technique is not employed; it becomes important, in the light of Simon's study, to ensure that the passages used to test reading comprehension present information which is essential if the testee is to answer the questions correctly.

Among experts in the reading field who have made comprehension their area for special study, agreement has yet to be reached on whether it consists of a unitary factor, five, or any other number of factors (Davis, 1971). For the purposes of summative evaluation, it may not be important. For diagnostic and instructional purposes, however, analysis of comprehension into its component skills seems essential Further research in cognitive psychology and psycholinguistics may demonstrate the importance of other factors, resulting in modification and refinement of the concept. Meanwhile, our definition must reflect current knowledge and armchair theorizing. From the practical standpoint, it appears that most teachers agree as to which skills are important

and the grade level at which they should be taught. Since norm-referenced tests are not designed to assess these skills in isolation, they cannot be used for the ongoing assessment of the effectiveness of the school's program for teaching the essential skills. Criterion-referenced tests based on behaviorally stated objectives which describe the specific skills in unambiguous terms appear much more suitable for the longitudinal monitoring of those skills which a school system deems to be an integral part of its reading program.

The various considerations which have been outlined above led to plans for the construction and testing of a new assessment model for the use of cooperating school districts in New York State. In the first phase of the study, which has been described extensively elsewhere (O'Reilly, 1973), a Bank of Reading Objectives (BRO), consisting of some 2000 objectives grouped into six areas (multisensory readiness, decoding, vocabulary, comprehension, location and study skills, and reading in the content areas) was compiled by a team of reading research and curriculum experts. To date, efforts have been concentrated in two of these areas, vocabulary and comprehension. From the objectives for vocabulary and comprehension, each of the nine cooperating school districts selected those which were most relevant to its reading program, determined the grade level or levels at which each objective should be taught and tested, and indicated the relative importance of each objective by designating the number of test items to be constructed. The complete test was designed to be administered in a period of 30 minutes. In order to allow for the continuous monitoring which was a major goal of the project, five equivalent forms of each test were constructed, and administered in the pilot phase at intervals of two to three weeks between March and June, 1974. Subjects were randomly assigned to the five forms in such a way that ultimately every student took all five forms of the tests. Initially, the study focussed on grades 4 through 6, but since the range of achievement within these three grades was considerably greater than three years, tests were constructed at

seven levels, corresponding roughly to grades 1 through 7, each student being assigned by his teacher to the level at which he was achieving.

A data-processing system was devised to complement the testing program. The feedback provided to the schools included group data on each item, each objective, and total scores at each level. In addition, every student received information on his own performance on the same measures. The total system, which is known as Comprehensive Achievement Monitoring (CAM)*, has several advantages: (1) It permits diagnostic evaluation of individual and group strengths and weaknesses on specific skills, thus enabling the teacher to deploy instruction time more effectively. (2) It allows for the continuous monitoring of every student on each skill, and is therefore a useful tool in the individualization of instruction. (3) It facilitates flexible ad hoc grouping for the teaching of specific skills on which identified students need further tutoring. (4) As additional data are gathered, correlational studies should indicate empirically which of the skills are indeed most important to the criterion of reading comprehension.

From our experiences with this initial pilot project, several modifications of the system have evolved. (1) In the first place, we found that the distinction between vocabulary and comprehension was not as clear-cut as it may appear on the surface. In fact, it was difficult to assign some of the skills to one or the other category. For this reason, the two were combined into a single test of comprehension. (2) In order to provide for still further input on the part of the cooperating teachers, a pool of items for each objective was constructed, from which they could select in order to construct their own tests, within the limitations imposed by the research design. Construction of these items is still in process, and the final product will be in the form of a Test Development Notebook consist-

*CAM was developed by National Evaluation Systems, Inc. of Boston and Palo Alto.

ing of 800 pages containing 4000 items referenced to the 150 objectives which were deemed most important by consensus of the participating school districts. Since the original seven levels of the pilot phase have also been extended to 20 ordinal levels with two levels per grade for grades 1 through 10, the notebook permits the construction by the school district of at least 100 test forms, distributed in units of five forms at each of the 20 levels. (3) Selection of passages on which the test questions were based followed three criteria: (a) Sampling of passages was made so as to represent the universe of reading materials at each age level, including instructional and leisure reading materials; (b) care was taken to ensure that the questions were unbiased in terms of previous information; (c) passages were calibrated for level of difficulty using a combination of the Dale-Chall readability formula and the Harris-Jacobson word lists. The outcome of these modifications following the field-testing phase is expected to be a valid instrument which is both relevant to school evaluation needs and sensitive to differential growth in the various skills subsumed under the rubric of reading comprehension.

A question of major interest in this project concerns the relative sensitivity of norm-referenced and criterion-referenced tests as measures of the influence of contributing school factors. Preliminary answers to this question are based on a series of multiple regression equations using the norm- and criterion-referenced measures as criteria, and student, teacher, and process variables as predictors. When the newly modified measures are available, it is anticipated that school factors will contribute more strongly to performance on the criterion-referenced measures. A related concern is the extent to which the criterion-referenced tests are perceived by students and teachers as fairer, less threatening, and more informative than conventional formal tests.

The longitudinal evaluation study using criterion-referenced measures of important reading-related skills which has been briefly described here is seen as offering a new model for tests development which allows for some user involvement in the construction process, as well as contributing significantly to the solution of problems raised in the context of this report.

References

Bormuth, J. On the Construction and Use of Achievement Tests.

Coleman, J. S. et al. On the Equality of Educational Opportunity.

Davis, F. B. Psychometric research on comprehension in reading. In F. B. Davis (Ed.), The Literature of Research in Reading with Emphasis on Models. Targeted Research and Development Program, U. S. Office of Education, OEC-0-70-4790, Project No. 0-9030, 1971.

Harris, A. J., and Jacobson, M. D. Basic Elementary Reading Vocabularies. New York: Macmillan, 1972.

Jensen

O'Reilly, R. P.

Simon, H.

Singer, H. Measurement of early reading ability using norm-referenced, standardized tests for differential assessment of programs in learning how to read and in using reading for gaining information. Paper read at the Conference on Early Reading Tests, Georgetown University, August 1973. (a)

Singer, H. IQ is and is not related to reading. Paper read at the annual conference of the IRA, Denver, May 1973.