DOCUMENT RESUME

ED 077 949 TM 002 756

AUTHOR Soar, Robert S.
TITLE Accountability: Problems and Possibilities. Problems in Accountability and the Measurement of Pupils.
PUB DATE 28 Feb 73
NOTE 7p.; Paper presented at the annual meeting of American Educational Research Association (New Orleans, Louisiana, February 28, 1973)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Achievement Tests; *Educational Accountability; *Measurement Instruments; Speeches; *Student Testing; *Teacher Behavior

ABSTRACT

One of the most significant revelations of this paper is a recent finding that indicates that a sizeable number of nonlinear relationships between pre-post gain scores of pupils are found when measured by newer, nonstandardized instruments, which are valid and reliable. When instruments/procedures of this sort are used, the fit for pupils at both the high and low ends of the scale tend to be thrown seriously out of line to the end that they exhibit difficulty in showing any significant gain, while the average students seem to generate normal, expected gains. It has also become fairly evident that a number of pupil characteristics tend to grow at a painfully slow rate so that it becomes almost impossible to realize an appreciable gain in the relatively short space of a year or two. One suggested resolution is to identify a reasonable, manageable number of pupil growth measures that have been found to be related to specific, measurable teacher behaviors and to build accountability programs that are predicated more on teacher behavior than on pupil growth factors. (Author/DB)

Accountability: Problems and Possibilities
Problems in Accountability and the Measurement of Pupils*

Robert S. Soar
University of Florida

Measurement of Pupil Growth

This is an assessment strategy which is immediately appealing to many. Probably there are a number of reasons for this -- since the business of schools is to produce change in pupils, it seems reasonable to assess the success of the school by measuring the growth of pupils; business pays workers, in some instances, in terms of production, why not pay teachers on the same basis? Such a solution is immediate and compelling, but examination of this possibility raises questions.

The Influence of the Classroom

A major difficulty in evaluating the teacher by measuring the growth of the student he teaches is the amount of influence the classroom can have in relation to other influences on the pupil. For example, the relations between attitudes and expectations of parents and both intelligence and achievement of their children have been documented repeatedly and found to be strong. They hold even within a single socio-economic group, and have been demonstrated in a number of ethnic groups (Wolf, 1964; Keeves, 1970; Garber and Ware, 1972, which includes an up-to-date review). Similarly, the influence of the peer group has been demonstrated.

If the teacher is only one of a number of influences on pupil growth, the correlation of growth for one pupil group with another the following year taught

*Presented at the American Educational Research Association meeting, New Orleans, La., February 28, 1973.

by the same teacher should not be high. This turns out to be the case. A series of studies (Soar, 1966; Rosenshine, 1970; Brophy, 1972) suggest that a typical value for this correlation would be in the low..30's. As test-retest reliabilities, correlations like these would not be acceptable.

To lay the pupil's growth, or lack of it, at the teacher's door, seems a major oversimplification, considering the many other factors which may be involved.

Measurement - Statistical Problems

The solution of measuring pupil growth which looks so simple and straightforward is deceptive, and involves a series of problems. The difficulties involved are ones which specialists in educational and psychological measurement have labored over for a generation or more, without final resolution. As Bereiter (1963) comments:

> "Although it is commonplace for research to be stymied by some difficulty in experimental methodology, there are really not many instances in the behavioral sciences of promising questions going unresearched because of deficiencies in statistical methodology. Questions dealing with psychological change may well constitute the most important exceptions. It is only in relation to such questions that the writer has ever heard colleagues admit to having abandoned major research objectives solely because the statistical problems seemed to be insurmountable." (p. 3).

It is relatively obvious that the procedure of measuring only where pupils stand at the end of the year would be inadequate. Whatever growth may have occurred during this school year would be such a minor element in the total amount of pupil knowledge that this possibility is easily dismissed. Nor does criterion referenced measurement avoid the difficulty. It seems obvious that some pupils will be near or above the standard before teaching begins, and others will have little hope of reaching standard within the time available whatever the teacher does. Tracking or grouping will be manifestly unfair

to teachers, and social status differences will be unfair to schools and systems. The alternative which comes to mind easily is that of testing pupils in the fall and again in the spring, so as to determine the change they have made during the time they have been with a given teacher.

This measure however, offers a surprise to those who have their first experience with it -- it too is related to the pretest score, except negatively. That is, students who initially score highest will show least gain during the year, and may very well show a loss on the average, and low scoring students will tend to show the greatest gain. Correlations between pretest scores and raw gain measures for well standardized tests are likely to fall in the -.30 to -.50 range, and less well standardized measures may have higher negative relationships. This is the regression effect, and it can be expected any time the pre and post-measures are less than perfectly related.

The procedures usually used to deal with this problem are to adjust out the relation of pretest by regressed gain or covariance analysis, but these procedures raise a further problem. Since the amount of the admustment made depends on how far the pretest score for a pupil is away from the mean, the question of what group is used to calculate the mean becomes important. Upper class and lower class pupils typically differ in pretest, so probably this difference should be recognized; but how many groups should there be -- how finely differentiated? The answer to this question is critically important, but not at all clear. And the growth a pupil appears to show is affected by it.

Further problems exist, at least occasionally. In our own work, we have not infrequently found that even on well-developed standardized tests it is not unusual for pupils to show ceiling effect. That is, the extent to which a pupil can show growth is limited by the number of items he missed in the fall,

to the extent that a test has this ceiling effect, high scoring pupils will be penalized, since they can't show the real gain they have made. As a further problem in some of the data we are currently analyzing (subtests assembled out of standardized tests), we have found relatively strong nonlinear relationships between pupils' initial scores and the gains they show. Pupils who initially make low scores gain little, pupils who make initially moderate scores gain greatly, and pupils who make initially high scores also gain little. So the classroom which happened to contain pupils who tested toward the middle of the scale will show considerably more gain than a classroom could in which pupils initially scored low or high. If pupils were ability-grouped, the teacher with the middle group would have a material advantage.

The general conclusion from these measurement problems appears to be that the growth a pupil shows is a function both of the growth he actually made, the test items which are used to reflect that growth, and the kind of score used to represent the growth. And since it is difficult to know the relative contribution of each of these sources, the measurement of gain remains uncertain. Also, it is relevant to note that the tests cited above are probably better developed than those to be used in state accountability programs.

Problems of Rate of Growth

Still further problems may exist. It seems reasonable to expect that at least some characteristics of pupils grow sufficiently slowly that change during the school year would not be measureable. (An AACTE task force on performance based teacher education has developed this point). As examples, it seems likely that learning sets towards complex problem solving and such things as responsible citizenship behavior probably change too slowly to be measureable within a single year.

Problems of Teaching and Test Administration

The St. Petersburg Times (Orsini, 1972) has reported on two other problems cited by teachers in the initial application of Florida's accountability program. One is the tendency for some teachers to concentrate on teaching the eight or ten children in the class who were tested in the fall and will be tested again in the spring. Small (1972) documents the parallel problem of teachers concentrating on pupils who were below criterion in an application of accountability measurement in England a century ago. In addition, the problem of teachers concentrating on the material to be tested was also reported in both articles. And there is, of course, always the problem of teachers "helping" pupils as they take the spring test in order to insure that they do well. The alternative of having a disinterested outsider do the testing raises cost-feasibility problems.

Problems of Levels of Complexity

Current evidence (Soar, 1972; Soar and Soar, 1972) suggests that the teacher behavior which supports relatively simple-concrete kinds of pupil growth is different from the kind which supports relatively complex-abstract pupil growth. It would seem important, then, to judge the competence of the teacher on his ability to promote higher level objectives as well as lower level ones.

A troublesome thought is the possibility that only measures of lower level objectives may be developed because of the difficulty of developing the higher ones, yet the accountability program will be forced to go into the field because of legislation which requires it. In that event, the likely result would be accountability testing which would overemphasize lower level objectives and under-represent higher level ones, if they are represented at all. The consequence, then, would be that teachers who stress lower level objectives would do well by the accountability criteria, and teachers who teach in ways which fa-

cilitate the growth of higher level objectives would appear to be less satisfactory teachers. It would not be surprising if this led, in turn, to greater numbers of teachers stressing low level objectives.

Another reasonable expectation is that the teacher who feels the accountability movement looking over his shoulder may very well "turn the screws" a bit, may put pressure on the pupils to achieve, so the teacher will make a satisfactory appearance in the spring testing. This is generally the sort of teacher behavior which is destructive of higher level objectives. So a number of pressures would converge on the teacher to teach for immediate effects, and for low level objectives (and to concentrate on low-achieving pupils, if a terminal measure is the criterion).

In summary, then, the measurement of teacher competence by way of pupil gain appears to be an uncertain route to travel at best. While there are problems in the use of pupil measures for lower level objectives, these problems are perhaps manageable. But the attempts to measure teacher competence through pupil gain in higher level objectives appears to be exceedingly difficult at best, and probably impossible in many cases. The dangers inherent in such an approach seem formidable.

References

Bereiter, C. Some persisting dilemmas in the measurement of change. In Harris, C. W. (Ed.) Problems in measuring change. Madison: University of Wisconsin Press, 1963.

Brophy, J. E. Stability in teacher effectiveness. R & D Report Series 77. The Research and Development Center for Teacher Education, The University of Texas at Austin, July, 1972.

Garber, M., & Ware, W. B. The home environment as a predictor of school achievement. Theory into practice, 1972, 11, 190-195.

Keeves, J. P. The home environment and educational achievement. Australian National University Research School of Social Sciences, Dept. of Sociology, October, 1970, p. 30.

Orsini, B. Perspective, St. Petersburg TIMES, Section D, p. 1, March 19, 1972.

Rosenshine, B. The stability of teacher effects upon student achievement. Review of Educational Research, 1970, 40(5), 647-662.

Small, A. A. Accountability in Victorian England. Phi Delta Kappan, 1972, 53, 438-439.

Soar, R. S. An integrative approach to classroom learning. NIMH project numbers 5-R11 MH 01096 to the Univ. of South Carolina, and 7-R11 MH 02045 to Temple Univ., Philadelphia, Pa., 1966.

Soar, R. S. The classroom: Teacher-pupil interaction. In J. S. Squire (Ed.), A new look at progressive education, Yearbook 1972. Washington, D. C.: Association for Supervision and Curriculum Development, Chap. V., 1972.

Soar, R. S., & Soar, R. M. An empirical analysis of selected follow through programs: An example of a process approach to evaluation. Chap. 11. In Gordon, I. J. (Ed.). Early Childhood Education. Chicago: National Society for the Study of Education, 1972, 229-259.

Wolf, R. M. The identification and measurement of environmental process variables related to intelligence. Unpublished doctoral dissertation, University of Chicago, 1964.