DOCUMENT RESUME

ED 150 187 TM 006 918

AUTHOR Soar, Robert S.; Soar, Ruth M.
TITLE Problems in Using Pupil Outcomes for Teacher Evaluation.
INSTITUTION National Education Association, Washington, D.C.
PUB DATE Apr 75
NOTE 20p.; Also appears as part of TM 006 639

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Academic Ability; *Academic Achievement; Achievement Gains; Educational Accountability; Effective Teaching; Elementary Secondary Education; *Evaluation Criteria; Evaluation Methods; *Performance Factors; Socioeconomic Status; Student Characteristics; Student Evaluation; *Teacher Evaluation; *Testing Problems; *Validity

ABSTRACT

Problems in the use of pupil achievement measures for evaluating teachers, schools or systems are reviewed, with the conclusion that they are disabling. The following reasons are cited: (1) What the pupil brings to the classroom in terms of ability, previous knowledge, home and peer influence, motivation, and other influences is clearly very powerful in determining academic standing at the end of the year. (2) Student achievement reflects only a small portion of the total set of objectives for which schools increasingly are being held accountable. (3) Taking pupil standing at year-end as an indicator of teaching effectiveness frequently does not recognize standing at the beginning of the year; in addition, the problems of adjusting for prior standing are extremely serious and rarely recognized. (4) This accountability system would reward the teacher who teaches to the test or who gives primary attention to those pupils who are below minimum standards but in reach of them. (5) Such an evaluation system would probably reward teaching behavior which promotes low cognitive level learning and penalize teaching which promotes complex learning. (Author/MV)

PROBLEMS IN USING PUPIL OUTCOMES FOR TEACHER EVALUATION

Robert S. Soar
Foundations of Education
Institute for Development of Human Resources
University of Florida

Ruth M. Soar
Florida Educational Research and Development Council

published by

NATIONAL EDUCATION ASSOCIATION
1201 16th Street, N.W., Washington, D.C. 20036

April 1975

FOREWORD

Increasingly state association leaders and staff, local teacher groups, and UniServ directors find themselves having to deal with the problems of inappropriate teacher evaluation. As most teacher leaders are aware, one aspect of such problems is reflected in attempts to evaluate teachers on the basis of student achievement.

The attached paper provides a combination of responses to these problems--research findings, technical problems in using test scores, and other considerations. It has been prepared by two nationally eminent researchers in the field. Citing the Soars can serve to increase the credibility of arguments against the use of student achievement for evaluating teachers. Their examples should be particularly useful in dialogues with school district research directors, testing and evaluation coordinators, and other administrators who are committed to using a year's growth in a year as a measure of teacher competence.

In some other NEA material we have called to attention the major reasons why teachers must not be evaluated on the basis of student achievement. Those reasons complement and are supported by the Soar paper and seem worth repeating here in that context:

1. The tests themselves are inadequate for such purposes. Banesh Hoffman has put it well:

> There is no generally satisfactory method of evaluating human abilities and capabilities.... Rough superficial evaluations are of course possible.... But the detection and evaluation of other than superficial ability is inevitably an art demanding insight, taste and knowledge. Current attempts to reduce it to a science and then mechanize it are not only dangerous but in a profound sense unscientific.[1]

[1] Hoffman, Banesh. "Psychometric Scientism." Phi Delta Kappan 48: 381; April 1971.

2. The nature of student populations is so varied that outcomes are often more influenced by those variables than by what teachers do. Gene Glass, ~~noted researcher~~, reminds us:

> Nothing short of random assignment of pupils to teachers as an iron-clad administrative necessity will ensure that the teachers were in a fair race to produce pupil gains.[2]

3. Many of the conditions which measurably affect learning outcomes are conditions over which teachers have little or no control and they vary widely among schools. Among them are: the number of students teachers must work with each day; time available to teach; planning time; up-to-dateness of curriculum; appropriateness of materials and media; students' physical and emotional readiness for learning; opportunity for teacher in-service education; and most important, decision-making power on curriculum matters.

Each of the reasons cited is considered in one form or another in the Soar paper. And even though some of the technical explanations may go beyond the needs of teacher leaders in responding to the issues, they serve as backups to commonly held teacher association positions.

--Bernard H. McKenna
Professional Associate
NEA Instruction and Professional Development

[2]Glass, Gene V. "Statistical and Measurement Problems in Implementing the Stull Act." Mandated Evaluation of Educators: A Conference on California's Stull Act. (Edited by N. L. Gage.) Washington, D. C.: Education Resources Division, Capital Publications, 1973. p. 54.

PROBLEMS IN USING PUPIL OUTCOMES FOR TEACHER EVALUATION

During the past few years there has been mounting pressure for measuring the outcomes of education, with movement toward holding the teacher, the school, and the school system "accountable" for producing the student learning expected by society. Decreasing enrollments, tighter budgets, and a general trend toward cost effectiveness have added to the pressure.

Measuring pupil achievement has increasingly been proposed as a way of assessing the effectiveness of teaching, and in fact has been mandated by a number of states. This approach is superficially reasonable and attractive, but it is fraught with problems which have not been generally recognized.

H. L. Mencken once commented, "There's always a well-known solution to every human problem--neat, plausible and wrong." The use of pupil achievement as a way of evaluating the teacher, the school, or the school system embodies this misleading simplicity. The solution seems so straightforward: If the job of the teacher is to promote learning in pupils, then it seems reasonable to evaluate the teacher in terms of the amount of learning he produces in his pupils.

The parallel with the industrial setting is clear: If the job of the worker is to assemble relays, then it seems reasonable to count the number

of relays the worker assembles and pay him or her accordingly. But in applying this procedure to teaching, a number of problems emerge which have not been widely recognized. The relay assembler receives parts which are identical (at least within very close limits) on which he or she performs a prescribed set of operations, also identical. Then the completed units leave the assembler, again almost identical from one to another.

But none of this is true for the teachers. Pupils appear in the classroom differing in ability, level of achievement, home background, interest, motivation, age--differing in numerous ways. The teacher must recognize these differences as he or she strives to help individual pupils grow toward their own potential. Consequently, the teaching process will differ from pupil to pupil. If the teacher has been successful, each pupil will have improved educationally when he or she leaves the classroom but each will probably be no more like the others than when the year began.

A major dimension, then, of the problem of evaluating teachers in terms of pupil outcomes is the recognition that what goes on in the classroom is not the only, or the most powerful, influence on where a pupil stands in achievement at the end of the year.

Influences Other Than the Classroom

Research has shown that the differences pupils bring with them when they enter the classroom have significant influence on achievement.

Entry level ability (pretest or fall score) and socioeconomic status are major determiners of what a pupil's standing will be at the end of the school year. These influences probably are more widely accepted than any other, but they are highly interrelated so that one overlaps the other. In practice they cannot be effectively separated.

The fact that IQ and achievement scores in the fall are highly related to spring achievement scores is widely accepted but seldom documented. In a study of 81 fifth-grade classes, Soar and Soar (1973) found correlations between class averages (means) for fall IQ and spring achievement ranging from +.85 to +.90, and correlations between fall achievement and spring achievement, ranging from .75 to .85. So the evidence is that as much as 80 percent of the variation in class averages for pupil achievement at the end of the year can be accounted for by pupil characteristics which existed at the beginning of the year, characteristics over which the teacher has no control.

The most extensive data on the influence of socioeconomic status on pupil achievement were presented in the Coleman Report, and more recently and more widely reanalyzed by Mosteller and Moynihan (1972) and Mayeske, et al. (1972). The studies show that as much as 80 percent of the variation in pupil achievement across schools (equal to a correlation of about +.90) can be accounted for by these factors.

Beyond these major influences there are others which help account for differences in pupil achievement and which should be considered. Although the research on family attitudes and support for learning in the home is not as extensive as that for pupil ability (pretest) and social status, it is consistent in indicating relationships between the educational values held by parents and their children's achievement in school. Garber and Ware (1972) found a relation of +.47 between achievement and a combined measure of support for learning in the home for a group of Black and Spanish American children. All students in the sample met federal poverty guidelines, so that socioeconomic status as usually measured was, in effect, held constant. The same authors cite similar findings from other studies.

Peer group attitude, although again the research is not extensive, has been identified as another important factor which can either support or hinder a pupil's achievement (Anderson, 1970).

Since there is compelling evidence that a number of influences over which the teacher has no control have powerful effects on pupil achievement, it cannot be expected that a teacher will have consistent results with successive groups of pupils. That is, the teacher will not be equally effective in producing growth with all groups because groups differ so widely. Studies by Rosenshine (1970) and Brophy (1972), for example, show that on the average only about 10 to 15 percent of the variation in achievement from group to group reflects the stable influence of the teacher, as shown by a median correlation in the low .30's.

As Medley (1974) has pointed out, and as commonly accepted methods of estimating reliability show (Chronbach, 1960, p. 131), data from about twenty classes would be required for making reliable decisions about individual teachers. Given this requirement necessitating collection of such large amounts of data, using the measurement of pupil achievement as a way to evaluate teachers is impractical as well as invalid.

What these findings seem to indicate is that the education of the pupil is dependent on many conditions in the society, not on the school alone. When the time the pupil spends in the classroom is compared with the time he spends under other influences, and when the degree of influence or control the teacher can exercise is compared with the power of other influences, the limited effect of the teacher is not surprising.

Because influences other than the teacher make a major difference in how much the child learns is not to say that the role of the teacher is

unimportant. The teacher is the only formal, institutionalized input the society has to the education of the child and the transmission of an established curriculum. And much of what the teacher does that contributes constructively to the child's future abilities, successes, and satisfactions may not be measured by currently common achievement instruments. It does say, however, that the teacher's influence is limited and that the teacher is most effective when he or she has the support of other elements in the society.

This whole constellation of other influences is usually not given consideration when measures of pupil achievement are proposed as the basis for evaluating teachers. It is reasonable that these influences are strong, since they accumulate over the life of the pupil. It is obvious, then, that pupil standing at the end of any school year is a completely inadequate and even misleading measure of the effectiveness of the teacher or the school. Yet the results of such achievement standings are frequently published by school or by school system.

"Standing" versus "Change" as Measures of Outcome

"Achievement," which is the most frequently used measure of student learning outcomes, usually refers to the amount of knowledge a pupil possesses at a given point--his or her "standing." The influences cited above show a strong relation to achievement as used in this sense.

An alternative to measuring achievement standing is to measure "change" in achievement from the beginning to the end of the year. When this is done, the influences cited are still likely to have an effect, although to a lesser degree, since change reflects their influence for a shorter period of time.

Although this alternative is appealing as another way of evaluating teaching, it raises still other problems. In a classic volume on the problems,

of measuring change, Bereiter (1963) commented:

> 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 problem seemed to be insurmountable. (p.3)

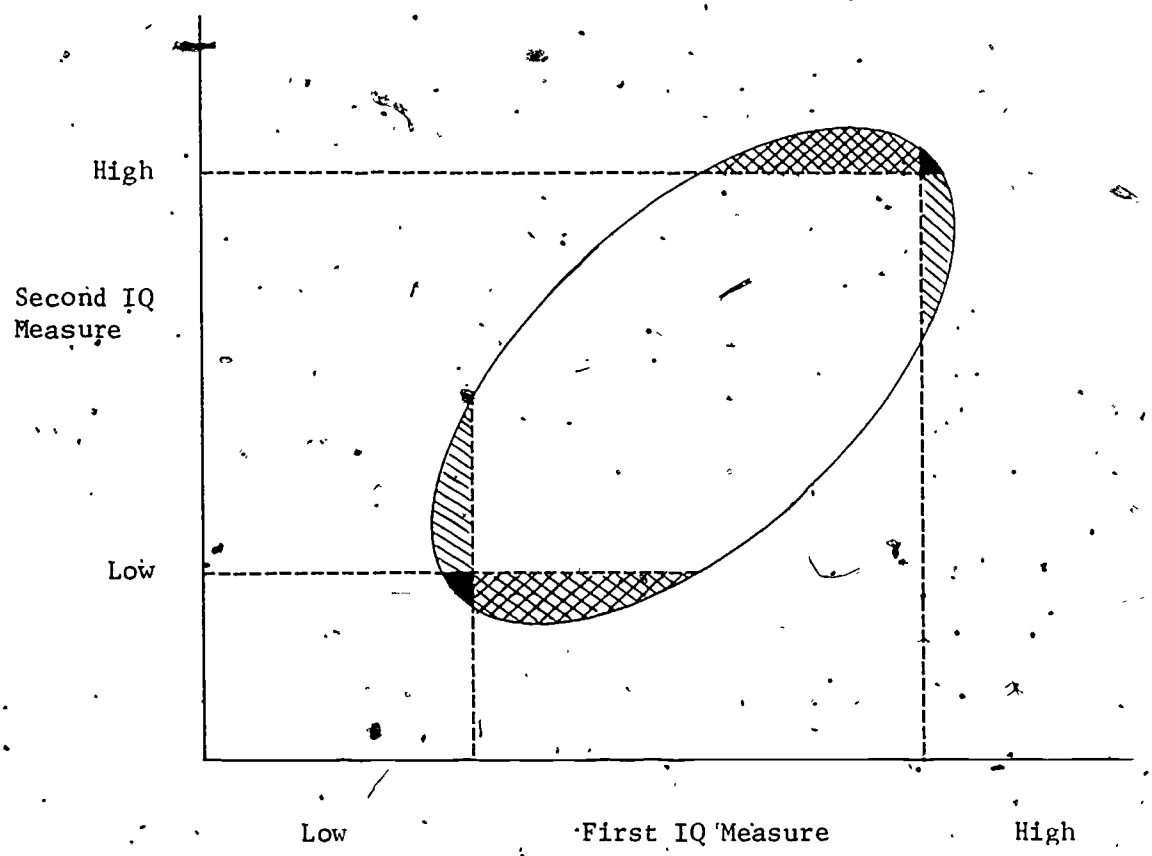
Difficulties in Measuring Change

If the fall score is simply substracted from the spring score so as to obtain a measure of net change, a new set of subtle but difficult problems is created. An illustration may serve to identify some of them. Figure 1 presents fictitious data from a group of pupils for whom measures of IQ from two forms of a test have been obtained 10 days apart. The initial IQ's are plotted on the baseline and the second IQ's on the vertical axis. Any point in the area outlined by the elipse represents simultaneously the IQ of a pupil on each of the testings, and the high and low 10 percent of the pupils at each of the two times has been indicated by shading and cross-hatching.

It is clear that the pupils who were in an extreme group on the first test were not, for the most part, in an extreme group on the second test. The blackened areas represent the small number of pupils who were extreme on both occasions.

At the upper right, the area is small because the pupils who make the highest scores at any testing are likely to do so on two bases: (1) they are bright (have high verbal skills), and (2) they are "lucky" (that is, they

Figure 1.--An Illustration of Regression Effect



happen to make good guesses on a few items for which they aren't sure of the answer, or the items on this test just happen to be ones for which they know the answers). But they are not likely to be lucky consistently when another form of the test is given, and so on another testing their scores are likely to be lower. Opposite influences will affect pupils at the lower left end of the elipse.

To put it another way, if the cutting point for the top 10 percent is an IQ of 120, there will be a number of pupils with true IQ's close to 120 who will sometimes be above that score on a series of tests and sometimes below it, depending on chance factors. So some fraction of pupils above 120 on the first test will fall below it on the second. Similarly, some fraction of the pupils scoring below 80 on a first test will be above it on a second.

In both cases, extreme pupils have "regressed," or moved, toward the mean. This regression effect can be expected whenever prediction is less than perfect, and the extent of the movement will depend on the inaccuracy of the prediction (Lord, 1963). With most psychological or educational predictions, the regression involved is considerable and may make up a significant proportion of the total range of scores.

The point to be stressed from this example has important consequences: Since pupils who were in the bottom 10 percent the first time were not, for the most part, in that group the second time, they must have moved upward. Similarly, the pupils in the top group must have moved downward. That is, there is a negative relationship between initial standing and the direction in which change is most likely.

As an example of this effect, the pupils who stand highest on an achievement measure at the beginning of the school year will probably show little if any increase in score at the end of the year, and may even show a

decline. On the other hand, pupils who score lowest at the beginning of the year will probably show considerable increase. Educators have sometimes been misled by this effect and have assumed that their programs were more functional for low achieving pupils than for high achieving pupils, when in reality all that was involved was the regression effect (the statistical tendency for scores to move toward the average). Similarly, a group of pupils placed in a remedial program because they stand low on a pretest can be expected to show considerable improvement; but again the improvement may be spurious, as a consequence of the regression effect.

This problem creates real difficulties if pupils are tracked on the basis of fall scores and teachers are evaluated on the basis of change in achievement of their pupils. For example, assume that pupils are tested in reading in the fall and the lowest third are put in Miss Jones' class, the middle third in Miss Smith's class, and the highest third in Mrs. Williams' class. We can anticipate that at the end of the year Miss Jones' class will show much improvement and Miss Smith's will show modest gain, but Mrs. Williams will be fortunate if her pupils show any growth at all. The problem is that the gain the pupils show is materially affected by regression effect, so to evaluate the teacher on the basis of pupil gain would be manifestly unfair.

There are statistical procedures for attempting to eliminate this effect, but as Bereiter (1963) commented, it is impossible to be certain that appropriate adjustments have been made; and the expertise to do even the best that can be done with the problem is not widespread. And, of course, all the out-of-school influences on achievement standing discussed earlier also influence gain, although to a lesser degree. So it is clearly inappropriate to use pupil change as a way of evaluating teachers where a teacher may suffer as a consequence of the error involved.

Teacher Performance Tests. A procedure for evaluating teachers which attempts to bypass the problems of change is the performance test or the evaluative teaching unit (Flanders, 1974). In it, the teacher teaches a prescribed brief unit (sometimes as little as a few minutes or as much as two weeks) and pupil knowledge is then tested. The attempt is made to minimize the problems of measuring gain by teaching material in which pupils should have little or no preknowledge, so that all presumably start at "ground zero." But the other problems of using pupil achievement to evaluate teachers still apply. In addition, there are questions of whether teaching material which does not have to be integrated into previous knowledge requires the same skills as the usual teaching setting and whether such short-term learning generalizes to long-term learning. There is the final difficulty that the performance of teachers on a unit of a few minutes does not predict their performance on a two-week unit (McDonald, 1974). Assuming that either can be used to predict year-long performance then seems risky.

Even if the measurement of standing or gain in achievement were a satisfactory way of evaluating teachers, there is still the problem of selecting the objectives to be measured.

What Objectives Should Be Measured?

Although subject matter achievement has been the primary focus of the discussion thus far, it is clear that schools are charged with and have accepted some degree of responsibility for many other kinds of pupil growth.

The Need for Multiple Measures. Over a long period schools have given attention to the social development and the moral values of pupils. And a broad view of the relationship between school and society suggests that when a problem emerges in the society, one of the first steps is likely to be to

involve the school in solving the problem. Traffic problems led to driver education; a concern for the loyalty of government employees led first to a ban on teaching about communism in the schools and later to the requirement that it be taught; problems of drug abuse have led to drug abuse education in the schools; concern about sexual attitudes has led to sex education; concern for occupational choice has led to career education in the schools; and when concern for segregation of the races became pressing for the society, the first and the major attempt to deal with the problem was delegated to the schools. To evaluate teachers and schools solely on the basis of the subject matter gains made by pupils grossly under-represents the broad range of objectives for which teachers and schools have been given some degree of responsibility. Yet for many of these objectives there are no measures which are immediately, for some even remotely, available.

Simple Versus Complex Learning. Even within the subject matter realm there are problems which are largely ignored. One of these problems is the need to distinguish complex achievement growth from simple growth and to provide appropriate measurement for each. Memory of facts (rote memory) falls at the simplest level and complex problem solving, abstracting, and generalizing fall at the complex level. The distinction is between retrieving information (memory) and processing information in its varying degrees of complexity. There is some evidence from a number of studies that the teaching behaviors which are associated with greatest growth in simple tasks are different from those which are associated with greatest growth in complex tasks (Solomon, Bezdek, and Rosenberg, 1963; Soar, 1968; Soar and Soar, 1972, 1973).

Most studies of pupil achievement fail to make this distinction; and the current stress on criterion-referenced measurement, emphasizing "small-

step" learning, seems likely to focus on simple kinds of learning. Measures of complex learning are slow and difficult to construct, in contrast to measures of simple learning, which can be more easily and quickly developed. Evaluating all subject matter at all grade levels would almost certainly require the construction of many new measures which would likely emphasize simple kinds of achievement, given the ease with which they can be constructed and the emphasis on criterion-referenced measurement. If teachers were to be evaluated on the basis of pupil achievement, then, it seems likely that the teacher who emphasizes simple learning would be more positively evaluated than the teacher who emphasizes more complex learning. This would be an unfortunate result.

A further problem related to the difficulty of measuring complex achievement growth is the likelihood that some highly valued objectives grow too slowly to show change within a school year --objectives such as complex problem-solving skills, citizenship, attitudes, learning to get along well with others, and creative expression. On the other hand, it seems likely that measures of short-term learning would tend to emphasize simpler kinds of learning.

Other Problems in the Use of Pupil Outcomes

A description of an application of accountability in England a century ago makes one of the problems clear (Small, 1972). In that setting, teachers were evaluated on the number of their pupils who attained the minimum level of achievement expected for the particular grade. The result was that teachers concentrated their efforts at the minimum level of proficiency, with a consequent lowering of the quality of instruction.

Another problem of serious consequence in the use of pupil measures

is raised by the OEO study of performance contracting, which found that the superior achievement of performance contracting programs disappeared when the teaching was controlled to eliminate the possibility of teaching the test (Page, 1972, 1973). The implication seems clear that, in a setting in which financial return follows from pupil achievement, teaching the test is likely to occur at least a portion of the time. This is a very reasonable finding and one which is well known, even in cases where a financial return is not involved -- teaching to the Regents Examination, for example.

A final problem is the possibility of bias if the teacher is the test administrator. Even outside test administrators have difficulty not "helping" pupils; but where a teacher is affected personally, it seems possible that his or her behavior might be influenced, even though unconsciously. This problem could be dealt with by using only specially trained test administrators, but this could be very costly.

Summary

When all these problems in the use of pupil achievement for teacher evaluation are considered, they become overwhelming. The influence of the teacher is minor compared to out-of-the-classroom influences -- pupil ability, previous knowledge, the home, the peer group, motivation, and others. What the pupil brings to the classroom in this respect is clearly a much stronger determinant of where he or she will stand at the end of the year than anything that has been done in the classroom. Influences on the development of future achievement measures seem likely to limit them to relatively simple measures for some time to come. Tests available for measuring the other objectives for which the teacher is to some degree responsible are relatively few. In addition to these problems, there are statistical difficulties in the measurement of

change which are extremely serious, if not disabling. They are still further exacerbated by the likely problems of teaching the test, of the teacher giving attention primarily to a small portion of the students, and of obtaining valid measurement in the classroom.

Taken all in all, this is an imposing array of difficulties, most of which have gone unrecognized when it is proposed that teachers be evaluated by measuring the outcomes of their pupils.

REFERENCES

Anderson, G. J. "Effects of Classroom Social Climate on Individual Learning." American Educational Research Journal 7: 135-53; March 1970.

Bereiter, C. "Some Persisting Dilemmas in the Measurement of Change." Problems in Measuring Change. (Edited by C. W. Harris.) Madison: University of Wisconsin Press, 1963.

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

Cronbach, L. J. Essentials of Psychological Testing. Second edition. New York: Harper and Brothers, 1960.

Flanders, N. A. "The Changing Base of Performance-Based Teaching." Phi Delta Kappan 55: 312-15; 55: January 1974.

Garber, M., and Ware, W. B. "The Home Environment as a Predictor of School Achievement." Theory Into Practice 11: 190-95; June 1972.

Lord, F. M. "Elementary Models for Measuring Change." Problems in Measuring Change. (Edited by C. W. Harris.) Madison: University of Wisconsin Press, 1963. Chapter 2, pp. 21-38.

McDonald, F. J. "The State of the Art in Performance Assessment of Teaching Competence." Performance Education: Assessment. (Edited by T. E. Andrews.) Albany: Multi-State Consortium on Performance-Based Teacher Education, New York State Education Department, 1974.

Mayeske, G. W., and others, A Study of Our Nation's Schools. U. S. Department of Health, Education, and Welfare, Office of Education, Report No. DHEW-OE-72-142. Washington, D.C.: Government Printing Office, 1972.

Medley, D. M. "Research and Assessment in PBTE." AACTE Leadership Training Conference on Performance-Based Teacher Education. St. Louis, April 30, 1974.

Mosteller, F., and Moynihan, D. P. On Equality of Educational Opportunity. New York: Random House, 1972.

Page, E. B. "A Final Footnote on PC and OEO." Phi Delta Kappan 54: 575; April 1973.

_____. "How We All Failed at Performance Contracting." Phi Delta Kappan, 54: 115-17; October 1972.

Rosenshine, Barak. "The Stability of Teacher Effects Upon Student Achievement." Review of Educational Research 40: 647-62; December 1970.

Small, Alan A. "Accountability in Victorian England." Phi Delta Kappan 53: 438-39; March 1972.

Soar, R. S. and Soar, R. M. Classroom Behavior, Pupil Characteristics, and Pupil Growth for the School Year and for the Summer. Grant numbers 5 ROI MH 15891 and 5 ROI MH 15626, National Institute of Mental Health, U. S. Department of Health, Education, and Welfare. Gainesville: University of Florida, 1973.

_____. "An Empirical Analysis of Selected Follow Through Programs: An Example of a Process Approach to Evaluation." Early Childhood Education. (Edited by I.J. Gordon.) Seventy-first Yearbook, Part II, National Society for the Study of Education. Chicago: University of Chicago Press, 1972. Chapter 11, pp. 229-59.

Soar, R. S. "Optimum Teacher-Pupil Interaction for Pupil Growth." Educational Leadership Research Supplement 2: 275-80; December, 1968.

Solomon, D.; Bezdek, W. E.; and Rosenberg, L. Teaching Styles and Learning. Chicago: Center for the Study of Liberal Education of Adults, 1963.