DOCUMENT RESUME

ED 091 292 SO 007 490

AUTHOR Porter, Andrew C.; McDaniels, Garry L.
TITLE A Reassessment of the Problems in Estimating School Effects.
PUB DATE 1 Mar 74
NOTE 38p.; Meeting of the American Association for the Advancement of Science (140th, Washington, D.C., March 1, 1974)

EDRS PRICE MF-$0.75 HC-$1.85 PLUS POSTAGE
DESCRIPTORS Criterion Referenced Tests; *Educational Accountability; Educational Testing; *Equal Education; Evaluation Criteria; *Evaluation Methods; Measurement; *Measurement Goals; Measurement Techniques; Program Effectiveness; Standardized Tests; *Testing

ABSTRACT

The purpose of this paper is to argue that the measurement of child outcomes is the major stumbling block in estimating the extent to which equality of opportunity exists for children in our nation's schools. This paper has three parts. The introduction discusses the concept of equal educational opportunity, as defined and then redefined over the past five years. The second part identifies the main focus of this paper, that members of the methodological community have ignored basic measurement problems and have directed their efforts towards analysis and design issues and to debates about what to measure, when assessing outcomes of schooling. Consideration of both is necessary to achieve educational reform and accountability. The third part describes the critical problems in measurement that must be resolved before effectiveness of schools can be assessed: when to measure, how to measure, how to interpret the size of an effect, and how to gauge the extent of program implementation when interpreting the results of social experiments about school effects. (Author/JH)

AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE, 140th MEETING

Subject..... A Reassessment of the Problems in Estimating School Effects

Authors..... Andrew C. Porter, Ph.D.
Visiting Scholar
National Institute of Education

Garry L. McDaniels, Ph.D.
Acting Assistant Director
National Institute of Education

Address ..... National Institute of Education
1200 - 19th Street, NW
Washington, D.C. 20208

Time..... 8:30 a.m., March 1, 1974

Place..... Sheraton Palace Hotel
Forty-Niner Room

Program..... AAAS Section of Statistics
Equal Educational Opportunity

RELEASE TIME

At time of Presentation

This paper presents a discussion of the problems in assessing school effects. Our review of current practice leaves us with the distinct impression that measurement issues have received far less attention than have the design and analysis issues. Several important measurement issues which deserve immediate attention are discussed.

The purpose of this paper is to argue that the measurement of child outcomes is the major stumbling block in estimating the extent to which equality of opportunity exists for children in our nation's schools. For the authors, this position represents a shift in emphasis. In our work in the Follow Through and Head Start Planned Variation experiments, the problems related to research design, data analysis, and large-scale research management were seen as paramount (Porter, 1972, 1973, 1974; McDaniels, 1972). In developing the arguments in this paper, we will, therefore, emphasize the reasons for our shifting concern from design, analysis and management problems to the measurement issue.

The paper has three parts. The introduction discusses the concept of equal educational opportunity, the manner in which this concept has redefined over the past five years, and the consequences of this redefinition for teachers and school administrators. The second part of the paper identifies the issues which were of initial concern to the authors and notes why our concerns have shifted. The third part describes critical problems which we now feel must be resolved before the effectiveness of schools can be assessed.

I. INTRODUCTION

A definition of equal educational opportunity has been troublesome, and it is fair to say that at present there is no concensus as to what is meant by the phrase. In the early 60's, the definition emphasized an equitable distribution of traditionally valued and rather easily measured educational resources (e.g., promoting small class size, hiring experienced teachers, expending equal dollars per student, encouraging schools to have libraries of similar size, etc.). The accountability of schools for providing equal educational opportunity was based on measuring quantities of these resource inputs, and the distribution of Title I funds was used to decrease their inequality across schools.

This general definition has fallen out of favor. First, the distribution of general resources was not nearly as uneven across subpopulations of children as had been believed (Coleman et al, 1966). Second, these general resources did not appear to be as strongly related to child outcomes as had been expected (Jencks et al, 1972; Mosteller, and Moynihan, 1972; and Coleman et al, 1966).

Consequences of New Definitions

There is wide-spread agreement that some subpopulations of children are not receiving an adequate education in the public schools. Subpopulations of children leave schools with few accomplishments in reading, mathematics, etc. Since these outcomes of schooling are uneven across subpopulations of children, unequal educational opportunity must exist.[1]

This general agreement is based on the assumption that we have adequate documentation of the outcomes of schooling that the test results which provide the basis for estimating the effects of schooling are reasonable. The second assumption is that the outcomes documented are, in fact, the result primarily of a school input (or the result of the lack of a school input).

The third assumption is that teachers and administrators can do something about school effectiveness and should, therefore, be held accountable for inequalities in these measured outcomes.

---

[1]It is likely that unequal or inequivalent educational programming and teaching resources should be employed among schols to achieve less variance among child performances.

The consequences of holding these assumptions now appear in State and Federal legislative initiatives. For example, 31 States are now considering laws to require all applicants for a teaching license to demonstrate their abilities as teachers. One of the leaders in this recent movement is California which has put into effect the Stull Act of September 1, 1972. Briefly, the Stull Act requires all California school districts to evaluate their teachers, administrators and other kinds of professional staff. Much of the evaluation is likely to be based on student behavior (outcomes). In addition there have been several States which have started major state-wide assessment programs based on student behavior. In the State of Michigan, there is a Chapter III Program which allots Title I funds on the basis of child achievement. At the Federal level, the Quie Bill, HR 5163, (in Committee) proposes to do the same thing with Title I funds for the nation.

There is a quality of unreality to these efforts which assume a known link between child outcomes and teacher behaviors. Several recent reviews of the research on school effects (Rosenshine and Furst, 1972; Rosenshine, 1971) have shown

that little work has been done. There are studies that show that teachers can be trained to behave in specific ways and that the teachers continue to act as trained for several years. However, there are few studies which link trained teacher behavior to child outcomes. And in the few studies which do exist, the methodological problems, especially in the definition of the child outcome variables (Heath and Nielson, 1973) are serious. Therefore, we now find teachers and school administrators in the bizarre situation of being told that they are accountable for equalizing educational outputs; outputs which are poorly measured (if measured at all). And in order to achieve these outputs, teachers are asked to be prepared in highly specific ways; ways which have not been related to the outputs expected. Similarly, school administrators are being asked to provide programming and materials to support teachers to produce these outcomes. This is becoming law.

II. EVALUATION OF SCHOOL EFFECTIVENESS

It is apparent that important policy decisions about education, and thus equal educational opportunity have been made with the assumption that research on school effects is either available or coming. Since available research is limited, it is therefore imperative that the research be conducted with reasonable dispatch. However, even the limited research on school effectiveness has uncovered a number of problems.

Three general categories of problems have occupied investigators (including ourselves) in this area: (1) the analysis of data in natural experiments; (2) the design and management of true experiments in field settings; (3) the apparently endless array of variables requiring measurement in social experiments. Each of these problem areas will be described briefly in the pages that follow.

Analysis of Data in Natural Experiment

Most field studies have been natural experiments (correlational studies) and do not provide direct information about aspects of schooling which cause changes in student behavior. Analytic arguments have centered around the utility of various analysis models for teasing out causal relationship from data

collected in natural (not experimentally manipulated) settings. Probably the two most common problems in these efforts have been the total or near total confounding of various in-school and out-of-school variables, and the sense that there was little natural variation on some of the in-school variables that might have their greatest potential impact on child behavior. Debate about the utility of various statistical models has centered around such issues as:

1. The ability of multiple-regression models to control for the partially confounded out-of-school variables such as family background;
2. the problem of multiple-fallible covariables in analysis of covariance, and the use of within treatment groups regression to predict the effects of between treatment group confounding;
3. the relative utility of matching versus regression adjustments for confounding variables;

4. the choice of units of analysis;

5. the utility of multiple data points, both prior to and after the introduction of the phenomenon under investigation.

Design and Management of True Experiments in Field Settings

It has been pointed out by numerous critics of correlational studies that none of these analytic solutions can rescue correlational studies so that causal statements can be made. As Box has said "To find out what happens to a system when you interfere with it, you have to interfere with it (not just passively observe it)." We agree with the notion that social experiments are needed; that potentially they provide more information; and that they are less likely to produce misleading information for policy makers and other observers.

Further, we are reasonably optimistic about the possibility of conducting such studies. There are three reasons for this optimism. First, a climate of opinion favoring the use of experimental design for studying social policy has been established. Articles supporting experimental research appear from academicians (Campbell, 1971), policy-oriented organizations (Brookings, 1974) and government officials at the State and national levels

(Evans, 1969; Goldberg, 1974; McDaniels, 1973). Second Federal initiatives in the Income Maintenance Program, Follow Through, Head Start Planned Variation and Home Start have shown a distinct move on the part of the government toward attempting to utilize experimental designs. Finally, it has been demonstrated that it is managerially possible to accomplish some of the more difficult features of experimental design such as random assignment in Federally-directed experiments.[2] As a result of this qualified optimism we will devote less energy to articulating the case for social experiments.

The Endless Array of Variables Requiring Measurement

In assessing the impact of schools, the first question is what are the goals or objectives that motivated having schools. Although this may appear to be a straightforwarded question, the difficulty in answering it is consuming.

It is argued that if the schools are to reflect the society that they serve, the answer to what are their objectives can't be simple. James (1971) comments: "We have been notably unsuccessful as a society in this Century in stating our aims for education," while Rosenshine and Furst (1971) concluded

---

[2]In the Follow Through Program, for example, random assignment of children to treatment and control groups was done in a study of the effects of summer schooling.

from their review of research on teaching: "Given the diverse goals of teachers, curriculum developers, students, and test developers, we question whether adequate designs can be developed to study achievement in the typical, uncontrolled situation."

A great deal of time has been devoted to debates concerning the relative emphasis which should be placed on the so-called cognitive and non-cognitive variables. Cognitive outcomes included the three R's, the sciences, social studies, etc. Non-cognitive outcomes include personality characteristics, self-perceptions, values, attitudes, etc. Physical education, vocation training, art and music get shoved back and forth between the two categories. It is generally argued that with such a broad array of child outcomes of potential interest the emphasis given to any one outcome varies greatly within States and even among schools within a school system. These cognitive and non-cognitive categories of school goals serve to polarize constituencies. Bereiter (1973) has argued that schools limit their goals to only the basic skills of the three R's. Others can be found that take nearly the opposite position. Jencks (1972) seems to have concluded that the schools should deemphasize the cognitive outcomes since they appear to be weak predictors of later life chances as he defines them, e.g. income and job satisfaction. He argues instead that school should emphasize short-term effects for their own sake.

We would like to argue that excessive time has been allocated to philosophical considerations regarding the emphasis on cognitive and non-cognitive outcomes. Clearly, if we ask the question, what are the goals or objectives of education, the answer is complicated, and the philosophical debates around this question will be going on long after we have left the professional scene. However, the energy consumed by this first question should be shared with another question: Are we satisfied with the measurement procedures where the objectives of schooling have consensus and the measurement of these objectives has a long history of development and use? This question is directed toward our use of standardized tests in the assessment of content related outcomes.

For example, we have a technology employed throughout the nation to assess student status in the areas of reading and mathematics. Setting aside any debate regarding whether or not the schools should have these subject matter areas as high priorities, how well are we assessing curricula and instruction in these areas?

Our feeling is that there are major problems with the content of our reading and mathematics tests; that there are major

deficiencies in our strategies for using testing instruments in these basic skills areas; that our methodology for test construction is weak; and that our resulting interpretations of child performance are often misleading. (Each of these problems will be discussed in more detail in the next section.)

It appears to us that more of the efforts of the methodological community might be addressed to basic measurement issues within the context of these rather widely-valued subject matters rather than being diverted by debates regarding the outcomes of schooling. Further, steps that are made to improve our measurement capacities in such areas as reading may generalize to other, but perhaps less commonly appreciated, outcomes of schooling.

Although it is impossible to fully document our impression that concerns for analysis have overshadowed concern about solutions to the problems of measurement, a few examples are illustrative. Anderson (1972) reviewed every issue of the Journal of Educational Psychology and the American Educational Research Journal from June 1964 to February 1971 in an effort to better understand types of instruments used in studies of reading comprehension. After having found 130 articles in

which one or more home-made achievement tests were employed he stated:

> "Most investigators reported nothing about their tests beyond such rudimentary information as the number of items and the response made. Several investigators did not hint that a test was used until the analysis of variance was described, at which point, the test was mentioned no more. One investigator characterized his test in a single sentence, 'Criterion achievement was measured by the final achievement test'." - p. 165

We suspect that the lack of detailed information supplied about the way in which comprehension was measured is suggestive of the amount of care and ingenuity put into the development of the tests.

As another illustration Heath and Nielson (1973) concluded after an extensive review of the research on teacher effects:

> "The operational definitions of student achievements are similarly shallow (as those for the variables of teacher process). For example, the criterion of student achievement used in several studies is a ten-question multiple-choice test of information based on Atlantic magazine articles on economics, political, and social conditions in Yugoslavia published between November 1964 and August 1965." - p. 12

At the very least, we are forced to conclude that current practices of reporting research on school effects do not place enough attention on describing what was measured and how.

Clearly unless analysts are explicitly thinking about the measures in their analyses, social experiments will not add much to the knowledge base for making informed policy decisions.

In the past several pages, we have argued that some efforts of the methodological community have been directed towards analysis and design issues and to debates about what to measure when assessing the outcomes of schooling. We have argued that basic measurement problems have not been addressed and feel that a climate of opinion must be developed which focuses attention on measurement issues.

In the next several pages, we will offer our recommendations about categories of measurement issues which deserve early consideration. These issues will include: when to measure, how to measure, and how to interpret the size of an effect. Finally, we will consider the problem of measuring the extent to which a school program is implemented and the importance of knowledge about implementation when interpreting the results of social experiments about school effects.

III. MEASUREMENT ISSUES

When to Measure

Schools sharing common goals may not be in agreement as to when those goals should be realized during a child's school exerience. For example, some schools operate on the belief that a positive self concept about school related activities should be fostered first and that cognitive skills will then follow. Other schools build programs on the assumption that if a child acquires basic skills from his school experience, he will automatically develop a more positive self concept related to schooling. In the case of the Follow Through Program, this has resulted in the dilemma that a comparison of schools emphasizing basic skills at the end of the first few grades will unfairly favor those schools which emphasize the early acquisition of basic skills, whereas a comparison on the affective outcomes may have the opposite bias. Since the ultimate goal is better life chances, neither comparison may be valid. The dilemma is pressed even further by the reality that better measures are currently available for the basic skills than are available for the affective outcomes.

A slightly different aspect of this, "when to measure problem" is illustrated by out-of-school experiences. The success of

Sesame Street children television program is a good example. Many of the children who are frequent viewers of Sesame Street do not need to be taught the same skills in school. The effect is that some teachers are free to spend time on other skills or more advanced levels of the basic skills in kindergarten. If some teachers are moving to more advanced levels of basic skills in the early grades, these skills should be measured earlier in a child's school experience than they have been in the past.

Still another variation of this "when to measure problem" concerns the schedule of testing in schools. Most schools only assess child performance in the spring of the year. However, this standard measurement schedule has limited usage. For example, classroom teachers cannot use such infrequent testing for diagnosis and corrective action for individual children. In addition, many general questions such as how much of a child's score actually represents the contribution of the school versus the contribution of the school plus or minus the contribution of the summer experience cannot be addressed by such standard assessment schedules. Clearly when to measure is a problem that requires more careful consideration than has heretofore been evident.

How to Measure

The two most popular ways to categorize measurement strategies are standardized tests versus criterion-referenced tests. Despite the common use of these two classifications of tests, the distinction between them is not totally clear. By in large, when the label standardized test is used, the user is referring to a test that has been developed for wide use (typically nationally) in assessing differences among students. Prior to being put on the market, the test publisher gives the test to a large sample representative of the population of children for which the test was developed. On the basis of their sample, norms are established so that in subsequent usage an individual child's score can be interpreted in light of how other children have done on the test. For example, a child's performance might be reported as comparable to the typical third grader at the end of the school year.

A criterion test differs from a standardized test in that it does not start with the objective of discriminating among children and does not interpret the performance of a child on the test relative to the performance of some known group of children. Instead a criterion-referenced test has

the objective of giving information on where a child's performance stands relative to some criterion performance. It's at this point that the label of criterion referenced becomes somewhat unclear. Some people wish to communicate that a minimal level of performance has been set a priori and the child's performance is either above or below that minimal level. For those people, a standardized test can also be a criterion-referenced test once the minimal level of performance has been set. Of course, the need for the standardized test norms are no longer. On the other hand, some people who use the label of criterion-referenced do not mean to imply that an a priori minimal level of performance has been specified; but rather that performance on the test has a direct and clear relationship to some continuous criterion. A child's performance on such a criterion-referenced test places the child somewhere on the continuum of the criterion rather than in some rank order position within a norm group.

Neither standardized tests nor criterion-referenced tests are necessarily well suited for use in assessing school effects since both have typically been constructed for assessing individual student behavior rather than differences among

variations of schooling. For purposes of assessing school effects we need tests which are consistent with the goals of schools being assessed and perhaps, to the extent possible, are constructed to tap those skills which are the sole domain of the schools.

The fact that researchers of school effects automatically turned to standardized tests is symptomatic of the lack of careful thinking about measurement problems involved in research on school effects. Since discriminating among schools was not an objective of the standardized tests, it would be sheer chance if they turned out to be useful for that purpose. In fact, we might argue that the purposes of standardized tests operate to make them particularly poor for use in assessing school effects.

One such argument might be as follows: Because most standardized tests have been developed by profit-making organizations, they have attempted to make the tests appropriate to the largest possible population of users. Therefore, in developing measures of student performance, the goal of standardized test developers is to make the test as "fair" to as many

varied subpopulations of students as possible. In making these te "fair," the instrument developers attempt to eliminate items which give an advantage to children with unique experience. Therefore, items which would be more easily answered by children who live in one region or another in the country are eliminated. Similarly, items which are sensitive to unique curricula are also eliminated. The extension of this argument suggests that standardized tests are designed in such a way that they will not be sensitive to many unique instructional interventions.

A second argument stems from the fact that standardized tests attempt to discriminate among individuals. This intent has led the test constructors to select items that are near 50 percent difficulty across subgroups of users in the target population of children and items that discriminate best among individual respondents on the variable being measured. First, items that are near 50 percent difficulty across most schools are not those likely to be sensitive to school effects. Instead valid items that are near 100 percent difficulty for some schools and zero percent for other schools would seem the better choice. Second, items that discriminate among

individual respondents are those items that correlate highest with the total score of the test. As Anderson (1970, p. 165) has pointed out, "People who do well on a test as a whole will have more verbal ability than people who do poorly. Items selected because they discriminate between these two groups will tend to contain difficult vocabulary or require references which are not necessarily critical to an understanding of the concepts and principles being tested. A test constructed to maximize discriminating power will emphasize aptitude and deemphasize achievement."

A closely related concern is that many of the standardized tests are constructed in such a fashion that they measure important prerequisite skills on the part of the respondent that are not an aspect of what the tests were built to measure. For respondents not possessing those prerequisite skills, the tests become a measure of the prerequisite skills rather than that which was intended. For example, many tests of arithmetic involve some reading. On this point, Elsa Roberts (1970, p.30) concluded after reviewing several standardized tests for use with preschool children.

> "Linguistic factors must be taken into consideration in tests for young children even if these tests are not specifically designed to test language. Until there is a great deal more research on the types of structures and operations acquired by age five and on the nature of cross-dialectal comprehension, we must be extremely careful in how we interpret the results of standardized tests and the uses to which we put them."

Finally, most standardized tests do not measure important higher cognitive processes such as creativity and abstract reasoning.

In considering the utility of standardized achievement tests for assessing school effects Klein (1970) concluded:

> "So far, the discussion has painted a pretty bleak picture regarding the utility of standardized tests for accountability. The major problems involve questionable test validity, poor overlap between program and test objectives, inappropriate test instructions and directions, and confusing test designs and formats. In short, a void exists between the demands of accountability and the present stock of standardized instruments. Further, this void will probably only widen as the pressure for accountability increases unless we start improving the methods of test construction and use."

Criterion-referenced tests have developed in part from the emphasis on behavioral objectives in education and so have been constructed to be consistent with the behavioral objectives that that they were designed to measure. They are not as likely to be confounded inadvertently by a mixture of aptitude and experience. Some criticism, however, has been launched

against criterion-referenced tests. They have tended to be too narrow in focus and so have been easy to teach too since they do not tap an entire domain of interest. It is very difficult to construct a set of behavioral objectives which is inclusive enough to adequately reflect the goals of schooling. However, banks of frequent behavioral objectives are being created throughout the country from which a test constructor can select those objectives and items which are desired. An example is the Instruction Objectives Exchange monitored by the UCLA Center for the Study of Evaluation.

Of the two types of tests, criterion-referenced tests seem best suited to the purposes of research on school effects. The criterion-referenced tests for a particular study of school effects should be constructed in such a manner that the battery of tests reflects the full set of common goals. Where this is not possible, the gaps in the battery of tests should be made explicit so that it is fully understood that some of the goals are not being assessed, and that the variations of schooling may or may not have important effects on those unmeasured goals.

Size of Effect

Even when the problems of when and how to measure have been solved an additional problem remains. How does the researcher aid the consumer of his research in interpreting the size of any effects due to the variations on schooling? All too frequently in the past the educational significance of an effect has been equated with whether the effect was large enough to be statistically significant for some popular choice of level of significance. Clearly statistical significance when testing a no-difference hypothesis is not a satisfactory way to judge the educational importance of the size of an effect. First, whether an observed difference is big enough to be statistically significant is a function of several factors other than its size, e. g., sample size, intrinsic variability of the units in the populations being investigated, utility of concommitant variables used to reduce the error variance. More importantly, however, educational policy makers are not interested in the null hypothesis per se. In fact, it is difficult for us to believe that any two serious variations on schooling have exactly equal effects on any outcome measure of interest to educators.

In that sense any comparison of variations on schooling that fails to reject the null hypothesis has made a Type II error.

What is needed to make informed policy decision about alternative models of education is some sense about the size of their difference in effect and how to judge the educational importance of that size of difference. The use of confidence interval estimates of the size of effects helps to solve the problem of reliability, but leaves unanswered the question of how to interpret the importance of differences falling within the interval. Two popular attempts at solving the interpretation of the size of difference problem are: (1) to state the difference in terms of standard deviation units; and (2) to state the difference as percent of variance explained by the variations on schooling. Neither of these approaches to the problem seems particularly useful. Our quarrel with them does not stem from the fact that they both ultimately require an arbitrary choice of size of difference such as saying that a half-standard deviation or more is educationally important or that 10 percent or more of the variance explained is important. Any interpretation of the size of an effect is

ultimately going to rest on some similar type of arbitrary decision and in fact different consumers of the research will (and should) have different standards for what constitutes the smallest difference that is educationally important. Rather, our quarrel with both approaches is their choice of metric on which to make such arbitrary decisions about minimum size of effect. Both metrics are a function of factors other than the level of performance of the children receiving the variations of schooling that are being compared. Two studies investigating the exact same variations on schooling for identical populations of children and identical outcomes measures will have quite different standard deviations and will typically yield quite different percentages of variation explained as a function of the choice of the unit of analysis. For example, if one investigator used school as the unit of analysis while the other used child as the unit of analysis in otherwise identical studies, each will have quite different metrics for interpreting the importance of the size of the observed differences. Using a half-standard deviation as the criterion for educational importance, the investigator using school as the unit has a much better chance of concluding that the difference in effectiveness of the variations was big enough to be important. Choice of units of analysis is not the only factor which can result

in different interpretations of the size of an effect. Studying homogeneous rather than heterogenous populations will result in similar differences and is, in fact, simply a more general statement of the choice of unit of analysis example. In addition, the reliability of the dependent variable will affect the interpretation of the size of an effect in standard deviation units or percent of variance explained.

Our preference is to express the size of the difference caused by variations on schooling in terms of the units of the outcome measure itself. Of course, there are difficulties in doing this since our outcome measures are rarely on an interval scale let alone a ratio scale. A mean difference of five points at one level of the continuum being assessed may have quite a different meaning then the same size difference at another level of the scale. Even if we did have measures that satisfy an interval scale, we frequently have difficulty translating what a mean difference of one point indicates in terms of a child's behavior in the classroom (and out). Yet, by concentrating on the size of an effect in terms of the units of the instrument, we are at least keeping the metric closer to that which is of interest. A five-point difference is a five-point difference regardless of the choice of units of analysis and regardless of whether the population of children was homogeneous or heterogeneous.

What type of information is necessary if the size of an effect is going to be judged in terms of the metric of

the instrument used to measure the child outcomes? We aren't sure we know the answer, though it would appear to be a function of understanding the validity of the test. Clearly the articles reviewed by Anderson for measuring reading comprehension which we cited earlier did not provide the necessary information. At a very minimum the interpreter would have to have a good feel for the types of items on the test and the information required of a child to answer those items correctly. Perhaps it will be necessary to conduct complex validity studies where groups of children who score at different levels of the tests are carefully described according to their school behavior and how their schooling needs differ. In any event, a better understanding of what a test really measures and how the tests measures it appears to be required. The use of standard deviation units and percent of variance explained appears to be in some sense an effort to give a statistical solution to what is basically a measurement problem.

Implementation

So far our concerns about measurement problems in assessing school effects have centered around the outcomes of variations on schooling. There is an additional set of measurement problems, however, that we feel have not received any where near adequate attention and are extremely important

to conducting useful research on the effects of schools. It is not enough to know that variations on schooling have differential effects, we must also know what the variations on schooling actually were. Again our point is an obvious and straightforward one; however, two problems arise. First, when educators develop different models of education, they frequently have difficulty fully describing their model as it should exist operationally when put into the schools. Second, even when the first problem is adequately solved, experience has taught us that it is difficult to get the model or major curriculum change in place in the schools. This second problem is typically referred to as the problem of implementation. To use the Follow Through example again, if two models differ in outcomes it might be due to differences among the models as orginally defined by the developers, or differences in level of implementation of the two models, or some combination. If models are fully implemented, the interpretation of results is reasonably straightforward. If the models differ in level of implementation, however, it is unclear whether the interpretation should be to favor the most successful model regardless of level of implementation or to concentrate

on better implementing the models. Clearly the results might be drastically different for better implemented models. Similarly if models do not differ, it may simply be due to the fact that none of the models really functioned consistent with their definition. Even if the decision is to favor the model with the best effects, disregarding how well the models were implemented, a problem of replication exists. Without really understanding the model as it functioned in the schools, it is impossible to know what it is that you have decided to replicate.

We, therefore, conclude that measuring implementation is necessary in order to provide a context for interpretation of the results of assessing school effects. But how is implementation to be measured? Some have suggested that a model of education has been implemented if it was seen to be effective in terms of child outcomes. For all of the reasons given above, we feel that estimating whether the model had an effect is not a satisfactory solution to the measurement of implementation.

Another suggestion is to use self report of those directly involved in the delivery of the models, e.g., the teacher and the children. Self-report data has the advantage of

being inexpensive and is undoubtedly better than no information at all. For example, a recent evaluation of the Hilda Taba method of education in the elementary grades used teacher interviews to discover that by their own report the teachers had not mastered many of the teaching skills which defined the TABA model. Nor had they attempted to systematically use the few skills that they did report having mastered. The results of the teacher interviews presented a useful context within which to interpret the results of the analysis of child outcome data which indicated that the children in the Taba Program achieved no better than the control children (Porter, 1974).

Had the teachers self reports been more positive about their mastery and use of the skills in the TABA model, however, the validity of the self-report data might have been questioned. In addition self-report data is limited to the extent to which the participants are in good positions to judge whether the full model has been implemented.

A third suggestion for measuring implementation of a model for schooling has been direct classroom observation. For models that are primarily delivered in the classroom setting, classroom observation seems to be the most

satisfactory strategy, but it has the disadvantage of being quite expensive. Stanford Research Institute has undertaken what is probably the largest classroom observation effort thus far. Their results have documented that the Follow Through models do differ systematically in terms of classroom activities and interactions. They also show that some models are more systematically implemented across sites than others.

In addition to cost, classroom observations have other limitations. For example, how often should a classroom be observed and at what times? Clearly a teacher may behave differently during the single time that her classroom is observed than she behaves on a typical day. At presnt there is little empirical evidence that is useful to help construct a schedule for classroom observation that is likely to produce valid data on implementation. Such data are needed.

The difficulties that educators have experienced in the past with implementing models of education in the schools also suggests that studies conducted to evaluate the effects of various models should allow sufficient time for those models to become implemented and stable in the

schools. During the first stage of such an evaluation, the evaluators should concentrate on assessing the extent to which the models are implemented, i.e., implementation is an end in itself. If the models are not implemented to a satisfactory degree after an acceptable period of time, the evaluation is completed. If the models,or at least some of the models, are judged to be satisfactorily implemented in stage one of the evaluation, those models are then compared on child outcomes in the second stage. The paradigm suggests two things. First, implementation should be studied in its own right. The result might be that giving greater attention to the importance of proper implementation, we would learn more about what it takes to implement a model in the schools. Second, the paradigm prevents the comparison of models in terms of their desired outcome prior to having been implemented. It is not difficult to imagine that attempting to implement some models is quite distracting from that which typically takes place in schools. Therefore, the first year of the model may be deterimental to the desired outcomes, even when in the long run the effect of the model will be positive.

Devoting a first stage of an evaluation to studying implementation is expensive. In light of what we know

about how to measure implementation at the present time the expenses may seem uncalled for. The alternative, however, seems even less satisfactory, i.e., to attempt to assess the effects of something; but you are not sure what.

In conclusion, we feel that issues of design and analysis of research on school effects have been given emphasis out of proportion to their importance. We are committed to the utility of social experiments and recognize that there are important design and analysis issues which need to be resolved, but their resolution will have little impact on important policy decision unless some basic problems of measurement are also resolved. Our review of current measurement practices leaves us with the distinct impression that the measurement issues have received far less attention than have the design and analysis issues and concomitantly that the quality of our knowledge of such issue has logged behind.

Although we recognize that schools have diverse goals, it is our opinion that the philosophical arguments about the goals of schooling will never be resolved. Therefore, research on the effects of schooling must proceed by concentrating on shared goals with appropriate caveats about diversity. For those who accept our position, several

important measurement issues deserve immediate attention. First, more thought and analysis must be given to the question of when to measure school outcomes and what are the implications of alternative testing schedules. Second, policy decisions need to be based on the size of effects of variations on schooling rather than whether or not they differ more than by chance. The problem of interpreting the size of an effect should be attacked in the metric of the test rather than in some metric which is a function of the population investigated. This suggests that a better understanding of test validity is necessary. Finally, we urge that greater attention be given to the problem of how to measure the extent to which a model of schooling has been implemented in a study of that model's effectiveness. Better knowledge about implementation is necessary, both as a context for interpreting studies of outcomes, as well as to facilitate studies of replication.

REFERENCES

Anderson, R. C. How to construct achievement tests to assess comprehension. Review of Educational Research. 1972, 42, 145-170.

Bereiter, C. Must We Educate. Englewood Cliffs, N.J.: Prentice-Hall, 1973.

Campbell, D. T. Methods for the experimenting society. Paper delivered at the meeting of the American Psychological Association, 1971.

Coleman, J.S., et al. Equality of Educational Opportunity. Washington: U.S. Government Printing Office, 1966.

Evans, J. and Williams, W. The politics of evaluation: the case of Head Start. Annuals of the American Academy of Political and Social Science, 1969, 385, pp.118-132.

Goldberg, M. School system perspectives on Follow Through in Alice Rivlin and P. Michael Timpane (ed.'s) Planned Variation Experiments: Should W ive up or Try Harder? Washington: The Brookings Inst n, 1974.

Heath, R. W. & Nielson, M.A. The of performance-based teacher education. Paper presented at the American Educational Research Association meetings, 1973.

James, H. T., Excerpts from the preliminary report to the National Academy of Education's Executive Council meeting May 6, 1971, on the feasibility of an Academy task force to explore the reporting of performance by educational institutions.

Jencks, C., et al. Inequality. New York: Harper & Row, 1972.

Klein, S. P. The uses and limitations of standardized tests in meeting the demands of accountability. UCLA Evaluation Comment, Center for the Study of Evaluation, 2, No. 4, 1971.

McDaniels, G. L., Attrition as an analytical and programatic problem for Head Start Planned Variation and Follow Through. Paper delivered at the meeting of the American Psychological Association, 1972.

McDaniels, Garry L. The Follow Through evaluation: design and issues. In Alice Rivlin and P. Michael Timpane (ed.'s) Planned Variation Experiments: Should We Give Up or Try Harder? Washington: The Brookings Institution, 1974.

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

Porter, A. C. Some design and analysis concerns for quasi-experiments such as Follow Through. Paper presented at the meeting of the American Psychological Association, 1972.

Porter, A. C. Analysis strategies for some common evaluation paradigms. Paper presented at the American Educational Research Association meetings, 1973.

Porter, A.C. An Evaluation of the TABA and BASIC Teaching Strategies Programs in the Lansing Public Schools, 1972-73. Michigan State University, Final Report to the Michigan Department of Education, 1973.

Porter, A. C. & Chibucos, T. R. Selecting analysis strategies. In G. Borich (ed.) Evaluating Educational Programs and Products. Educational Technology Press, 1974. pp. 415-464.

Rivlin, Alice and Timpane, P. Michael (ed.'s) Planned Variation Experiments: Should We Give Up or Try Harder? Washington: The Brookings Institution, 1974.

Roberts, E. An evaluation of standardized tests as tools for the measurement of language development. Unpublished paper, Northwestern University, 1970.

Rosenshine, B. Teaching Behaviors and Student Achievement. London: National Foundation for Educational Research in England and Wales, 1971.

Rosenshine, B. & Furst, N. Research on teacher performance criteria. In B.O. Smith (ed.) Research in Teacher Education: A Symposium. Englewood Cliffs, N.J.: Prentice-Hall, 1971.