DOCUMENT RESUME

ED 114 261 95 SE 018 802

AUTHOR Berliner, David C.
TITLE A Status Report on the Study of Teacher Effectiveness.
INSTITUTION ERIC Information Analysis Center for Science, Mathematics, and Environmental Education, Columbus, Ohio.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE Mar 75
NOTE 21p.; Occasional Paper Series

EDRS PRICE MF-$0.76 HC-$1.58 Plus Postage
DESCRIPTORS *Accountability; *Educational Research; Effective Teaching; Evaluation Criteria; *Evaluation Methods; *Performance Based Teacher Education; Research Methodology; Science Education; *Teacher Behavior; Teacher Influence
IDENTIFIERS Research Reports

ABSTRACT

This report discusses the fact that many educators are committed to competency based teacher education and teacher accountability systems in spite of the lack of empirical evidence linking teacher behavior to student outcomes in the classroom. Some of the difficulties associated with research in this area are identified as problems in instrumentation, methodology, and statistics. Specific problem areas include the inadequacy of standardized tests, the unknown predictive ability of tests from special teaching units, the problem of building multivariate outcome measures, the problem of measuring the appropriateness of teacher behavior, the lack of experience in choosing an appropriate unit of analysis for describing teacher behavior, and the lack of stability of many teacher behaviors. Further research is recommended on how student backgrounds affect measures of teacher effectiveness, what subject matters should be examined, how normative standards and volunteer teachers affect what we can say about teachers and teaching, how individual students react to teaching skills, how students monitor and interpret a teacher's behavior in ways which may or may not coincide with how educational theorists interpret the phenomenon, and studies on the validity and generalizability of measures of teacher effectiveness. (Author/MLH)

SCIENCE EDUCATION INFORMATION
REPORTS

OCCASIONAL PAPER SERIES

A STATUS REPORT ON THE STUDY OF TEACHER EFFECTIVENESS

By

David C. Berliner
Far West Laboratory for Educational Research and Development
San Francisco, California 94103

ERIC Information Analysis Center
for Science, Mathematics, and Environmental Education
The Ohio State University
400 Lincoln Tower
Columbus, Ohio 43210

March, 1975

CONTENTS

SCIENCE, MATHEMATICS, AND ENVIRONMENTAL EDUCATION INFORMATION REPORTS

The Science, Mathematics, and Environmental Education Information Reports are being developed to disseminate information concerning documents analyzed at the ERIC Science, Mathematics, and Environmental Education Information Analysis Center. The reports include four types of publications. Special Bibliographies are developed to announce availability of documents in selected interest areas. These bibliographies will list most significant documents that have been published in the interest area. Guides to Resource Literature for Science, Mathematics, and Environmental Education Teachers are bibliographies that identify references for the professional growth of teachers at all levels of science, mathematics, and environmental education. Research Reviews are issued to analyze and synthesize research related to science, mathematics, and environmental education over a period of several years. The Occasional Paper Series is designed to present research reviews and discussions related to specific educational topics.

The Science, Mathematics, and Environmental Education Information Reports will be announced as they become available.

Occasional Paper Series - Science

The Occasional Paper Series (Science) is designed to present reviews of literature or discussions related to specific topics or educational programs related to the teaching and learning of science. We hope these papers will provide ideas for implementing research, suggestions for areas that are in need of research, and suggestions for research design.

The ERIC Science, Mathematics, and Environmental Education Information Analysis Center has cooperated with the National Association for Research in Science Teaching in sponsoring this paper as a General Session presentation at the 48th Annual Meeting in Los Angeles, California, March 17-19, 1975.

This paper will serve as the basis for a journal article to be published in the near future.

Stanley L. Helgeson
and
Patricia E. Blosser
Editors

The material in this publication was prepared pursuant to a contract with the National Institute of Education, U. S. Department of Health, Education, and Welfare. Contractors undertaking such projects under government sponsorship are encouraged to express freely their judgment in professional and technical matters. Prior to publication, the manuscript was submitted to the National Association for Research in Science Teaching for critical review and determination of professional competence. This publication has met such standards. Points of view or opinion, however, do not necessarily represent the official view or opinions of either the National Association for Research in Science Teaching or the National Institute of Education.

A STATUS REPORT ON THE STUDY OF TEACHER EFFECTIVENESS*

David C. Berliner
Far West Laboratory for Educational Research and Development
San Francisco, California 94103

Advocates of performance or competency based teacher education, state mandated evaluation programs, such as the Stull Bill in California, and teacher accountability systems, all suffer to some degree from ostrichism. Ostrichism is a common disease often afflicting education. It's etiology is in a premature commitment to a particular educational movement. Behavioral symptoms include the practice of sticking one's head into the sand when problems appear, in the hope that the problems will go away.

The particular educational movement which is inducing the current epidemic of ostrichism is the commitment of educators to competency training and evaluation without the existence of empirical evidence linking teacher behavior to student outcomes in classroom settings. The Coleman report (1966), and its offshoots (Jenks, 1972; Mosteller and Moynihan, 1972), have minimized the role of the teacher in accounting for educational outcomes. These investigators claim that family background, socioeconomic status, ethnicity and the like, are the major causal variables affecting between school differences in achievement.

In that same tradition is the criticism of Heath and Nielson (1974). Their review of the studies of teacher clarity, use of student ideas, criticism, enthusiasm, and other variables commonly accepted as skills or competencies, has revealed serious flaws in the extant research. They concluded first that there is no established empirical relation between teacher behavior and student achievement. Second, that the flaws in the research are due to nonsensical statistical analyses, weak research designs, and sterile operational definitions of teacher behavior and student outcomes. And third, that because of the strong association between omnibus measures of student achievement and socioeconomic and ethnic status, the effects of teachers and techniques of teaching on achievement are bound to be trivial.

These are serious criticisms of the effects of teaching on student achievement. Yet unless replicable findings relating teaching behavior to student achievement in natural classroom settings can be found, the performance and competency based teacher education, evaluation, and accountability programs will not be believable. Let us remember that the heart of the performance and competency based approaches to teacher education, teacher evaluation and teacher accountability has to be the

*The ideas presented in this paper have emerged from discussions with the staff of the Beginning Teacher Evaluation Study of the Far West Laboratory for Educational Research and Development. This is a project of the California Commission on Teacher Preparation and Licensing, funded by the National Institute of Education. The comments of Margaret Bierly, Leonard Cahen, Nikki Filby, Charles Fisher and Marjorie Powell are gratefully acknowledged.

empirically established relationship between teacher behavior as an independent variable and student cognitive and affective outcomes as dependent variables. Whether we are interested in effective science teaching, as this group is, or effective mathematics or home economics teaching, establishing empirical relationships between teacher behavior and student outcomes has to be our goal.

Ferment exists because performance and competency based education, in all its forms, has been sold before it really exists (cf. Shanker, 1974). Those who use research to criticize teachers, teaching, and performance based teacher education, as well as those who defend teachers, teaching and performance based approaches have all taken positions before they have the necessary empirical backing. There is not now, and there will not be for sometime, any empirical evidence to take any firm position on these issues. Extremely important problems hamper the study of teachers and teaching in all subject matter areas. I believe it will take years before these problems can even be understood well enough to do classroom research properly. I think you should keep in mind that the first step in the systematic study of any phenomena is the recognition of what problems exist in that research area. Addressing these problems, rather than assuming they will go away, or that they do not apply, will enhance the likelihood that studies of teacher effectiveness will be fruitful. The problems, as I see them, are loosely grouped into three categories concerned with the instrumentation, methodology and statistics used in studying how teachers affect the achievement of students.

INSTRUMENTATION PROBLEMS

There are serious instrumentation problems connected with both the independent and dependent variables commonly used in research on teacher effectiveness. Six of those issues are discussed here.

Dependent Variable Problems

Our work at the Laboratory has been hampered by an inability to satisfactorily resolve three problems connected with development of dependent variables. These problems are connected with standardized testing, tests of special teaching units, and development of multivariate outcome measures.

Standardized testing. In studies of how teachers affect students, standardized achievement tests are extensively used as criteria or outcome measures. These tests are, as a group, highly reliable instruments. They usually have adequate curriculum content validity, and seem predictive of future academic success. These tests have, however, one overwhelming flaw. They simply may not reflect what was taught in any one teacher's classroom. The tests are designed to be used in all kinds of courses within a curriculum area, and therefore cannot be completely sensitive or appropriate for any one teacher's teaching (Gall, 1972). They simply lack content validity at the classroom level.

The standardized achievement tests are also highly correlated with standardized intelligence tests, thus causing us to wonder exactly what

kinds of items are really used in these tests. Furthermore, the tests are usually group administered multiple-choice tests. When working with young, bilingual, or lower socioeconomic status children, there is a serious question about whether many of the children are being appropriately tested.

In our own work, when standardized tests must be used, we try to refine the items in a number of ways. We try to choose items where there is evidence of substantial change in difficulty level over some instructional period. In this way we hope to identify items that are reactive to instruction. We try to pick items that correlate weakly with a measure of general intelligence, like the Raven's Progressive Matrices test, rather than picking those items with higher saturations of general intelligence. We try to have teachers rate items on how much time it would take them to teach that idea, or, how much emphasis they put on material like that addressed by the item. Unless items on a standardized test are put through a systematic screening of this type, the test is not going to be particularly reactive to teaching. Off-the-shelf standardized tests make poor dependent variables for studies of teaching. This is part of the difficulty in interpreting the Coleman report. The tests they used in that study were more reactive to family background and ethnicity than they were to instructional events within the school. It does not directly follow from this kind of evidence that teachers have no effect on student achievement.

Tests for special teaching units. To insure the use of tests that are content valid for a particular classroom, many investigators of teaching have created special teaching units, or content vehicles to study teaching (Berliner and Ward, 1974; Joyce, 1975; Popham, 1971). An experimental unit of this type contains curricula materials, objectives, and sample test items. The teacher is asked to teach to the objectives. The unit could be a single 30-minute lesson, or require daily work over three weeks. Under these conditions every teacher has similar materials and objectives to work with. Students are pre and post tested with carefully constructed tests designed to tap many dimensions of the material in the experimental teaching unit. The dependent variable in this situation is much more valid and much more reactive to classroom teaching. In comparative studies of teaching effectiveness, these experimental teaching units, and their tests, have much to commend them. Each teacher has a similar chance to try to produce gains in student achievement. Some teachers will be better at this than others.

Unfortunately, at this time in our research efforts, we do not know if the measures of teaching effectiveness arrived at over a short period of time provide an estimate of teacher effectiveness over a longer period of time. This methodology, which is used in our research on teaching, allows us to use tests of high content validity that seem to accurately reflect classroom practice for a short period of time. But this methodology may not have any predictive validity. We do not know if the ranking of teachers on effectiveness, as determined by the relationships between student pre and post test scores associated with an experimental teaching unit, is at all correlated with a ranking of those teachers over the whole school year. We will have information on this issue later in the year. Frankly, we do not now expect a measure of teaching effectiveness obtained over a short period of time to correlate

very highly with a measure of teaching effectiveness for an entire school year. Thus studying teacher effectiveness with dependent measures tied to special teaching units may not, in our estimation of the state of the art, be a fair characterization of teaching over the long haul. Predictive validity with such materials appears to be too low.

Multivariate outcomes. There are at least two dependent variables in any instructional interaction that should be of interest to us. One of these is the achievement of the learner in the situation. This has been a commonly used measure of instructional outcomes. The other, less often examined, is the learner's feelings about the instructional situation. We do not always ask students questions which probe their liking for their teacher or the subject matter. We overlook inquiring about their enjoyment of their classmates, the degree of threat felt in the class, and whether or not they would take more courses in that area. When such issues are addressed in research studies, the affective set of dependent measures is kept separate from the achievement measures.

Our problem in the research we do is to find ways to use multivariate outcomes so that many kinds of achievement and affective responses are used as indicators of the quality of classroom life for a child. I think the problem is something like the difficulties in teaching reading. You can get high comprehension at slow reading rates. Or you can get low comprehension at high rates of reading. But it is obvious that there must be some optimum multivariate outcome that simultaneously considers both reading comprehension and speed. The same kind of multivariate outcome measures, simultaneously considering both achievement and affective outcomes is needed for research on teaching. If we do not consider what is learned and what is felt about that learning, simultaneously, we stand to fractionate school learning into pieces that do not resemble the students' view of reality.

Independent Variable Problems

Our work has also been hampered by problems connected with the independent variables used in studies of teacher effectiveness. A major difficulty we have encountered is related to the issue of appropriateness of teacher behavior. A second issue is related to the determination of a unit of analysis for the independent variable. A third issue is concerned with the stability of teacher behavior.

Appropriateness of teaching behavior. My colleagues and I have spent a good deal of time counting teacher behaviors. We know something about the number of higher and lower cognitive questions asked per unit time, we have counted the rate of positive verbal priase, the number of criticisms made, the number of probes, the frequency of explaining links, etc. For many of these variables we have found a low correlation with some student outcomes measures. But in our classroom observations we have become acutely aware of the difference between a higher cognitive question asked after a train of thought is running out, and the same type of question asked after a series of lower cognitive questions has been used to establish a foundation from which to explore higher-order ideas. We have seen teachers ask inane questions. We have seen teachers direct questions to what we believe was the wrong child. We have seen positive verbal reinforcement used with a new child in the class, one

who was trying to win peer group acceptance, and whose behavior the teacher chose to use as a standard of excellence. We watched silently as the class rejected the intruder, while the teacher's count in the verbal praise category went up and up and up. We have seen teachers respond to student initiated questions with irrelevant information. We have seen teachers achieve a high rate of probing student responses to questions, seemingly without regard for the student or the kind of initial response given to a question. Some students were embarassed by the probing, with other students probes occurred at inappropriate times, and sometimes probes were not used when the situation seemed to cry out for them. Similarly, we observed skillful probing where a student's knowledge about an issue was brought out and shared with the class, after a weak first response was given by that student. The teacher's questioning was as skillful as Plato's, but we had recorded only its frequency.

All these events have led us to reassess our strong behavioristic stance in the study of teaching. We still regard frequency counts as very useful information. But we now feel quite strongly that the qualitative dimension, dealing with value judgements about appropriate use of skills, must enter into our observations of teaching. We must address the appropriateness issue in order to study the information processing and decision making skills of human teachers. It is precisely these skills that provide the most important rationale for having human teachers in the classroom.

The unit of analysis for the independent variable. Something else we have become acutely aware of in our studies of teacher effectiveness is the problem of the unit of analysis for characterizing the independent variable. Is the single teacher question the unit of interest? Is the question, along with the wait-time, the unit? Or is the teacher question, wait-time, and student answer the unit which best characterizes the independent variable? And if the latter is most appropriate, does that transaction become part of an episode or strategy of even more complex dimensions and longer duration? Teachers follow strategies of questioning and of discussion. In an inductive lesson the meaningful unit of analysis may be a one-hour or one-week episode that is concerned with the conservation of matter. The individual questions, reinforcers, probes and student responses may be trivial aspects of the overall episode. We certainly need to think about new conceptions for the units underlying independent variables used in studies of teacher effectiveness.

Something else about the nature of an instructional episode has perplexed us. We have found very little data describing the nature of the instructional activities and episodes a child engages in each day. Since instructional time appears to be an important variable in the learning process (Wiley, 1973; Harnischfeger and Wiley, 1975) we need to obtain accurate records of how time has been allocated to the various instructional activities and episodes we might identify. The work of Gump (1967) and the techniques of Barker (1968), are useful starting points for obtaining this kind of information. This perspective yields accurate descriptions of the time a child spends in various activities and the time he is exposed to instructional episodes of various types.

These activities and episodes can be treated as independent variables and may be causally related to various types of student outcomes.

Stability of teacher behavior. Before an observer enters a classroom to code teacher behavior in any sensible way, he has to be sure of two things. First, that the frequency of the events he is trying to observe is high enough so that at least one instance of the event will occur during the observation period. Second, the behavior to be coded should represent the teachers usual and customary way of behaving. Only if these conditions are met can a teacher's behavior be sensibly characterized by the frequency count or rating scale description obtained in observation of classroom activities. These basic requirements for observation must be examined closely.

Many studies relating teacher behavior to student outcome have examined teacher behavior that did not occur frequently. For example, among 32 primary-grade science teachers the use of questions calling for identifying relationships, hypothesizing, and testing hypotheses are extremely rare events on any given occasion of observation (cf. Moon, 1969; 1971). Another case of low frequency events, in an important area of teaching, has to do with the management skills of teachers. We find that in some communities classroom management is not too difficult. The students are motivated and parents exert tight behavioral control, so that traumatic disturbances are quite infrequent. In other communities serious management problems exist all day long. So we find that to observe instances of teacher behavior in the area of classroom management, we must remember to take into account ecological factors. Furthermore, we have learned that even in settings where management problems usually occur with high frequency, certain teachers are so quick to establish a non-disruptive social system that, by the time the observer enters the class, particular kinds of events have been precluded from occuring.

How then can one study teacher behavior when important variables in the study rarely occur? One answer, of course, is in denser observation than is customary. Five one-hour observations of teacher behavior, which is unusually high for most studies of teaching, may simply not provide all the information an investigator may want. In addition, part of the answer is in knowing when and where to observe. For example, the first two weeks of schooling would be important for a study of management skills in inner city schools. Simply trying for denser observation, later in the year, in other types of schools, might be wasted.

The problem of estimating behavioral stability is partly related to the problem of the frequency of occurrence of behavior. When the frequency of a behavior is low the correlations between the frequency of occurrence for certain events, over occasions (that is, a coefficient of stability for the behavior), will be low. But part of the problem in looking at stability of teacher behavior is quite distinct from the frequency issue. Think for a moment about the characteristics you prize in a teacher. Usually, people think of "good" teachers as flexible. Such teachers are expected to change methods, techniques, and styles to suit particular students, curriculum areas, time of day or year, etc. That is, the standard of excellence in teaching that we hold implies a teacher whose behavior is inherently unstable. Needless to say, there

is a problem for an observer who is trying to measure a teacher's customary and usual ways of teaching.

For our study of teaching we have reviewed teacher stability, over occasions, for a great many variables (Shavelson and Dempsey, 1975). The results are fascinating. On the laughable side are coefficients of stability from Campbell's (1972) analysis of science teaching at the junior high school level, over two occasions. The Flanders Interaction Analysis System was used, and the stability coefficient, that is, the correlation between a teacher's standing on a measure across two occasions was, for a measure of indirectness in teaching (i/d ratio), -.90. On five occasions Moon (1969; 1971) studied 32 primary grade science teachers trained in the Science Curriculum Improvement Study (SCIS). The stability coefficient for the Flanders indirectness measure went all the way up to +.18; for the frequency of fact or recall questions, the stability coefficient was -.12; and for amount of teacher talk, only +.12. In Borg's (1972) study, the behavioral stability of teachers was measured after training in questioning techniques had taken place. The stability of the ratio of higher-order to fact questions was .07. The rather large number of low and even negative stability coefficients which exist in the literature confirms our belief that the independent variables we often work with in studies of teacher effectiveness are not fair indicators of a teacher's typical behavior. We are so eager to capture variables for data analysis with our rating scales and frequency counts, that we seem to have forgotten to check if our methodology is appropriate to the phenomena we are interested in studying!

Of course there are many exceptions to the trend for teacher behavior to be unstable. We have found ratings of variables over 10 occasions that yield high stability coefficients. These include stability coefficients of .92 for teacher warmth; .79 for teacher enthusiasm; and .83 for teacher sensitivity (Wallen, 1969). We have found frequency counts demonstrating that a global variable composed of all types of reinforcement is reasonably stable over occasions, yielding a stability coefficient of .64 (Trinchero, 1974). In the latter study, however, we find considerable evidence pointing to the lack of generalizability of stability coefficients across different teacher populations, curricula areas and student populations. For example, the stability coefficient over two occasions for the frequency of positive verbal teacher behavior was .04 for English teachers, and .57 for social studies teachers.

By examining the stability of teachers' behavior, which is used as the independent variable in studies of teacher effectiveness, we conclude that: 1) some teacher behaviors that we think are important to study occur infrequently. To study them requires extensive observation in particular settings at appropriate times; 2) some teacher behaviors that we think are important to study are basically unstable over occasions. No practical amount of observations will result in a reliable estimate of a teacher's use of these behaviors. Perhaps we need to develop measures of variance instead of measures of central tendency to describe those behaviors; 3) some teacher behaviors are stable over occasions. In general, but not always, ratings or high inference variables, rather than frequency counts or low inference variables, are the more stable; 4) stability coefficients for many teacher behaviors will

not demonstrate ecological or population validity. Teacher behavior is moderated, as it should be, by the kinds of students and the variety of settings that teachers work in. Until we know more about which teacher behaviors fluctuate, and how and why they fluctuate over time, settings, curricula, and populations, studies relating teacher behavior to student outcomes must remain primitive.

METHODOLOGICAL PROBLEMS

A loosely related set of issues has been grouped under the title problems in methodology. Each of the problems and issues mentioned is in some way hampering the development of reliable knowledge about the relationship between teacher behavior and student outcomes.

Student Background and Teacher Effectiveness

One problem in studying the teaching process is estimating how much can legitimately be expected of teachers or schools as an influence on student growth. This problem is debated in educational philosophy, sociology and economics, as well as educational psychology. And this issue has already been mentioned when we discussed how procedures are needed to reduce the influence of intelligence and ethnicity on test performance in studies of teacher effectiveness. But the problem is even more pervasive. Can a teacher be held accountable if a perfectly appropriate prescription is given, and then not followed by students? Suppose a teacher says, "read this chapter and come to my office so we can discuss it." Among sub-cultures that see schools as hostile or useless, students will not read the chapter and will not come in to discuss it. Classes of such students may show minimum growth in achievement at the end of the year. And these low achieving classes may very well be made up of lower socioeconomic status children and ethnic minorities. Under these conditions, how much responsibility is to be placed on teachers for the low student performance?

On the other hand, with high intelligence, high socioeconomic children, growth in achievement takes place almost in spite of the teachers and teaching. Can the achievement of students in those settings be attributable to teachers, or is it a product of genetic and environmental advantage, relatively unaffected by what teachers do?

Since some children, often whole groups of children, may be unwilling to learn in the institutions we now use to educate them, and some children learn in those institutions regardless of what happens to them, how do we go about attributing student achievement to what teachers do? In the case of low achieving students we feel we may have to evaluate teachers against some other criteria than student achievement, yet to do so denies that teachers can and should make a difference in the achievement of lower socioeconomic and minority children. I have no solutions to this problem. I only know it exists and must be thought about as people naively discuss teacher effectiveness without qualifying what they say by noting the students' background characteristics, particularly socioeconomic status and intelligence.

The Subject Matter and Teacher Effectiveness

That student background characteristics influence test performance and almost all other aspects of schooling is well established. What was not so well understood, until recently, is that student performance in different curriculum areas is differentially affected by those background characteristics. In the International Education Association's (IEA) cross-cultural study of student achievement (Postlethwaite, 1973), the variance accounted for by student background characteristics, such as intelligence and social class, was estimated for a number of subject matter areas. Clearly highlighted, around the world, was that home influences on subjects like reading and social studies are very powerful. Those influences are so powerful, in terms of their accounting for student achievement, that there may not be enough variance unaccounted for in the performance of students to attribute to the influence of teachers.

But in other curriculum areas, student background accounts for much less variance. Physics, chemistry, French, Spanish, geometry, and trigonometry are not typically learned at home, and therefore the schools account for more variance in these measures of achievement than for achievement measures in reading, social studies or language arts. This does not mean that socioeconomic status and intelligence are not related to performance in science, foreign language or mathematics. It simply means that the influence of those background factors is much less, thus leaving more variance to potentially attribute to school and teacher effects.

If we want to study teaching we should study it in those areas where we are most likely to be able to attribute an effect to teachers, after the influences of test unreliability and home background have been removed. Instead we typically study teaching in those subject areas where we are hardest pressed to causally relate teaching behavior to student outcomes. New approaches are called for.

Normative Standards and Volunteer Samples in the Study of Teacher Effectiveness

Our own work and that of many of my colleagues, is, in simplest form, a comparison of the post-instruction test scores of classes that had similar pre-instruction test scores. These comparative differences in outcomes are believed to discriminate between more and less effective teachers. Our research approach is entirely normative. And in a norm referenced research study some teachers will always appear to be better than others. In fact, the whole sample of teachers in any study may be quite poor when judged against some absolute standards, and we would never know.

More likely, since studies of teacher effectiveness in natural environments require the informed consent of volunteer teachers, we are likely to do research with a sample of self-confident, relatively open teachers, almost all of whom may be superior to a non-volunteer sample on an unknown number of unidentified dimensions. But in a norm referenced system, where teachers are evaluated against other teachers, we will judge some of our sample to be less effective than others. This

is a silly research strategy, but one we cannot easily change. To bring about change in this approach we would need to impose criterion referenced achievement standards for teachers, and require all teachers to participate in research of the type we are talking about. Until we can do that, and I doubt we ever will, we should never talk of effective and noneffective teachers. We are, at best, dealing with more and less effective teachers, which is quite different from the absolute criteria implied by the terms effective and noneffective. And because our norm referenced research is done with volunteer samples, our statements about teacher effectiveness should also include some reference to the fact that these are more or less effective teachers from a sample of teachers that are themselves probably superior to the average teacher in an unknown number of ways.

Individual Differences Among Students and Teacher Effectiveness

All teachers known that some of the things they do will not be effective with some of the children they teach. There is no feeling of failure when this occurs, that's just the way things are. Most teachers recognize this problem and modify instruction accordingly. They customize their behavior, as best they can, to fit the individual styles of students. Our research on teacher effectiveness, however, usually ignores this phenomena. We rarely collect enough individual difference measures on students to find out if particular teaching behaviors are differentially effective with different types of children. For example, from what we know about how aptitudes and treatments interact (cf. Berliner and Cahen, 1973), we can expect that a highly structured course in science, taught by a well organized somewhat dominant teacher, will yield greater achievement for high anxious students than for low anxious students. On the other hand, the low anxious student will probably perform better than the high anxious student in the class of a science teacher providing only small amounts of guidance and using an inductive approach. In research on teacher effectiveness we ordinarily find no relation to student achievement outcomes for teacher behaviors that help to define constructs like inductive or deductive teaching style. Relationships may not appear because we do not know how to partition students into meaningful subgroups for whom the two different treatments might be uniquely applicable. If we could have divided students into high and low anxious individuals, to follow our example, we might have found that teacher behaviors within each teaching style had important effects on student achievement.

I have no doubt that the styles of teaching and teaching behavior recommended by, say, the curriculum guides accompanying new science curriculum projects are appropriate recommendations for some teachers, when interacting with some students. But not all students! By not focusing on the individual aptitudes, styles, personality, and traits of the student, we mask the effects of teachers, thus making it almost impossible to establish empirical relations between teaching behavior and student outcome.

An equally important reason to use the aptitude-treatment interaction approach is to find teacher behaviors that in general have positive relationships with student outcomes, but are, in fact, negatively affecting the performance of small numbers of students. Research

on teacher effectiveness has to begin searching for interactions as it continues trying to establish more general links between teacher behavior and student outcomes.

Mediation of Teacher Effectiveness Through the Student's Behavior

Another aspect of classroom reality that must be brought into our designs for research on teaching skills and competencies, is the fact that teacher behavior does not influence student achievement directly. That is, a teacher's indirectness, or questioning, or reinforcement does not simply result in greater mathematics, reading, or science achievement. The link that must be understood is the behavior of the student in the instructional setting. We are now convinced that the mediating link so necessary to consider is a students active time-on-task. If teacher questions, reinforcement, warmth, and clarity are to affect outcomes, they can only do so by engaging and then keeping the student's attention. If the student will attend, the possibility of learning exists. We need to look at teacher behaviors that affect student active learning. To do so means putting much more effort into clinical studies. In this way an investigator can work one-to-one with students, trying to understand how the student allocates his attention, and how nominal stimuli emitted by the teacher, become effective stimuli for that student. To think that there is a direct link between, say, a teacher's questions which require the generation of hypotheses by students, and the students' achievement on a science test is overly simple. Intermediate links in that causal flow require us to examine the student's attending and information processing behavior.

Another aspect of the student that must be thought about for research in teaching is the student's perspective of the events that impinge upon him in classrooms. We do not know how much of what we call skilled teaching is even perceived by the learner. From the learners perspective, perhaps "analysis" and "synthesis" level questions are not distinguishable. Students may differentiate only "memory" and "thinking" questions. From the learner's perspective the rate of reinforcement may be irrelevant. The teacher either is "nice" or "not nice" to students. I believe that some variables thought to be quite important by educational theorists are in fact unimportant, unperceived or unperceivable by students (cf. Dunne, 1974). Students exposed to variables they cannot perceive or to variables they believe to be unimportant, may be unaffected by such variables. We certainly need to follow Snow's (1974) advice to researchers that urges more detailed accounts of what learners do in response to experimental treatments.

Construct Validation and Teacher Effectiveness

Through the writings of the logical positivists, and particularly the physicist Bridgman, social scientists became aware of the critical nature of language and operations in science. An initial development to further scientific understanding of some phenomena is a descriptive language that uses concepts having common meaning among the scientists working in the same area. The intensive and extensive meaning of key concepts needs to be shared by the members of the scientific community the less the overlap of shared meaning, the less rigor the science can

develop. A case in point would be a term like "withitness" from the study of teaching by Jacob Kounin (1970). The teacher who can spot trouble before it begins has "withitness." Such a teacher can be working with one group of students and call out a student's name at the other end of the room because he is beginning to cause a disturbance. That is "withitness." I recently went into a classroom and one of the concepts that helped me organize what I saw was the concept of "withitness." I felt perfectly at home using the concept. It helped me make sense out of the different styles of two teachers I was observing. Yet the concept itself cannot be rigorously defined and relies upon very subjective interpretation of phenomena. The construct of "withitness," like many of the concepts we work with, is useful, but inadequately defined.

One way to increase the preciseness of our concepts is to tie them through clear operations to the measurement of their occurrence. For example, we can take a concept like teacher warmth, and define it as the number of times per day the teacher smiles. But is that what we want to measure when we measure warmth? It seems that the phenomena we are interested in is fragmented beyond recognition when we use the occurrence of some molecular behavior to operationally define our terms.

What we need to do in the study of teaching is to being incorporating multiple methods of measurement into the studies we do (Campbell and Fiske, 1959). If we want to work with the concept of "withitness" or "warmth," we need to measure the concept from as many different perspectives as we can. For example, we should measure a teachers warmth by self-report, student report, observer rating, frequency count of smiles, percent of gestures regarded as affectionate, and anything else we can think of. Then, from the intercorrelations of the various imprecise and imperfect measures of warmth, we can begin to understand the construct we so glibly use, but cannot clearly define. Extensive construct validation must take place or the impreciseness of our language for describing the phenomena we are interested in will keep the empirical study of teaching at its present primitive level.

The Generalizability of Measures of Effectiveness

If we are going to try to characterize teachers as more or less effective; in order to see if the behavior of those teachers differ, we need to know if the teachers themselves maintain their rank ordering on measures of effectiveness over time and over subject matter areas. As part of our research, we reviewed studies that addressed this problem. There are about eight studies of teacher effectiveness over lengthy periods of time (see Shavelson and Dempsey, 1975). The mean of these correlations between teacher effectiveness measured two or more times is about .30. This is based on data from predominantly primary age children tested with standardized reading and mathematics achievement tests. Brophy's (1973) study presents some interesting data to consider. Residual gain scores over 3 years were examined for 165 elementary teachers. Twenty-eight percent of the teachers were consistent in their effects on students three years in a row. Approximately 14 percent of the teachers in the study were consistently effective in producing higher than predicted reading and math achievement. And 14 percent of the

teachers were consistent in being associated with classes that had scores lower than predicted in reading and mathematics three years in a row. Thirteen percent of the teachers showed linear increases in residual gains over the three years. That is, they appeared to be getting more effective in their teaching. Similarly, 11 percent of the teachers showed a linear decrease over that time period. They seemed to be getting less effective over time. Forty-nine percent of the teachers in this sample were inconsistent in the patterning of their residual scores over time.

In our review of short term studies of teacher effectiveness, ranging across grade levels and all kinds of curriculum areas, we find that when the same content is taught to similar students (for example, teaching and reteaching an ecology lesson to two samples of urban students), moderately stable estimates of teacher effectiveness are obtained. But when different content is taught to two or more groups of similar students, the effectiveness measures were not found to be stable. Similarly, when different content is taught to the same students, estimates of effectiveness from occasion to occasion are unstable. Our own research, just completed, involved about 200 elementary school teachers, each of which taught a two-week, specially designed teaching unit in reading and mathematics. Residual gain scores for each subject matter were calculated. These measures of effectiveness using different content and the same students were correlated. From these data we find that measures of effectiveness in the two curriculum areas correlate about .30.

It appears that teachers do not, by and large, remain in a stable ordering on measures of teacher effectiveness. If, as we have discussed, the independent variables we typically look at are often unstable, and measures of teacher effectiveness also show instability, the possibility of correlating teacher behavior with student achievement to determine effective teaching behavior is quite limited. In fact, unless we reconceptualize much of what we do in this research area, it is ludicrous!

STATISTICAL PROBLEMS*

We have examined instrumentation and methodological problems, and turn now to a brief discussion of the statistical problems associated with the study of teacher effectiveness. The strategy we use in our research is to identify groups of teachers that differ in effectiveness and then to analyze the teaching behavior of the teachers in the contrasting groups. Our choice of statistical techniques is limited to those that apply when a single achievement test is administered to students prior to, and following, some teaching; and the teaching is considered an intervention that takes place with students who were not randomly assigned to classes. Under these conditions a statistical method is required to discriminate between groups of teachers that differ significantly in average pupil gain. The basic problem is one

*Robert W. Heath and Richard Marliave, performed the analyses that addressed the problems discussed in this section of the paper.

addressed over and over in educational research. How do you measure change without a true experimental design?

We have examined the whole range of statistical techniques based on regression approaches. We looked at the advantages and disadvantages of residualized raw scores, residualized true scores, curvilinear adjustments and methods that correct for non-homoscedastic bivariate distributions. We have also examined ways to define effectiveness based simply on post test raw score differences, for classes that had similar pre test scores. And we find much to recommend this simplest of methods, which avoids all pretense of sophisticated statistics. We have also found interesting possibilities in the new scaling methods, which avoid many of the assumptions of classical test theory. Groups of teachers that maximally differ from each other can be identified with these techniques, providing samples or more and less effective teachers within curriculum areas.

CONCLUSION

I stated above that the heart of performance and competency based teacher education, evaluation and accountability programs is the establishment of empirical relationships between teacher behavior as an independent variable and student achievement as a dependent variable. But before we can adequately establish those relationships we need to deal with the problems of instrumentation, methodology and statistics. We must come to grips with the inadequacy of standardized tests, the unknown predictive validity of tests from special teaching units, the problem of building multivariate outcome measures, the problems of measurement of appropriateness of teacher behavior, the lack of experience in choosing an appropriate unit of analysis for describing teaching behavior, and the lack of stability of many teacher behaviors.

We need time to consider the problems of how student background affects measures of teacher effectiveness, what subject matters should be examined, how normative standards and volunteer teachers affect what we can say about teachers and teaching, how individual students react to teaching skills, how students monitor and interpret a teacher's behavior in ways which may or may not coincide with how educational theorists interpret the phenomena, and we need time and resources to do construct validation and studies of the generalizability of measures of teacher effectiveness.

Finally, we need guidance on what techniques to use for measuring changes in the achievement of students in natural classrooms.

When we have finished examining this potpourri of problems, issues, and concerns, we will be ready to begin the scientific study of teaching. And if we cannot deal with all of these problems, perhaps we should simply acknowledge that teaching is, after all, a very complex set of events which cannot be easily understood.

REFERENCES

Barker, R. G. (1968) Ecological Psychology: Concepts and Methods for Studying the Environment of Human Behavior. Stanford, California: Stanford University Press.

Berliner, D. C. and L. S. Cahen (1973) Trait-Treatment Interactions and Learning. In F. N. Kerlinger (Ed.), Review of Research in Education, 1. Itaska, Illinois: F. E. Peacock Publishers.

Berliner, D. C. and B. A. Ward (1974) Proposal for Phase III Beginning Teacher Evaluation Study. San Francisco, California: Far West Laboratory for Educational Research and Development.

Borg, W. R. (1972) The Minicourse as a Vehicle for Changing Teacher Behavior: A Three-Year Follow-up. Journal of Educational Psychology, 63, 572-579.

Brophy, J. E. (1973) Stability of Teacher Effectiveness. American Educational Research Journal, 10, 245-252.

Campbell, D. T. and D. W. Fiske (1959) Convergent and Discriminant Validation by the Multi-Trait-Multimethod Matrix. Psychological Bulletin, 56, 81-105.

Campbell, J. R. (1972) A Longitudinal Study in the Stability of Teachers' Verbal Behavior. Science Education, 56, (1), 89-96.

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

Gall, M. D. (1973) The Problems of "Student Achievement" in Research on Teacher Effects. San Francisco, California: Far West Laboratory for Educational Research and Development (Report A73-2).

Gump, P. V. (1967) The Classroom Behavior Setting: Its Nature and Relation to Student Behavior. U. S. Office of Education, Department of Health, Education and Welfare. Final Report, Project No. 5-0334, Contract No. OE-4-10-107. Lawrence, Kansas: University of Kansas (mimeo).

Harnischfeger, A. and D. E. Wiley (1975) Teaching-Learning Processes in Elementary School: A Synoptic View. Beginning Teacher Evaluation Study, Technical Report No. 75-3-1. San Francisco, California: Far West Laboratory for Educational Research and Development (mimeo).

Heath, R. W. and M. A. Nielson (1974) Performance-Based Teacher Education. Review of Educational Research, 44, 463-484.

Jenks, C., et al. (1972) Inequality: A Reassessment of Family and Schooling in America. New York: Basic Books.

Joyce, B. R. (1975) Vehicles for Controlling Content in the Study of Teaching. Paper given at the meeting of the American Educational Research Association, Washington, D.C., April, 1975.

Kounin, J. S. (1970) Discipline and Group Management in Classrooms. New York: Holt, Rinehart and Winston.

Moon, T. C. (1969) A Study of Verbal Behavior Patterns in Primary Grade Classrooms During Science Activities. Unpublished doctoral dissertation, Michigan State University.

Moon, T. C. (1971) A Study of Verbal Behavior Patterns in Primary Grade Classrooms During Science Activities. Journal of Research in Science Teaching, 8, 171-177.

Mosteller, F. and D. P. Moynihon (1972) On Equality of Educational Opportunity. New York: Vintage Books.

Popham, W. J. (1971) Performance Tests of Teaching Proficiency: Rationale, Development, and Validation. American Educational Research Journal, 8, 105-117.

Postlethwaite, T. N. (1973) A Selection From the Overall Findings of the IEA Study in Science, Reading Comprehension, Literature, French as a Foreign Language, English as a Foreign Language, and Civic Education. Paris, France: International Institute for Educational Planning (mimeo report no. IIEP/STU/MISC/73.3 [Rev. 1]).

Shanker, A. (1974) Competency-Based Teacher Training and Certification: Acceptable and Unacceptable Models. QUEST Consortium Yearbook. Washington, D.C.: American Federation of Teachers.

Shavelson, R. S. and N. Dempsey (1975) Generalizability of Measures of Teacher Effectiveness and Teaching Process. San Francisco, California: Beginning Teacher Evaluation Study, Technical Report #2, Far West Laboratory for Educational Research and Development.

Snow, R. C. (1974) Designs for Research on Teaching. Review of Educational Research, 44, 265-291.

Trinchero, R. L. (1974) Three Technical Skills of Teaching: Their Stability and Effect on Pupil Attitudes and Achievement. Unpublished doctoral dissertation, Stanford University.

Wallen, N. E. (1969) Sausalito Teacher Education Project. Annual Report. San Francisco, California: Division of Compensatory Education, Bureau of Professional Development, San Francisco State College (mimeo).

Wiley, D. E. (1973) Another Hour, Another Day: Quantity of Schooling, A Potent Path for Policy. Studies of Educative Processes, No. 3, Chicago, Illinois: University of Chicago, July 1973.

Winne, P. H. (1974) Teacher Effectiveness and Student Perceptions of Teacher Cues. Dissertation proposal, Stanford School of Education, (mimeo).