DOCUMENT RESUME

ED 111 810 95 SP 009 503

AUTHOR Gage, N. L., Ed.
TITLE NIE Conference on Studies in Teaching; Panel 9, Research Methodology.
INSTITUTION National Inst. of Education (DHEW), Washington, D.C.
PUB DATE May 75
NOTE 47p.; For related documents, see SP 009 494-504

EDRS PRICE MF-$0.76 HC-$1.95 Plus Postage
DESCRIPTORS *Educational Research; Measurement Techniques; *Research Design; *Research Methodology; *Research Problems; Teaching

ABSTRACT

This panel's goal was to improve the validity and utility of measurement, design, and analysis in research on teaching through the stimulation of new methodological knowledge and through the identification and translation of useful existing knowledge from other descriptions. This panel tried to identify as many methodological problems as possible which limit the productivity of research on teaching, and then adopted four "approaches" which it believed to be solutions that encompass all the methodological problems of research on teaching. These four approaches were (1) to develop and test new analysis and design strategies appropriate for research on teaching; (2) to increase understanding of existing measurement strategies for research on teaching and, where appropriate, develop new measurement strategies; (3) to identify, demonstrate, and disseminate methodologies from other research disciplines which appear to have merit for research on teaching; and (4) to consider the utility of standards for improving methodological practice in research on teaching. (BD)

NIE CONFERENCE ON STUDIES IN TEACHING

PANEL 9

RESEARCH METHODOLOGY

GOAL STATEMENT

To improve the validity and utility of measurement, design, and analysis in research on teaching through the stimulation of new methodological knowledge and through the identification and translation of useful existing knowledge from other descriptions.

PARTICIPANTS

Dr. Andrew Porter (Chairperson), Department of Educational Psychology, Michigan State University

Dr. T. Anne Cleary, College Entrance Examination Board, New York

Dr. Chester Harris, Graduate School of Education, University of California at Santa Barbara

Dr. Richard Light, Graduate School of Education, Harvard University

Dr. Donald L. Meyer, Department of Psychology, University of Pittsburgh

Dr. Barak Rosenshine, Bureau of Educational Research, College of Education, University of Illinois

Dr. Marshall Smith, Graduate School of Education, Harvard University

Dr. Susan B. Stodolsky, Department of Education, University of Chicago

Ms. Linda Glendening (Secretary), Office of Research Consultation, Michigan State University

CONSULTANTS

Dr. Albert Beaton, Educational Testing Service, Princeton

Dr. Lawrence Hubert, Department of Educational Psychology, University of Wisconsin

Dr. Maryellen McSweeney, Department of Educational Psychology, Michigan State University

Dr. Melvin Novick, Department of Educational Psychology, University of Iowa

Dr. Niel Timm, Department of Educational Research, University of Pittsburgh

Editorial comments were received from Dr. Daniel Antonoplos, Dr. Jane David, and Mr. Carlyle Maw.

Washington, D. C.

May, 1975

N. L. Gage, Editor
Kent Viehoever, Coordinating Editor

TABLE OF CONTENTS

TABLE OF CONTENTS
(continued)

PREFACE

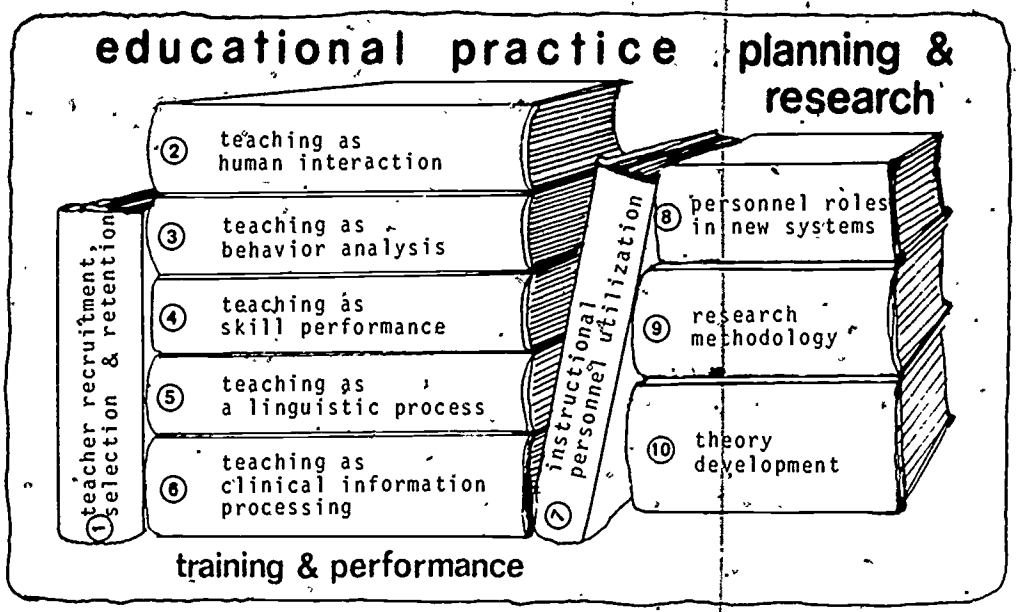
The volume before you is the report of one of ten panels that participated in a five-day conference in Washington during the summer of 1974. The primary objective of this Conference was to provide an agenda for further research and development to guide the Institute in its planning and funding over the next several years. Both by the involvement of some 100 respected practitioners, administrators, and researchers as panelists, and by the public debate and criticism of the panel reports, the Institute aims to create a major role for the practitioner and research communities in determining the direction of government funding.

The Conference itself is seen as only an event in the middle of the process. In many months of preparation for the Conference, the staff met with a number of groups--students, teachers, administrators, etc.--to develop coherent problem statements which served as a charge to the panelists. Panel chairmen and others met both before and after the Conference. Several other panelists were commissioned to pull together the major themes and recommendations that kept recurring in different panels (being reported in a separate Conference Summary Report). Reports are being distributed to practitioner and research communities. The Institute encourages other interest groups to debate and critique relevant panel reports from their own perspectives.

The Conference rationale stems from the frank acknowledgment that much of the funding for educational research and development projects has not been coordinated and sequenced in such a way as to avoid undue duplicat[illegible] yet fill significant gaps, or in such a way as to build a cumulativ[illegible]mpact relevant to educational practice. Nor have an agency's affected constituencies ordinarily had the opportunity for public discussion of funding alternatives and proposed directions prior to the actual allocation of funds. The Conference is thus seen as the first major Federal effort to develop a coordinated research effort in the social sciences, the only comparable efforts being the National Cancer Plan and the National Heart and Lung Institute Plan, which served as models for the present Conference.

As one of the Conference panels points out, education in the United States is moving toward change, whether we do anything about it or not. The outcomes of sound research and development--though enlisting only a minute protion of the education dollar--provide the leverage by which such change can be afforded coherent direction.

In implementing these notions for the area of teaching, the Conference panels were organized around the major points in the career of a teacher: the teacher's recruitment and selection (one panel), training (five panels), and utilization (one panel). In addition, a panel was formed to examine the role of the teacher in new instructional systems. Finally, there were two panels dealing with research methodology and theory development.



Within its specific problem area, each panel refined its goal statement, outlined several "approaches" or overall strategies, identified potential "programs" within each approach, and sketched out illustrative projects so far as this was appropriate and feasible.

Since the brunt of this work was done in concentrated sessions in the space of a few days, the resulting documents are not polished, internally consistent, or exhaustive. They are working papers, and their publication is intended to stimulate debate and refinement. The full list of panel reports is given on the following page. We expect serious and concerned readers of the reports to have suggestions and comments. Such comments, or requests for other panel reports, should be directed to:

Assistant Director
Program on Teaching and Curriculum
National Institute of Education
1200 19th Street, N. W.
Washington, D. C. 20208

As the organizer and overall chairman for the Conference and editor for this series of reports, Professor N. L. Gage of Stanford University richly deserves the appreciation of those in the field of teaching research and development. The panel chairpersons, singly and together, did remarkable jobs with the ambitious charge placed before them: Special acknowledgments are due to Philip Winne of Stanford University and to Arthur Young & Company for coordination and arrangements before, during, and after the Conference. But in sum toto, it is the expert panelists--each of whom made unique contributions in his or her respective area--who must be given credit for making the Conference productive up to the present stage. It is now up to the reader to carry through the refinement that the panelists have placed in your hands.

Garry L. McDaniels
Program on Teaching and Curriculum

LIST OF PANEL REPORTS AND CHAIRPERSONS

1. Teacher Recruitment, Selection, and Retention, Dr. James Deneen, Educational Testing Service
2. Teaching as Human Interaction, Dr. Ned A. Flanders, Far West Laboratory for Educational Research and Development
3. Teaching as Behavior Analysis, Dr. Don Bushell, Jr., University of Kansas
4. Teaching as Skill Performance, Dr. Richard Turner, Indiana University.
5. Teaching as a Linguistic Process in a Cultural Setting, Dr. Courtney Cazden, Harvard University
6. Teaching as Clinical Information Processing, Dr. Lee S. Shulman, Michigan State University
7. Instructional Personnel Utilization, Dean Robert Egbert, University of Nebraska
8. Personnel Roles in New Instructional Systems, Dr. Susan Meyer Markle, University of Illinois
9. Research Methodology, Dr. Andrew Porter, Michigan State University
10. Theory Development, Dr. Richard Snow, Stanford University

- Conference on Studies in Teaching: Summary Report, Dr. N. L. Gage, Stanford University

INTRODUCTION

Statement of Goal

As expressed by Panel 9, the goal for research methodology within the context of research on teaching is:

> To improve the validity and utility of measurement, design, and analysis in research on teaching through the stimulation of new methodological knowledge and through the identification and translation of useful existing knowledge from other disciplines.

The Panel agreed that although much useful research on teaching has been conducted, the value of some of the research has been limited because of methodological problems. In some cases appropriate methodology was not available; in other cases existing best practices were not followed. There have also been cases where methodologies were borrowed from other research disciplines without a careful examination of the assumptions involved.

Thus, the intent of the Panel was to identify as many as possible of the methodological problems which limit the productivity of research on teaching. Because of the breadth of the area considered, the time constraints, and the limited number of panel members, it is likely that important problems are omitted in the present report. Even those problems identified are described with varying degrees of specificity. It is hoped, therefore, that this document will stimulate productive written criticisms as to the relevance of the problems identified, the adequacy of the descriptions of those problems, and the identification of important problems that were omitted.

Issues and Dimensions of the Panel's Work

One of the first concerns of the Panel was that of how to identify and discuss potential solutions to methodological problems without the context of a specific research project. One suggestion was to identify methodological problems that appear to cut across much of the research on teaching. Some of these were relatively easy to identify from the numerous reviews of literature that have been critical of past research.

Another suggestion was to use categories of research on teaching as the contexts for discussion. Some dimensions useful for categorizing research on teaching are:

Types of variables, e.g.,
- Variables antecedent to learning situations
- Variables which describe the process of the teaching/learning situation
- Contextual variables
- Variables which describe outcomes of the learning situation

Types of learning environments, e.g.,
- One-on-one tutorial
- Structured classroom
- Open classroom

Participants to be measured, e.g.,
- Students
- Teachers
- Parents

The Panel concluded that "generic" methodological issues would serve as a starting point, but that all discussions would refer in general terms to the above dimensions. In addition, progress reports from other panels and discussions with members of other panels were used as vehicles for insuring that the methodological concerns were relevant to the needs of research on teaching.

General Discussion of Approaches

The Panel agreed upon four general Approaches for achieving its stated goal:

Approach 9.1 Develop and test new design and analysis strategies appropriate for research on teaching.

Approach 9.2 Increase understanding of existing measurement strategies for research on teaching and where appropriate develop new measurement strategies.

Approach 9.3 Identify, demonstrate, and disseminate methodologies from other research disciplines which appear to have merit for research on teaching.

Approach 9.4 Consider the utility of standards for improving methodological practice in research on teaching.

These four Approaches were adopted as a set believed to encompass all the methodological problems of research on teaching and are, therefore, necessarily broad. The first two Approaches emphasize the need for new methodological developments which specifically address the needs of research on teaching. These two Approaches consider problems of design and analysis (Approach 9.1) and measurement (Approach 9.2), respectively. Together, they cover the full range of new methodological developments.

Approach 9.3 is based on the recognition that existing methodologies developed in other research disciplines may be relevant to research on teaching, but are as yet untried in that context. Finally, Approach 9.4 is a response to the criticism that some research on teaching has suffered from a failure to use the best existing methodology. It was suggested to the Panel that a statement on standards of methodological practice for research on teaching would be useful in alleviating this problem.

APPROACH 9.1

DEVELOP AND TEST NEW DESIGN AND ANALYSIS STRATEGIES APPROPRIATE FOR RESEARCH ON TEACHING

The development of principles for the design and analysis of studies has a long history, much of it stimulated by problems of research in specific fields. For example, during the early and middle parts of the twentieth century, problems of analysis of agricultural data played an important role in the development of techniques commonly used today. Generally, there has been less input to this literature from education and teaching than from agriculture, the biological sciences, etc. Even today, we see many new developments coming from areas other than education. For example, analysis of covariance and index of response methodology have come primarily from agricultural problems. The Panel felt that it is time for more systematic efforts toward the development of principles for the design and analysis of studies within the special and possibly unique context of problems of education, generally, and the study of teaching, in particular.

Perhaps the major impression left by reviews of current research on teaching is that problems of design and analysis are encountered at many stages, and are solved--if at all--in an imitative or derivative fashion drawing on analogies with earlier studies, especially those in agriculture. The current need is to treat seriously the unique problems posed by attempts to describe and relate processes of teaching to types of outcomes of teaching. To do so, serious attention will have to be paid to many problems of measurement (which are considered in a separate approach) and to the development of new design and analysis procedures.

Much is known, especially at the theoretical level, about characteristics of various design and analysis procedures. What is missing, however, is more detailed knowledge of specific applications to research on teaching and of the limitations of the usefulness of the procedures within that research context. In general, it is understood that a doctrine of specificity applies to problems of design and that this doctrine requires that designs be developed for particular situations and inquiries. It is true, however, that at least rough categories of types of applications can be developed and used as guides.

As might be expected, several of the programs within Approach 9.1 reflect the ongoing debate about designs and analyses useful for investigating causal relationships. Clearly the most convincing evidence comes from designs which include variable manipulations controlled by the experimenter. And for much of the research on teaching it was generally agreed that arguments for causal relations are strengthened when random

assignment of subjects to levels of an independent variable is accomplished. Still, history suggests that such designs are difficult to implement, particularly when the subjects are people. A better understanding of how to implement such designs is needed.

Nevertheless, researchers will continue in their attempts to "tease out" causal relationships from correlational data. When cautiously interpreted, the results from correlational designs can be useful. However, analytical models that support such efforts are not as yet fully understood and for some designs more useful models may be developed.

Several other programs within this Approach reflect the Panel's concern with the interpretation and generalization of results. For example, one program was concerned with the problem of introducing explicitly into both design and analysis the use of prior and collateral information about the context and participants of a study--information which can, if successfully used, yield more efficiently designed studies. Another program was aimed at the development of methods for making research on teaching a cumulative enterprise. As Light and Smith (1971) have observed, significant knowledge in the social sciences accrues ever too slowly. A major reason is that various research studies on a particular question tend to be of dissimilar designs, making their results difficult to compare. An even more important factor is that social science studies frequently produce conflicting results, which hinder theoretical developments and confuse those responsible for the implementation of social policies. At a minimum, what is needed are (a) criteria for determining when data from dissimilar studies can be pooled, and (b) methods for recognizing fundamental differences in research design, and avoiding the creation of artificial differences.

This Approach is intimately related to Approach 9.2, which is aimed at increasing the understanding of existing measurement strategies for research on teaching and, where appropriate, developing new measurement strategies. In addition, this Approach receives direction from the problem areas of all other panels in the Conference. For example, problems of selection (Panel 1) involve estimation of statistical relations; problems of conceptualizing and observing teaching (Panels 2-6) involve sampling; and problems of theory development (Panel 10) involve consideration of the roles of data.

Program 9.1.1: Analysis Problems Related to Hierarchically-Nested Data

Much of the data in educational research is hierarchically nested (Porter, 1973). For example, students are nested within classrooms which are, in turn, nested within schools. Such hierarchical nestings give rise to a variety of methodological problems.

Project 9.1.1.1: Models for Estimating Relations among Variables at a Lower Level of Aggregation. Given data on a set of aggregate units, what models are useful in the estimation of relations among variables for subunits (Iverson, 1974; Robinson, 1950)?

Project 9.1.1.2: Models for Data Aggregation. How should aggregation proceed when measurements are taken on several variables for units at one level, but the researcher wishes to aggregate both across variables and across units to a higher level unit? If one first aggregates across variables and then across units, results can (and frequently will) differ from those obtained when aggregation across units precedes aggregation across variables. Are there contexts and purposes when one order of aggregation is more useful than another?

Project 9.1.1.3: Analysis of Unbalanced Designs. What methods are most useful for analyzing data from unbalanced hierarchically-nested designs?

Project 9.1.1.4: Consequences of Violating Assumptions of Independence. What are the consequences for various interval estimation procedures of violating the assumption of independence because of an incorrect choice of the unit of analysis (not appropriately specifying the aggregate units in the analysis model)?

Project 9.1.1.5: Analysis of Non-Independent Student Data. Many instructional situations apply a "treatment" to a class of individual students. The classic methods of analyzing an experiment for comparing different treatments can be used with the classroom as the unit of analysis, and the conventional probability statements can be meaningful when it is possible to assign treatments at random to classrooms. Although the students have not been treated independently, their individual scores can contain useful information. What analyses are possible to utilize this information? What models and assumptions would be necessary to permit a valid analysis using individual scores?

Program 9.1.2: The Utility of and Methods for Conducting "True Experiments" in Research on Teaching

There was a strong consensus within the Panel that to understand the effect of an aspect of teaching it is necessary to manipulate that aspect. This requires an active role on the part of the researcher which might best be accomplished by randomly assigning participants to conditions of interest (Campbell, 1971). Although variable-manipulation studies are frequently labelled experiments, the word experiment is also used more broadly. Because of the importance of variable manipulation to the future productivity of research on teaching, the Panel recommends clearer language in the research literature. Therefore, the Panel recommends the adoption of standard terminology which communicates clearly that a study has manipulated the variable of major interest through random assignment.

Project 9.1.2.1: Use of Incentives for Participation in "True Experiments." This project would examine the use of incentives to encourage participation in variable-manipulation investigations for research on teaching.

Project 9.1.2.2: Ethical Issues in Conducting "True Experiments." This project would consider ethical issues where random assignment can infringe upon the rights of participants in an experiment:

1. Denial or temporary deferral of treatment to persons in need of it as a consequence of the use of random assignment;

2. Compromising the participant's right of informed consent to participate or not.

Project 9.1.2.3: "True Experiments" within Quasi-Experiments. This project would examine alternative procedures for embedding small randomized studies within large ongoing nonrandomized studies. Campbell and Stanley (1963) have considered some possibilities, but more work seems to be needed.

Program 9.1.3: Data Analysis Procedures for Quasi-Experimental or Correlational Studies

Much research on teaching consists of selecting a number of classrooms, testing the students on some criterion variable before and after instruction, and relating those scores to the type of instruction. Information from this research strategy may be useful for understanding the instructional process and for suggesting hypotheses for "true experimental" research. Two problems, however, are evident: (a) How should the data be analyzed; and (b) What is the utility of the results?

The confusion about methods of analysis stems from at least two concerns. First, since pupils have not been assigned to classes at random, the posttest scores are usually adjusted for pretest scores on one or more measures. Historically, several methods of adjustment have been used. One method adjusts on a separate within-class regression equation for each class. This method is not as restrictive as some in terms of assumptions, but it ignores the collateral information available from similar classes. Another method uses a pooled within-class regression line for adjustment. A third method ignores the individual scores and merely uses mean posttest scores and pretest scores across classes.

Second, other aspects of data from such studies are often ignored. Two examples are (a) the possibility that teaching strategies may affect the slopes of the within-class regression lines themselves; and (b) the possibility that the performance of a class may be affected by the distribution of pretest scores.

It is evident that different analyses reflect different conceptualizations and models. The confusion over which analysis is "best" stems from a lack of making explicit the underlying model and relating it to the purpose of the study. A full explication of models and analyses appropriate for studies of teaching of this type is needed.

Finally, it must be realized that quasi-experimental studies of this type are not as useful for inferring direct cause and effect as "true experiments," but they can suggest models which may be useful in understanding the teaching-learning process. Still, the relative utility of the two approaches (quasi-experiments and true experiments) is not well understood or agreed upon.

Project 9.1.3.1: Adjusting on Multiple Fallible Covariables. Researchers often wish to adjust outcome variables for differences across conditions on some set of antecedent variables. A specific example is found in attempts to assess teaching performance in terms of student outcomes. Such adjustment is of interest because students are typically not randomly assigned to teachers. One model for making adjustments is to use the structural relations refined on the latent true variables. Cochran (1968) has provided useful statements about the relationships between least square estimates of regression coefficients and the coefficients defined on the underlying latent true variables. Econometricians

have provided a variety of methods for estimating the structural relations given fallibly-measured variables. At least one useful solution exists for a single fallible covariable (Porter & Chibucos, 1974). What remains to be done is to incorporate the knowledge about estimating structural relations of multiple fallible covariables to predict one or more dependent variables with the subsequent analysis of adjusted outcomes.

Program 9.1.4: Development and Exploration of Formal Models for Incorporating Information about the Extent of Implementation of Teaching Strategies into the Evaluation of Those Strategies in Terms of Outcomes

A critical question in the evaluation of teaching strategies is the extent to which the strategies were implemented by the teachers. Clearly, a strategy can look ineffective simply because it was not used, yet this possibility may be overlooked in an evaluation that concentrates on student outcomes. The problem of measuring implementation is addressed later in Program 9.2.8. Program 9.1.4 focuses on how to formally incorporate implementation data into the evaluation of strategies in terms of outcomes.

Analytic models that can be used to predict outcomes for a variety of levels of implementation are needed. Such models would help researchers unconfound the effect of level of implementation from the effect of the strategy given full implementation. For example, if Strategy A at the observed level of implementation has better outcomes than Strategy B across all levels of implementation, the conclusions are clear. If a more fully implemented Strategy B might exceed Strategy A in outcomes, however, then the researcher might want to concentrate on methods for improving the implementation of Strategy B.

Program 9.1.5: Investigation of the Utility of Longitudinal (Time-Series) Designs for Various Types of Research on Teaching and Concomitant Analytic Problems

Longitudinal data-collection efforts are sometimes held as a panacea for research on teaching. Although this is not likely to be the case, the question remains: For what research questions are longitudinal designs necessary? In addition, a variety of problems with the analysis of longitudinal designs appear to require further work, e.g., changing metrics of the dependent variables over time, unevenly spaced time points, and methods for collapsing data across time points. Glass (1972) and Anderson (1971) have done recent work on some related problems.

Program 9.1.6: Empirical Selection of Models of the Teacher-Student Interaction Process

If researchers have difficulty specifying an underlying model of the teacher-student interaction process, what sorts of statistical procedures can be used to choose among several competing models? More specifically, what are some useful alternatives to the least squares criterion?

Program 9.1.7: Procedures for Combining the Results of Related Studies over Time

How can studies of teaching formally build into future designs the results of earlier studies so that future studies can be more effective or powerful?

Program 9.1.8: Procedures for Studies of Teacher Effectiveness

Prior to consideration of the design and analysis procedures for research on teacher effectiveness, a caveat is necessary. Several members of the Panel questioned the utility of such research even given satisfactory methodology. The reasoning was that numerous past efforts have not been productive. Many teacher characteristics have been shown to be unstable over time and those that are stable appear to be unrelated to student outcomes. In addition, studies which simply attempt to establish that teachers do have consistent effects over time do little to help understand the causes of those effects (Rosenshine, 1970; Brophy, 1973; Acland, 1974). The Panel concluded, however, that improved methodology would rule out one rival explanation for the lack of utility of such studies and might, therefore, be of value.

Given the above, the ideal strategy for studying the general question, "Do teachers make a difference?" requires something close to the following design. First, a large number of students would be randomized over N teachers. Then, class means and variances on some index of change, e.g., posttest scores or gain scores, would be compared across the N teachers. This could be done over several years to determine whether there are any consistent outliers. The existence of one or more outliers would imply some structural difference in teacher effectiveness.

The analysis would be performed to examine each teacher's change scores over the several years. Any teacher whose change scores consistently fell above the average for all teachers would represent a positive teacher effect. This is similarly true for negative effects. A hodgepodge of above and below the mean results for most teachers would indicate a lack of teacher differences. The intraclass correlation could be used to detect consistent differences in teacher performance (Veldman & Brophy, 1974).

Project 9.1.8.1: Problems Due to Lack of Random Assignment. Suppose that because of political or administrative realities, large-scale random assignment of children across many teachers for several years is not possible. The question then emerges: How can the broad program goal of searching for consistent teacher effects be examined? This goal creates a need for some kind of sensible "adjustment" to determine the change scores achieved by each teacher in each year.

How should these adjustments be made? The answer is not obvious. For example, one possibility would be to run a grand regression equation using all the pretest-posttest scores for any given year. Then, for each teacher, a residual (the observed minus predicted) final score could be obtained. But this involves implicit assumptions about learning curves. What precisely are these assumptions and are they reasonable? This question is similar to that posed in Program 9.1.3.

Now, suppose this method of obtaining residuals was applied over several years, fitting a new grand regression equation each year, and computing each of the N teachers' residuals from the new regression each year. This process would lead to a set of M residuals for each teacher over M years. Once again, the intraclass correlation would be useful to determine whether consistent differences in teacher performance can be detected. A high, positive correlation implies strong, consistent differences in teacher effects.

Project 9.1.8.2: Following Students over Time. The description of the program for searching for teacher effects has so far considered each year's change score within each of the N classrooms as an independent entity. This approach ignores an important question: If a group of students for one teacher has a mean change score in Year 1 that far exceeds the mean change score over all N teachers, what happens to those students in the following years? Do they maintain their lead? Do they increase it? Or does that lead dissipate? Answering this question would require following students over time. One design would be to keep each group of students in any class in Year 1 together for several subsequent years. This would tend to preserve any contextual effects of students interacting positively with one another. A second strategy would be to break up the classes from one year to the next. If this breaking up were done randomly, new information could be developed about other teachers in future years. Several questions must be dealt with here, and a thoughtful consideration of the implications of alternative designs would be useful.

Project 9.1.8.3: Procedures for Combining Several Intraclass Correlations into a Single Estimate. Assume the earlier projects in this program have been completed; i.e., assume we have available an intraclass correlation coefficient based on M years of data from N teachers. Then, the value of this coefficient will give information about differential teacher effects. But the correlation coefficient would be coming from a single study, for example, in a single city. Imagine that, because of interest in getting good multisite (multicity or multischool) data, a similar study is conducted in each of R cities. This gives us R intraclass correlation coefficients that may well be based on different sample sizes. What is the most effective way of combining the set of R intraclass correlation coefficients into an overall estimate (Votaw, 1948; Olkin, 1965)?

There are at least three alternative ways of combining the data from the R studies. First, the raw data could be pooled. Second, a median of all the intraclass correlations could be computed. Third, Fisher's Z transformation, which is simply a function of the correlations and their associated sample sizes, could be used. Is one procedure always preferable to the others?

Probably, each procedure has a setting in which it is most effective. A reasonable guess is that the most effective procedure depends upon an assumption about the form of the population of correlation coefficients that arise from different sites. For example, if one assumes that all sites have a true underlying coefficient and that this coefficient is an identical parameter over all R sites, one method may be best. A second circumstance involves assuming some distribution of true coefficients over the R sites. Then, the best way of combining the R observed coefficients may well depend upon the distribution of true coefficients. If so, what procedures are useful for describing the distribution? A final case would be that in which researchers develop a series of R estimated coefficients and we have a modest prior probability that several of them are outliers. In this event, depending upon our prior estimate of both the probability of an outlier and also its estimated magnitude, we would probably want to weigh outlying observations less than coefficients clustering around a measure of central tendency.

Program 9.1.9: A National Study of Current Educational Practice Analyzed at the Behavior-Setting or Organization-of-Instruction Level

Research on teaching has been conducted in a wide variety of settings and types of classrooms and schools. An important question about most research on teaching is the extent to which the conditions studied are representative of a larger population. In reading and conducting research on teaching and instruction, therefore, it would be useful to have knowledge of the distribution of basic types of educational practice. One could be interested in such an issue at many different levels: district organization, subject matter coverage, etc. In this program, however, interest centers on the organization of classroom settings, i.e., the instructional organization (Gump, 1967).

Many researchers have intuitive hunches about the distribution of instructional organization, that is, about how typical or atypical a particular situation is. But data which speak directly and systematically to this descriptive end are not available. For example, how frequently does recitation as an instructional setting occur in high schools? In elementary schools? How often does free choice of activity or individual work occur in high schools? In elementary schools?

Knowledge about instructional organization is important because it relates to the behavioral options of teachers and students; behavior setting structure has been shown to be systematically related to philosophical curricular differences (Grannis, 1973). If one had knowledge of the instructional organizations of a representative sample of schools, generalizations about teaching procedures could be more systematically related to other factors, for instance, to interpreting evaluation outcomes.

In addition to its immediate purposes, such a study would facilitate systematic sampling plans, policy decisions, and historical research, particularly if it could be efficiently collected periodically. Such a survey would provide a convenient way of getting evidence about educational innovation and change.

Project 9.1.9.1: Development of Behavior-Setting Types. There is need to develop an inclusive set of behavior-setting types or instructional types for use in future studies. These typologies could be based on a small empirical study of classrooms, and reviews of literature and concepts (Gump, 1967; Grannis, 1973).

Project 9.1.9.2: Economical Ways of Acquiring Information on Behavior Settings. In the past, behavior settings have been studied by direct observation, which is costly. There is a need to compare the validity and reliability of behavior-setting (type of instructional organization) information obtained via teacher questionnaire and direct observation of classrooms. The aim would be to develop and validate an economical means of obtaining reliable and valid information for a national study.

Project 9.1.9.3: National Survey of Classroom Behavior Settings. The objective of this project would be to conduct a national study of classrooms at various educational levels to ascertain the distribution of various instructional organizations within and across schools, districts, etc. The survey would utilize the strategies developed in Projects 9.1.9.1 and 9.1.9.2. Grade levels, subject matter, etc., should be included as relevant information.

APPROACH 9.2

INCREASE UNDERSTANDING OF EXISTING MEASUREMENT STRATEGIES FOR RESEARCH ON TEACHING AND, WHERE APPROPRIATE, DEVELOP NEW MEASUREMENT STRATEGIES

There is a long and productive history of psychometrics which has supplied theory and guided test construction for research on teaching. Much of this history, however, relates to concerns with measuring the aptitudes and achievement of individual students. Although this work has been, and will continue to be, of value to research on teaching, other aspects of measurement appear to need greater attention. For example, better measures of so-called noncognitive outcomes of teaching, including personality characteristics, self-perceptions, values, and attitudes are required. There is a need for better theories about such constructs, but development of measures is also constrained by the need for better methodology. A second example is the need for better measures of the teaching process, particularly in natural settings. A third example is the need for group assessment measures as contrasted with measures designed to assess individual differences.

Current and pending legislation has given a sense of urgency to the need for assessing effects of teaching. Thirty-one states are now considering laws requiring all applicants for a teaching license to demonstrate their teaching effectiveness. One example is the Stull Act, effective in 1972 in California, which requires all school districts to evaluate their teachers. Many of these evaluations will be based on student outcomes, yet existing measures of student outcomes are largely restricted to cognitive achievement and aptitudes. Even these measures may not be appropriate since most were designed to distinguish among individuals (students), not groups (classrooms).

The programs within this Approach can be roughly categorized as dealing with concerns for measuring dimensions of the process of teaching or of the outcomes of teaching. There are several motivations for measuring dimensions of the process of teaching. First, knowledge about what actually takes place in a learning situation is useful in stimulating new theories about teaching strategies. Second, much research is devoted to providing teachers with new strategies believed to facilitate student learning. If student outcomes do not reflect the attempts to change teaching strategies, then there are at least two explanations. One is that the strategies were not effective, and the second is that

the strategies were not implemented by the teachers. Better measures of the teaching process are necessary to narrow the alternative explanations.

With respect to measurement of outcomes, there is a need to develop or select measurements which are valid for assessing the effectiveness of an intervention. This need stems from the inappropriateness of many current and widely used standardized achievement tests. These measures are inappropriate for assessing teacher (and curriculum) effects for a variety of reasons:

1. They were not designed to measure the outcomes of interventions.

2. They tend to measure relatively stable characteristics.

3. Functionally, the major purpose of these tests is the sorting and selection of individual students.

Recent efforts have been at least partially responsive to the above outlined measurement needs. First, numerous classroom observation instruments have been developed to measure the teaching process, and some useful data banks describing classroom activities are now available, e.g., the SRI Follow Through classroom observations. Nevertheless, the properties of existing observation schedules are generally not well understood, and problems of validity and reliability remain. Second, the recent surge in the development and use of criterion-referenced measures should alleviate some of the concerns about existing achievement measures. Still, most of the work is concentrated on assessing individual student performance, while one of the major needs for research on teaching is to assess the impact of interventions.

As stated previously, this Approach is related to Approach 9.1 to develop and test new design and analysis strategies. Clearly, the reliability and validity of measures can limit the utility of a research study. Design and analysis strategies must be sensitive to the weaknesses of the measures, but they cannot turn useless data into useful data. There is some reason to believe from recent literature that concerns for solutions to design and analysis problems have overshadowed concerns for solutions to problems of measurement. If so, this imbalance should be corrected.

Program 9.2.1: Educational Significance of an "Effect"

Historically, the issue of what an instrument measures has been approached from two points of view: (a) the content of the instrument (face or content validity) and (b) the interrelationships between the instrument and other variables (predictive, concurrent, or construct validity). Typically, these points of view have not been used differently for various types of measuring instruments, e.g., for norm vs. criterion-referenced tests or multidimensional vs. single-trait tests. Though it is not clear whether such a differentiation should be made, it seems reasonable to think about the conditions under which the two approaches are most useful.

The two validity approaches are similar in that both beg the issue of causality. They differ, however, in that correlation studies of the interrelationships between the instrument and other measured behaviors depend upon existing distributions of scores on all variables considered. This last point involves the distinction between an "effect" defined as a difference between two points on a natural scale and a standardized measure of the effect. Standard deviation units, correlations, and percentages of variance explained are examples of standardized metrics which have been used to define the educational significance of an "effect."

The first purpose of this program is to suggest strategies for assigning meaning to measurement--strategies which are independent of the original distribution of the measurements (Porter & McDaniels, 1974). A secondary purpose is to attempt to give meaning to the "impact" of an intervention through the mechanism of giving meaning to the particular measures used to assess the outcomes of the intervention. In a sense then, the function of this program is to move the field from defining "educational significance" of an effect as, say, a one-half standard deviation difference between an "experimental" and a "control" group, the standard used in the Westinghouse-Ohio evaluation of Head Start (1969). It is also intended to move the field away from defining "educational significance" as a statistically "significant" difference. Instead, it is intended that the field begin defining the "educational significance" of an effect in terms of either the measured consequences of the size of the effect for that instrument or the content validity of the instrument and the chosen criterion level.

o projects are suggested. The first is to explore the possibility of d ning the meaning of the size of difference between two points on a nat l scale empirically by estimating the impact of a change from one po to another on a broad range of other possible concurrent and future outcomes. This strategy will be labelled as indirect validation. The second project involves the determination of the meaning of particular criterion levels on instruments and is designed to provide direct understanding of a phenomenon through content validity.

Project 9.2.1.1: Indirect Validation. The "size of an effect" is defined in terms of the raw score difference between two points on a scale. For example, this might translate into the difference between means. What is called for is to give meaning to effects of different sizes by relating those effects to other measured aspects of a person's behavior or experience. Thus, how does a ten-point difference in Binet IQ scores relate to differences in one's chances of attending a college or one's being assigned to a special remedial class or one's future income? Here, meaning would be given to the size of effect through its relationships with other outcomes. In the context of no intervention, this would translate into giving empirical meaning to a particular distance on a particular measuring instrument (difference in scores on IQ tests take meaning from predicted differences on other outcomes). As a start, a limited set of widely used instruments might be studied, e.g., the Stanford-Binet Intelligence Scale and the Metropolitan Achievement Tests. Existing data could be used to attempt to give meaning to the instruments and new data could be suggested where necessary.

The following issues should be considered in carrying out the project:

1. Would use of a standardized measure of effect (such as explained variance, correlation, or standard deviation units) yield the same kinds of conclusions as the indirect validation approach? Under what conditions do these two approaches lead to different conclusions?

2. In the context of giving meaning to the size of an effect of an intervention, does a single score distance translate into different sizes of effect for different contexts and populations?

3. Consider the same problem as 2 for giving meaning to a raw score distance where difference in size of effect may not be attributed to a particular intervention.

4. Consider the problem that "effects" of the same size at different points on a scale may have to be assigned different meanings depending on the context and population. For example, in Boston, the cut-off for assignment of students to special classes is an IQ score of 80. In this situation, an intervention which results in a two-point change in IQ scores has different meaning if the change is from 79 to 81 than if the change is from 104-106.

5. Consider the possibility that two interventions, each raising IQ scores by 10 points (say, from 100 to 110) on a short-term outcome measure, may have very different meanings if the two increases in scores are accompanied by changes in different characteristics and, therefore, by different impacts on other outcomes.

Project 9.2.1.2: Direct Validation. This project would describe existing measuring instruments and particular criterion levels in terms of a theoretically-based understanding of the content of the measuring instruments. The intent of the project is to give meaning to an instrument by describing what the instrument requires of the respondent in terms of knowledge or skills. Thus, the test and criterion level would be used in a theoretical framework to give direct meaning to reaching or failing to reach criterion on the instrument.

The following issues should be considered in carrying out the project:

1. Consider the logic of the test as well as other characteristics. For example, in reading it would be useful to differentiate among the following: (a) labored decoding skill, (b) fluent decoding skill, (c) understanding of the logic, syntax, and internal structure of discourse, and (d) extent to which the respondent shares the concepts and purposes of the test constructor.

2. In the context of interventions, consider the possibility of using this direct validation strategy to assess interventions without reference to comparison groups.

3. For a given instrument, consider the meaning of different criterion levels for (a) a single context and population, and (b) across different contexts and populations. Use existing data where possible and suggest new data where necessary.

Program 9.2.2: Analysis of the Desirable Properties of Tests Stratified by the Purposes of the Tests

What are the desirable properties of tests serving different purposes? Some examples of tests having different purposes are:

Mastery tests -- These lead to dichotomous decisions (Harris, Alkin, & Popham, 1974).

Diagnostic tests -- Should they be multidimensional measures of skills plus measures of other characteristics that influence those skills?

Measures of outcomes -- Should they sample the common core of objectives or sample the multitude of differential objectives?

Program 9.2.3: Construction of Tests with Face Validity

What test construction strategies are most useful in developing measures that have the face validity required by the courts? Given current emphasis on accountability, this concern seems particularly important (Klein, 1971).

Project 9.2.3.1: Development of New Measures That Are Tied to the Purposes of Instruction. For the study of teaching, what is the role of specialized tests designed to be sensitive to different teaching strategies? All too often researchers use general standardized achievement tests that were designed for purposes other than differentiating among teaching strategies and which, therefore, cannot be expected to be sensitive to that end.

Project 9.2.3.2: Development of Measures Dealing with Non-Cognitive Outcomes. In addition to achievement measures, there is need for the development of measures of important non-cognitive outcomes. The construction of these measures should be tied to well-developed theories (Walker, 1974).

Project 9.2.3.3: Development of Measures for Observations of Classroom Process Variables. The development of measures to assess classroom process variables such as time spent on a task holds promise for research on teaching. The rate of progress in this area of research over the past few years suggests a need for a totally new approach. Perhaps greater concern for the relation of process to outcomes would be useful.

Program 9.2.4: Analysis of Crossed Design Achievement Tests

Achievement tests may consist of a set of items that exist in a completely crossed design. The dimensions of such a design might be types of content and types of tasks. The complete set of items then exists in a two-way design with one item per cell; An example of a crossed design achievement test and an item analysis appears in Harris and Harris (1973).

The problem of how to score and analyze such items appears to be one that deserves considerable study. Unidimensional latent trait models are probably inappropriate. What other models need to be developed? To what extent are multi-mode analyses appropriate?

Program 9.2.5: Test Bias

Questions of test bias may be relevant to several aspects of research on teaching. The possibility of test bias (differential predictive validity across subgroups) may be an important consideration in designing systems for the prediction of teacher effectiveness, a concern of Panel 1. Similarly, test bias is an important consideration when achievement measures (mastery, diagnostic, etc.) are used in tracking students.

What are the implications of various definitions of test bias for differential treatment of students and teachers? Several studies (Linn & Wertz, 1971; Schmidt & Hunter, 1974) address this question for the Cleary (1968) definition of test bias, but similar work is required for other definitions, such as those of Thorndike (1971) and Cole (1973).

An additional concern is that almost all test bias studies have been conducted using criteria that may be presumed to be biased to the same degree as the predictors. Can other criteria be developed that are less subject to the same biases? For example, previous research shows that verbal tests predicted success in gunnery classes when the criterion was grades received from the class but not when the criterion was performance measurement. At the college level, test bias studies have used early, i.e., freshmen, performance exclusively. Evidence, although it is not very systematic, is accumulating that suggests that if later performance were used as the criterion, different results would be obtained.

Finally, almost all test bias studies have been conducted at the higher educational levels. There is a great need for this type of research at the lower educational levels. Appropriate criteria, however, must be developed for this research. (See Project 9.3.11.2 for an additional aspect of test bias.)

Program 9.2.6: Evaluation of Profiles

Comparative studies of teaching methods encounter technical problems in the evaluation of profiles of outcomes. Technical characteristics of the measures must be considered in the development of any composite indices of outcomes (Harris, 1955). In addition, problems of weighting the importance of a variable a priori must take into account the differing metrics of the variables.

Program 9.2.7: Defining Desired Teacher Performance

When defining teacher performance for purposes of accountability, is there any agreement among significant groups such as parents, teachers, students, and legislators? What kinds of consistencies can be found in the objectives underlying existing statewide teacher assessment programs? (For work on a related issue, see Hoepfner, Bradley & Doherty, 1974.)

Program 9.2.8: Development of Measurement and Observational Procedures for Describing the Degrees and Types of Implementation of the Components of Various Teaching Processes and Programs

The problem of estimating and describing the degree of implementation of programs and the components of programs is a critical one for research

on teaching. The need for data related to the implementation issue is most salient in two contexts. First, when a program developer, teacher trainer, supervisor, etc., is attempting to train persons to carry out a particular program or type of instruction, he needs to know the extent to which the implementation is occurring. He needs also to be able to analyze the ways in which the program is and is not being implemented. In this context, such information might primarily serve a feedback function. Second, in evaluation studies or any study tying program (treatment) to student outcomes, information on the degree and type of implementation is essential. For example, in the analysis of presumed replications of a given curriculum it is essential to know how comparable the classroom procedures (treatments) really were. In evaluation studies of a comparative nature (as argued in Program 9.1.4), implementation data are even more essential for interpretation and analysis of effects (Stodolsky, 1972; Bissell, 1971).

Project 9.2.8.1: Measuring Implementation. While the methodologies and measurements needed for implementation research may be somewhat program specific, the following general approach might be useful:

- Explore the means for collaboration between curriculum developers and methodologists in order to develop operationalized descriptions of essential components of a curriculum.
- Specify tolerance levels for acceptable or unacceptable levels of implementation.
- Explore means for identifying nonessential or unintended components of a curriculum.
- Repeat the above steps for a few diverse curricula.

In carrying out this approach or a similar approach, the following types of issues should be considered:

1. For what components of programs or types of programs can implementation be assessed without direct observation or with minimal observation?
2. For what components of a program or types of programs is direct observation essential for estimating implementation? (See also Project 9.1.9.2.)
3. How much data are necessary? How much and how frequently should monitoring be done? (The answer will probably vary for different classes of programs.)

Project 9.2.8.2: Stability of Student and Teacher Behaviors. In dealing with the issue of how much data are necessary for implementation studies, an important related issue is the accumulation of knowledge about the stability of student and teacher behavior in general. It would be helpful to have a better empirical basis for estimating the stability of behavior and, therefore, for obtaining guidance as to the frequency and extent of data collection. In addition, empirical data on such matters

would facilitate interpretation of data on classroom phenomena. Thus serious attention should be given to the questions:. How inherently stable are student and teacher behaviors in classroom setting? Under what conditions are the behaviors relatively stable and relatively unstable?

One approach to this question might take the form of studies in which teachers and students are observed intensively over a period of time (say a month). If sufficient data were available, various estimates of stability could be made regarding behaviors of different types and their relations to subject matter, setting, etc. For an example of such data see Karlson (1972).

Program 9.2.9: Studies to Improve the Reliability of Observational Procedures

Even when the stability of the phenomena being studied is known, the reliability of observational procedures can be problematic. When using on-the-spot category systems, the major concern is field reliability, i.e., observer agreement. In this connection, studies to explore the effective training of observers deserve support. While there is some accumulated wisdom on this subject (Gellert, 1955; Weick, 1968), empirical studies comparing the utility of certain alternative procedures for training should be carried out.

In observational studies which use open-ended procedures, e.g., narrative records, there are two types of reliability: (a) field reliability, i.e., agreement of observers in the field; and (b) coding reliability, i.e., reliability of applying coding categories to narratives. These two types of reliability are interdependent. In particular, field reliability cannot be assessed without coding. Exploration of methods for assessing the two types of reliability as well as their interdependence should be supported. Finally, in the case of closed systems, alternative training procedures for field observers should be studied.

More generally, certain technical studies of the utility of various approaches to recording data should be launched. For example, under what conditions does videotape or audiotape recording improve the precision of observations? What are the costs and benefits of various procedures for recording data?

Program 9.2.10: Psychometric Properties of Criterion Referenced Tests and Concomitant Test Construction Strategies

Although the need for criterion referenced tests is apparent, the methodology for developing them is lagging badly behind the aspirations of potential users. Much of classical test theory does not apply. New models need to be developed to deal with such problems as the fidelity of measures to the performances represented, the stability and generalizability of the measures, and the probability of misclassification under various conditions. As theory develops it must be translated into test construction strategies.

APPROACH 9.3

IDENTIFY, DEMONSTRATE, AND DISSEMINATE METHODOLOGIES FROM OTHER RESEARCH DISCIPLINES WHICH APPEAR TO HAVE MERIT FOR RESEARCH ON TEACHING

The first two Approaches reflected a concern for the development of new design, analysis, and measurement techniques that serve the unique needs of research on teaching. Most of the methodology currently used in research on teaching, however, was originally developed in other research disciplines. There are at least two reasons why continued identification, translation, and dissemination of methodologies from other research disciplines seems warranted. First, in many cases, these borrowed methodologies have served research on teaching well. Second, where existing useful methodologies are available, duplication of development should be avoided.

Panel members observed that historically there has been a time lag between the development of methodological and analytic strategies in one discipline, and the use of those strategies in another discipline. During the present period of rapid development of methodologies across a variety of research disciplines, it is becoming increasingly difficult for workers in research on teaching to stay abreast of what is available. At a minimum, Approach 9.3 calls for an awareness of methodological developments in econometrics, sociology, psychology, anthropology, as well as applied and mathematical statistics. These methodological developments need to be screened for their potential utility in research on teaching, and the more promising methodologies should be tried out. As a start, the Panel attempted to identify (in the form of programs) a few methodologies that at least on the surface appeared to have utility for research on teaching.

This Approach is intimately related to the fourth Approach, which calls for considering the utility of standards for improving methodological practice in research on teaching. Both Approaches differ from the first two in that they are designed to improve research on teaching through the use of existing methodologies rather than through the development of new methodologies. The difference between the third and fourth Approach is that the third Approach attempts to capitalize on methodologies virtually unknown to the community of researchers on teaching, while the fourth Approach is concerned with increasing the level of methodological awareness within that research community.

Program 9.3.1: Optimal Designs for Research on Teaching

Evidence on a particular research problem or question usually can be collected in several ways. Unfortunately, the choice among designs is often made on the basis of what other investigators have done, irrespective of whether their choice was optimal or whether the setting of the earlier study was similar to that of the present one. A good design should, however, maximize the probability of obtaining useful results. Although the term useful must be defined by each investigator, the definition should consider a variety of factors. For example, choosing a design solely on the basis that it has sufficient power to reject a false null hypothesis may be too restrictive. Clearly, the choice of design must be made within the constraints imposed by factors such as financial and administrative feasibility.

Existing textbooks on statistical design provide only broad statements about the utility of alternative designs and little or no guidance as to their application in real-world research settings such as schools and classrooms. The Bayesian approach, however, has the potential for combining relevant factors into a model which allows the researcher to select a design in a rational and clearly-defined way (Raiffa & Schlaifer, 1961). Technically, this process is called pre-posterior analysis. Given prior experience, alternative designs and their probable results are analyzed relative to the utility of those results, and the design having the maximum utility is chosen. Another advantage of pre-posterior analysis is that it focuses attention on the important factors in choosing a design. The model facilitates the identification of critical points where precise information is necessary and, hence, where research efforts should be directed.

While some theoretical methods for pre-posterior analysis are available, few practical methods have been developed. What is needed are ways to make the methodology accessible to the performer of research on teaching, with his perhaps unique knowledge and experience. One way to achieve this goal is through the production of computer programs which interrogate the researcher at critical points and present not only the optimal design, but also an analysis of the relative importance of each critical point to the final choice of design.

Program 9.3.2: Problems in Developing Measurement Procedures to Describe Various Teaching Processes or Programs (Including Behaviors of Teachers and Students)

The following is a collection of partially related questions about problems in developing measurement procedures to describe the teaching-learning process.

- To what extent and under what conditions is the notion of "sequence" useful in describing processes?
- How and under what circumstances can the more complex time-series analyses be applied to the description of teaching processes?
- How and under what circumstances can signal detection or quantal response theory be applied to the description of teaching processes?
- How and under what circumstances can Markov processes be applied to description of teaching processes?
- To what extent can present multi-dimensional scaling procedures, both metric and non-metric, be employed for meaningful reduction of extensive collections of data describing teachers and students?

Program 9.3.3: Evolutionary Operation

In what way, if any, is the concept of evolutionary operation (Box & Draper, 1968) useful for investigations of the teaching process?

Program 9.3.4: Organizational Development Methodology for Use in Formative Research on Teaching Strategies

Over the past ten years organizational development, as a field of inquiry into the analysis of the adequate functioning of groups, has developed a systematic methodology which, at present, is primarily used in industry and government. Work like that of March (1965) and Argyris (1971) may offer considerable insight into attempts to carry out adequate formative research on teaching strategies.

Program 9.3.5: Computer Simulation

The computer simulation of human behavior carried out by political scientists such as Newell and Simon (1961, 1968) and by psychologists such as Abelson (1963) might yield insights useful in research on teaching. Such insights may result in providing more resources for dynamic modeling of the teaching process.

Program 9.3.6: Path Analysis and Other Models for Estimating Causal Relationships

The objective of this program would be to consider the variety of techniques used to estimate causal relationships by people in political science and sociology and determine their applicability to research on teaching. The simplest of the approaches is "path analysis"--an approach which has already been disseminated somewhat, at least in its most

primitive form (Werts & Linn, 1970; Duncan, 1966). A serious discussion of the application and misapplication of path analysis in research on teaching would be useful. In addition, research techniques from Blalock (1964, 1971) and others, using partial correlational analyses and certain types of multi-stage least squares analyses given assumptions about causal ordering, might be useful to research on teaching.

Program 9.3.7: Scaling Methods from Consumer Research

The objective of this program would be to consider the variety of scaling methods developed in consumer research for possible application to research on teaching (Crespi, 1961; Green, 1970; Gallup, 1972).

Program 9.3.8: Generalizing from Non-Random Samples

When data are collected on a non-random sample of teachers and students, is the possibility of valid inference to the complete population eliminated? Recently, techniques have been developed for estimating relationships among variables even when marginal distributions have been biased (Goodman, 1972, 1973). When are such procedures appropriate for research on teaching?

Program 9.3.9: Investigation of Potential Uses of Exploratory Data Analysis

Modern data analysis entails a philosophical reorientation of statistical practice. A scientific ideal--formulate hypothesis, design and execute experiment, accept or reject hypothesis--is still honored, but the scientist is also encouraged to explore all available data looking for new hypotheses, unusual phenomena, and re-expressions of information. Much emphasis is placed on graphic displays and other simple techniques which enable a data analyst to know his data more intimately and can be used without the aid of the computer (Tukey, 1972). Another emphasis, one that takes maximum advantage of new computers, is on robust resistant methods which are useful in a wide variety of real-world situations where the usual statistical assumptions are questionable.

Project 9.3.9.1: Stem-and-Leaf Plots. An example of a simple data-analytic technique is the stem-and-leaf plot (Tukey, 1970), which is a way of rearranging data to get the pictorial advantage of a histogram without the usual loss of information. The stem-and-leaf is about as easy to form as a histogram, and the computing of medians and quartiles (hinges) and the identifying of outliers is then greatly facilitated.

Project 9.3.9.2: Robust/Resistant Regression. Robust/resistant regression (Beaton & Tukey, 1974) is an example of an attempt to avoid the emphasis on "fitting the unfittable" that is intrinsic to the least squares methods of squaring residuals before minimizing. It is easy to find or construct problems where least squares procedures fail to fill any points well, whereas estimation and/or smoothing approaches may fit the fittable very well while signalling, but not fitting, the outliers. Robust/resistant regressions fit almost as well as least squares in the ideal (Gaussian) case, and require only about 2 to 6 times as much computer time as classical regression (Miller, 1968; Mosteller & Tukey, 1968; and Quenouille, 1949). Other analysis methods, not specifically

discussed by Tukey but applicable to outliers or data points which do not seem to "fit" the model, are analyses using trimmed means (means which do not include the outlying points) or medians.

Project 9.3.9.3: Jackknife Procedures. The jackknife procedure is another general-purpose tool for estimation and hypothesis testing which holds up well in a variety of situations while losing little efficiency in ideal cases. In addition, the jackknife can be used in a number of situations in which other methods are unavailable or incomputable. The cost of the jackknife is fairly modest in typical situations.

Program 9.3.10: Analysis Models for the Estimation of Non-Additive Effects of Teaching in Other than Factorial Designs

In most experimental studies, the parameter of interest is one of location, i.e., whether or not groups differ with respect to their means. Sometimes, inequality of variances is also observed. Such inequality can indicate a non-additive model, e.g., $Y = O_1 X + O_2$, where X is a control value for a particular student and Y is the experimental value for that student. The model specifies both additive and multiplicative effects where O_1 can be thought of as a learning rate parameter. This and other models for non-additive effects may be useful for research on teaching (Lohnes, 1972). It should be noted that concern for non-additive effects is related at least in part to concern for aptitude-treatment interactions.

Program 9.3.11: Development of Statistical Decision Theory Models for Monitoring the Instructional Process

Statistical decision theory has been found to have important application in business and economics and was introduced to education by Cronbach and Gleser in 1957. An advantage of decision theory (Novick, 1971; Novick & Jackson, 1970; Pollack, 1968) is that it permits several aspects of the decision problem to be considered simultaneously in a coherent manner. Its drawbacks are the complexity of its mathematical formulation and the difficulty of providing some of the judgmental input required for its implementation. The first difficulty of decision theory (complexity of its mathematical formulation) has succumbed to repeated attack by a large number of able statisticians. Also, greater skill on the part of educational statisticians in formulating their problems in relatively simple, but realistic, ways has helped simplify decision theory. The second difficulty (input required for implementation) is being reduced as interactive computer systems become available to help investigators quantify coherently their utilities and prior probabilities.

Project 9.3.11.1: Monitoring Individualized Instruction Programs. One area in which decision theory is useful is that of monitoring individualized instructional programs (Hambleton, 1973). In such programs, decision points are continually appearing and a rational and coherent procedure for making the advance-return decision is required. While some work has been done, much more is needed. Methods for choosing among various instructional modes are needed, as are methods of combining serially-gathered data on individual students.

Project 9.3.11.2: Decision-Theoretic Approach to Problems of Test Bias. The area of bias in selection, or culture-fair testing, is another in which a decision-theoretic formulation can have general applicability. While simple solutions are possible, much needs to be done to study the relationship between students' and institutions' utility structures and to ascertain how differences between these structures affect acceptability of selection and self-selection fairness. Also, much work needs to be done with sophisticated utility structures and with multiple predictor and multiple outcome formulations.

APPROACH 9.4

CONSIDER THE UTILITY OF STANDARDS FOR IMPROVING METHODOLOGICAL PRACTICE IN RESEARCH ON TEACHING

Two reasons were suggested for attempting to develop methodological standards for research on teaching. The first was that some research on teaching contains methodological flaws, many of which are common across time and across studies. The second reason was that much research on teaching has not been cumulative. It is difficult, and sometimes impossible, for teachers or educational researchers to pool results from studies dealing with common interest areas.

Setting methodological standards has been a fairly common practice, motivated by the hope that through the establishment of a set of minimal levels or standards of acceptable quality, the consumer will be protected. Perhaps the most relevant example is the APA-AERA-NCME set of standards for test publishers. Several groups are also considering the possiblity of standards for program evaluation. The consensus of the Panel however, was that it is not possible nor desirable to legislate through standards the methodological quality of research on teaching.

Researchers must take a creative approach to data analysis and be willing to use multiple strategies in order to obtain the full utility of their data. It seems likely that methodological standards for research on teaching would militate against such practices and, instead, promote rather routine and unthinking analyses. Further, research on teaching has special yet varying methodological needs which a single set of standards could not begin to address. It was decided, therefore, to discourage the development of methodological standards. In place of standards, the Panel recommended several programs to facilitate communication of information about how to handle methodological problems that are of major concern in research on teaching.

This Approach is based on two major ideas: (a) the establishment of archival data that can be used for secondary analysis and for illustration of the results of alternative design and analysis strategies; and (b) the establishment of procedures for disseminating the results of Approaches 9.1, 9.2 and 9.3 to persons engaged in research on teaching. The goal is to encourage those doing research on teaching to use the "best known practices" in measurement, design, and data analysis.

Program 9.4.1: Secondary Analyses and Alternative Designs

It seems desirable to commission competent educational research methodologists to review and critique past studies in the field of research on teaching. These reviews of past research should contain a variety of secondary analyses and compare the utility of those strategies to the initial analysis. In addition, they should identify and describe alternative design and analysis strategies for addressing the research question that could not be illustrated through secondary analysis. This information should be documented and made available for wide dissemination, particularly to the educational community interested in research on teaching.

Program 9.4.2: Research Data Archive

Professional journals have editorial policies and formats that greatly restrict the amount of information and exploration that might be of interest to other researchers. While it is not appropriate (nor perhaps desirable) to attempt changes in existing publication practices, it is nevertheless true that some interested consumers of the literature could profit from more complete reports. The Panel suggests that an archive be created which would allow researchers to submit a more inclusive summary of their total research findings and their actual research data at some summary level. Two main concerns are directly related to this. First, what kinds of research results and summary data are most useful to archive? Second, where should this archive be placed and how can it be made readily accessible to researchers of teaching? These archival data are directly related to facilitating Program 9.4.1 on Secondary Analyses and Alternative Designs.

Program 9.4.3: Training Programs

It was suggested that professional organizations such as the American Educational Research Association be encouraged to sponsor methodological training sessions for researchers with on-going projects in research on teaching. These training sessions should be applied and project-based, not theoretical in nature. Another training suggestion was that fellowship programs be created specifically for mid-career researchers. These would be structured programs that would bring into the university community persons who are actively involved in research on teaching. At the university, these mid-career researchers would be able to take research methodology courses and tap faculty ideas relating to their specific research.

Program 9.4.4: Providing the Methodological Capacity to Support Research on Teaching

Ideally, a person conducting research on teaching should have not only an interest in and understanding of the research issue, but also the methodological sophistication to identify and, where necessary, adapt methodology for his particular research needs. Unfortunately, this is not always the case. The previous programs in Approach 9.4 dealt with potentially long-run solutions to the problem through training. It would be helpful, however, if there were some short-run strategies. One possible strategy would be to make competent methodologists more readily available to persons engaged in research on teaching, and to do so in a way that sustains their availability over the duration of a research project. This might be accomplished by partially supporting methodological specialists on the staffs of state departments of education, research and development centers, or laboratories--specialists who would have specific assignments to research projects on teaching.

Program 9.4.5: Test Evaluation Manuals

To guide selection from existing measurement strategies for research on teaching, it is suggested that test evaluation manuals be published for different areas of the teaching-learning process. The U.C.L.A. Center for the Study of Evaluation has completed several manuals on tests of student characteristics. Similar manuals could be devised stressing other areas of the teaching-learning process. Within each area, such as measuring teacher effectiveness, an extensive search should be made for relevant instruments, both published and experimental, in order to ascertain the number and quality of instruments that have already been developed to assess variables in that particular area.

Each test evaluation manual should give critical information about the relevant instruments such as:

1. a summary of the purpose of the test
2. the type of instrument (i.e., interview schedule, self-report, etc.)
3. evaluations of the quality of the instrument
4. a sampling of actual test items.

TENTATIVE PRIORITY ESTIMATES

At the close of the Conference, Panel members were asked to rate each program (or each project, where such was specified for a particular program) on the basis of its judged importance to research on teaching. The criteria for judged importance were left to the discretion of the individual members, but clearly the ratings must be interpreted within the context of the Panel's concerns, i.e., research methodology. Since the Panel was small, since its members were subject to shifts in set as they focused on specific problem areas, and since the ratings were done at the end of an exhausting set of sessions, they should not be over interpreted. Nonetheless, they are presented here as a stimulus to the reader to make similar comparisons among programs.

The ratings were made on a scale ranging from 1 (of little importance) to 3 (of great importance). Table 1 shows the resulting order of programs within each of the four Approaches. Programs which were not rated by the Panel and which cannot therefore be located in the ordering are nonetheless included at the bottom of the listings.

TABLE I. TENTATIVE PRIORITY ESTIMATES
(Programs listed in order of descending importance, by short title, on scale of 1.0 to 3.0)

APPROACH 9.1: DESIGN & ANALYSIS STRATEGIES	
9.1.1: Hierarchically-Nested Data	
9.1.1.1: Estimating Relations	---
9.1.1.2: Models for Data Aggregation	---
9.1.1.3: Unbalanced Designs	---
9.1.1.4: Violating Independence Assumptions	---
9.1.1.5: Non-Independent Student Data	2.8
9.1.4: Use of Implementation Data	2.6
9.1.5: Longitudinal Designs	2.6
9.1.3: Quasi-Experimental/Correlational Studies	
9.1.3.1: Multiple Fallible Covariables	2.5
9.1.9: Current Educational Practice	
9.1.9.1: Behavior-Setting Types	2.3
9.1.9.2: Ways to Observe Behavior Settings	2.1
9.1.9.3: Survey of Classroom Behavior Settings	2.1
9.1.7: Combining Related Studies over Time	2.2
9.1.2: Conducting "True Experiments"	
9.1.2.1: Incentives for Participation	1.7
9.1.2.2: Ethical Issues of "True Experiments"	2.0
9.1.2.3: "True" within Quasi-Experiments	---
9.1.8: Studying Teacher Effectiveness	
9.1.8.1: Lack of Random Assignment	1.7
9.1.8.2: Following Students over Time	1.7
9.1.8.3: Combining Intraclass Correlations	1.8
9.1.6: Teacher-Student Interaction Models	1.4

APPROACH 9.2: MEASUREMENT STRATEGIES	
9.2.1: Educational Significance of an "Effect"	
9.2.1.1: Indirect Validation	3.0
9.2.1.2: Direct Validation	2.8
9.2.5: Test Bias	2.7
9.2.8: Program Implementation	
9.2.8.1: Measuring Implementation	2.5
9.2.8.2: Stability of Behaviors	2.5
9.2.9: Reliability of Observation	2.4
9.2.3: Tests with Face Validity	
9.2.3.1: Measures Tied to Instruction Purposes	2.3
9.2.3.2: Measures of Non-Cognitive Outcomes	2.4
9.2.3.3: Observations of Process Variables	2.4
9.2.2: Desirable Properties of Tests	2.3
9.2.6: Evaluation of Profiles	2.3
9.2.4: Crossed Design Achievement Tests	2.2
9.2.7: Defining Desired Teacher Performance	1.7
9.2.10: Criterion Referenced Tests	---

APPROACH 9.3: METHODS FROM OTHER DISCIPLINES	
9.3.2: Measuring Teaching Processes	2.8
9.3.10: Non-Additive Effects of Teaching	2.8
9.3.1: Optimal Designs for Research	2.6
9.3.8: Generalizing from Non-Random Samples	2.5
9.3.5: Computer Simulation	2.0
9.3.3: Evolutionary Operation	1.7
9.3.4: Organizational Development Methodology	1.3
9.3.6: Path Analysis	1.0
9.3.7: Scaling Methods from Consumer Research	1.0
9.3.9: Exploratory Data Analysis	---
9.3.11: Statistical Decision Theory Models	---

APPROACH 9.4: IMPROVING METHODOLOGICAL PRACTICE	
9.4.1: Secondary Analyses & Alternative Designs	2.5
9.4.2: Research Data Archive	2.3
9.4.3: Training Programs	2.0
9.4.4: Methodological Support Capacity	2.0
9.4.5: Test Evaluation Manuals	---

SUMMARY

Although much useful research on teaching has been conducted, the utility of some of the research has been limited because of methodological problems. In some cases, appropriate methodology was not available; in other cases, established best practices were not followed. In addition, there have been cases where methodologies were borrowed from other research disciplines without a careful rethinking of the assumptions involved. Thus, the goal adopted by Panel 9 was to improve the validity and utility of measurement, design, and analysis in research on teaching. To that end, four Approaches were adopted covering the stimulation of new methodological knowledge as well as identification and translation of useful methodological knowledge from other disciplines.

The first Approach called for the development and testing of new design and analysis strategies. Perhaps the major impression left by reviews of current research on teaching is that problems of design and analysis are encountered at many stages, and are solved, if at all, in an imitative or derivative fashion drawing on analogies with earlier studies, especially those in agriculture. The Panel felt that it is time to put forth more systematic efforts toward developing principles for the design and analysis of studies within the special and possibly unique context of problems of education in general and the study of teaching in particular. Solutions are needed for design and analysis problems such as cumulating results from distinct but related studies, controlling the influences of confounding variables, and studying longitudinal effects.

The second Approach called for an increased understanding of existing measurement strategies for research on teaching and where appropriate the development of new measurement strategies. Much of the history of measurement in the behavioral sciences relates to concerns for measuring individual student aptitudes and achievement. Although this work has been and will continue to be of value to research on teaching, there are other important aspects of measurement. Greater attention should be given to problems of

test bias. Psychometric theory must be developed to support the criterion referenced test movement. Better measures of the teaching/learning process are required, particularly in natural settings. Yet, another example is the need for group assessment measures as contrasted with measures designed to assess individual differences.

Current and pending legislation has given a sense of urgency to the solution of these design, analysis, and measurement problems. Thirty-one states are now considering laws requiring all applicants for a teaching license to demonstrate their teaching effectiveness. One example is the Stull Act, 1972, of California, which requires all school districts to evaluate their teachers.

The first two Approaches reflected a concern for the development of new design, analysis, and measurement techniques which serve the unique needs of research on teaching. Most of the methodology currently used in research on teaching, however, was originally developed in other research disciplines. There are at least two reasons why continued identification, translation, and dissemination of methodologies from other research disciplines (Approach 3) seems warranted. First, in many cases, these borrowed methodologies have served the researchers of teaching quite well. Second, where existing useful methodologies are available, duplication of development should be avoided. Several potentially useful methodologies were identified.

The fourth Approach considered the utility of setting standards of methodological practice within research on teaching. The consensus of the Panel was that it is neither desirable nor possible to legislate (through standards) the methodological quality of research on teaching. Researchers must take a creative approach to data analysis and be willing to use multiple strategies in order to obtain full utility of their data. It seems likely that methodological standards for research on teaching would militate against such practices and, instead, promote rather routine and unthinking analyses. Further, research on teaching has special yet varying methodological needs which one set of standards could not begin to address. It was decided, therefore, to discourage the development of methodological standards. In place of standards, the Panel recommended several programs to facilitate communication of information about how to handle methodological problems that are of major concern in research on teaching.

REFERENCES

Abelson, R. P. Computer simulation of "hot" cognition. In S. S. Tomkins & S. Messick (Eds.), Computer simulation of personality. New York: Wiley, 1963.

Acland, H. A study of teacher effects based on student achievement scores. Cambridge, Mass.: Huron Institute, Unpublished Manuscript, 1974.

Anderson, T. W. Statistical analysis of time-series. New York: Wiley, 1971.

Argyris, C. Management and organizational development: The path from XA to YB. New York: McGraw-Hill, 1971.

Beaton, A. E., & Tukey, J. W. The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. Technometrics, 1974, 16, 147-174.

Blalock, H. M., Jr. Causal inferences in nonexperimental research. Chapel Hill, N.C.: University of North Carolina Press, 1964.

Blalock, H. M., Jr. Causal models in the social sciences. Chicago: Aldine-Atherton, 1971.

Bissell, J. S. Implementation of planned variation in Head Start. Bethesda, Md.: National Institute of Child Health and Human Development, 1971.

Box, G. E., & Draper, N. R. Evolutionary operation: A method for increasing industrial productivity. New York: Wiley, 1968.

Boyer, E. G., Simon, A., & Karafin, G. (Eds.). Measures of maturation: An anthology of early childhood observation instruments (3 vols.). Philadelphia: Research for Better Schools, 1973.

Brophy, J. Stability of teacher effectiveness. American Educational Research Journal, 1973, 10, 245-252.

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

Campbell, D. T. & Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963, 171-246.

Cleary, T. A. Test bias: Prediction of grades of Negro and White students in integrated colleges. Journal of Educational Measurement, 1968, 5, 115-124.

Cochran, W. G. Errors of measurement in statistics. Technometrics, 1968, 10, 637-666.

Cole, N. S. Bias in selection. Journal of Educational Measurement, 1973, 10, 237-255.

Crespi, I. Use of a scaling technique in survey. Journal of Marketing, 1961, 25, 69-72.

Cronbach, L. J., & Gleser, G. C. Psychological tests and personnel decisions (1st edition). Urbana, Ill.: University of Illinois Press, 1957.

Duncan, O. D. Path analysis: Sociological examples. American Journal of Sociology, 1966, 72, 1-16.

Gallup, G. The sophisticated poll watcher's guide. Princeton, N.J.: Princeton Opinion Press, 1972.

Gellert, E. Systematic observations: A method in child study. Harvard Educational Review, 1955, 25, 179-195.

Glass, G. V. Estimating the effects of intervention into a non-stationary time-series. American Educational Research Journal, 1972, 9, 463-477.

Goodman, L. A. A model for the analysis of surveys. American Journal of Sociology, 1972, 77, 1035-1086.

Goodman, L. A. Causal analysis of data from panel studies and other kinds of surveys. American Journal of Sociology, 1973, 78, 1135-1191.

Grannis, J. C. Columbia classroom environments project. New York: Institute for Pedagogical Studies, Teachers College, Columbia University (U.S. Office of Education, Contract No. 71-0593), 1973.

Green, P. Multidimensional scaling and related techniques in marketing analysis. Boston: Allyn & Bacon, 1970.

Gump, P. The classroom behavior setting: Its nature and relation to student behavior. U.S. Office of Education, Contract No. 5-0334, 1967.

Hambleton, R. K. A review of testing and decision-making procedures for selected individualized instructional programs. ACT Technical Bulletin No. 15, Iowa City: Research and Development Division, American College Testing Program, 1973.

Harris, C. W. Characteristics of two measures of profile similarity. Psychometrika, 1955, 20, 289-297.

Harris, C. W., Alkin, M. C., & Popham, W. J. Problems in criterion-referenced measurement. CSE Monograph Series in Evaluation #3, University of California at Los Angeles, 1974.

Harris, M. L., & Harris, C. W. A structure of concept attainment abilities. Madison: Wisconsin Research and Development Center in Cognitive Learning, 1973.

Hoepfner, R., Bradley, P. A., & Doherty, W. J. National priorities for elementary education. CSE Monograph Series in Evaluation #2, University of California at Los Angeles, 1974.

Iverson, G. R. Recovering individual data in the presence of group and individual effects. American Journal of Sociology, 1974, 79, 420-434.

Karlson, A. L. A naturalistic method for assessing cognitive acquisition of young children participating in preschool programs. Doctoral dissertation, University of Chicago, 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.

Light, R. J., & Smith, P. V. Accumulating evidence: Procedures for resolving contradictions among different research studies. Harvard Educational Review, 1971, 41, 429-471.

Linn, R. L., & Werts, C. E. Considerations for studies of test bias. Journal of Educational Measurement, 1971, 8, 1-4.

Lohnes, P. R. Statistical descriptors of school classes. American Educational Research Journal, 1972, 9, 547-556.

March, J. G. Handbook of organizations. Chicago: Rand McNally, 1965.

Miller, R. G., Jr. Jackknifing variances. Annals of Mathematical Statistics, 1968, 39, 567-582.

Mosteller, F., & Tukey, J. W. Data analysis, including statistics. In G. Lindzey & E. Aronson (Eds.), Handbook of social psychology (Vol. II). Reading, Mass.: Addison-Wesley, 1968.

Newell, A., & Simon, H. A. Computer simulation of human thinking. Science, 1961, 134, 2011-2017.

Newell, A., & Simon, H. A. Simulation: individual behavior. In D. L. Sills (Ed.), International encyclopedia of the social sciences (Vol. 14). New York: Macmillan, 1968, 262-268.

Novick, M. R. Bayesian considerations in educational information systems. In Proceedings, Invitational Testing Conference. Princeton, N.J.: Educational Testing Service, 1971.

Novick, M. R., & Jackson, P. H. Bayesian guidance technology. Review of Educational Research, 1970, 40, 459-494.

Olkin, I. Correlations revisited. In Seventh Annual Phi Delta Kappa Symposium on Educational Research, Improving experimental design and statistical analysis. Madison: University of Wisconsin Press, 1965.

Pollack, I. Information theory. In D. L. Sills (Ed.), International encyclopedia of the social sciences (Vol. 7). New York: Macmillan, 1968, 331-337.

Porter, A. C. Some comments on the summative evaluation of compensatory education programs. Paper presented at the Fifth Annual Michigan Conference on Compensatory Education, 1973.

Porter, A. C., & Chibucos, T. R. Selecting analysis strategies. In G. Borich (Ed.), Evaluating educational programs and products. Educational Technology Press, 1974.

Porter, A. C., & McDaniels, G. L. A reassessment of the problems in estimating school effects. Paper presented at the American Association for the Advancement of Science, 1974.

Quenouille, M. Approximate tests of correlation in time-series. Journal of the Royal Statistical Society, 1949, Series B 11, 68-84.

Raiffa, H., & Schlaifer, R. Applied statistical decision theory. Boston: Division of Research, Graduate School of Business Administration, Harvard University, 1961.

Robinson, W. S. Ecological correlations and the behavior of individuals. American Sociological Review, 1950, 15, 351-357.

Rosenshine, B. Evaluation of classroom instruction. Review of Educational Research, 1970, 40, 279-300.

Schmidt, F. L., & Hunter, J. E. Racial and ethnic bias in psychological tests: Divergent implications of two definitions of test bias. American Psychologist, 1974, 29, 1-8.

Simon, A., & Boyer, E. G. (Eds.). Mirrors for behaviors: An anthology of observation instruments (3 Vols.). Philadelphia: Research for Better Schools, 1967-1970.

Stodolsky, S. S. Defining treatment and outcome in early-childhood education. In H. J. Walberg & A. T. Kopan (Eds.), Rethinking urban education. San Francisco: Jossey-Bass, 1972.

Thorndike, R. L. Concepts of culture-fairness. Journal of Educational Measurement, 1971, 8, 63-70.

Tukey, J. W. Exploratory data analysis. Reading, Mass.: Addison-Wesley, 1970.

Tukey, J. W. Some graphic and semigraphic displays. In T. A. Bancroft (Ed.), Statistical papers in honor of George W. Snedecor. Ames: Iowa State University Press, 1972.

Veldman, D. J., & Brophy, J. E. Measuring teacher effects on pupil achievement. Journal of Educational Psychology, 1974, 66, 319-324.

Votaw, D. F., Jr. Testing compound symmetry in a normal multivariate distribution. Annals of Mathematical Statistics, 1948, 19, 447-473.

Walker, D. K. Socioemotional measures for preschool and kindergarten children. San Francisco: Jossey-Bass, 1974.

Weick, K. E. Systematic observational methods. In G. Lindzey & E. Aronson (Eds.), Handbook of social psychology (Vol. III). Reading, Mass.: Addison-Wesley, 1968.

Werts, C. E., & Linn, R. L. Path analysis: Psychological examples. Psychological Bulletin, 1970, 74, 193-212.

Westinghouse Learning Corporation. The impact of Head Start: An evaluation of the effects of Head Start on children's cognitive and affective development. Bladensburg, Md.: Westinghouse Learning Corp. and Ohio University. U.S. Office of Economic Opportunity, Contract No. B89-4536, 1969.

NATIONAL PLANNING CONFERENCE ON STUDIES IN TEACHING

Sponsoring Program Dir.: Garry McDaniels, NIE
Conference Chair: N. L. Gage, Stanford U.
Asst. to Chair: Philip Winne, Stanford U.
Panel Coordinators (Staff, Arthur Young & Co.): Sandra Lafe Smith, William Callahan, Lillian Handy, Mary Carey, Albert Schreiber, Mark Versel, Blair Curry, Gerald Decker, Joseph Ryan, Elsa Graitcer

Contract Project Dir.: Alan Pittaway, Arthur Young & Co.
Conference Coord.: Robert MacDicken, Arthur Young & Co.
Participant at Large: Arthur Coladarci, Stanford U.

1. Teacher Recruitment, Selection, & Retention

Chair: James Deneen, ETS
Members: Dale Bolton, U. Washington
William Demmert, USOE
Goldine Gleser, U. Cincinnati
Sonja Nixon, Wildwood Elem. Sch., Mahtomedi, Minnesota
Robert Peck, U. Texas
Nathan Quinones, Board of Educ., Brooklyn
Advisory Members: Robert Bhaerman, AFT
Roy Edelfelt, NEA
David Imig, AACTE
James Scharf, EEOC
Richard Sharp, Shea & Gardner
Sec.: Susan Sherwin, ETS

2. Teaching as Human Interaction

Chair: Ned Flanders, Far West Laboratory for Educational R&D
Members: Bruce Biddle, U. Missouri
Jere Brophy, U. Texas
Norma Furst, Temple U.
Bryce Hudgins, Washington U. of St. Louis
Donald Medley, U. Virginia
Graham Nuthall, U. Canterbury, New Zealand
Doris Ray, Lathrop H.S., Fairbanks, Alaska
Melvyn Semmel, Indiana U.
Robert Soar, U. Florida
Sec.: Christopher Clark, Stanford U.

3. Teaching as Behavior Analysis

Chair: Don Bushell, Jr., U. Kansas
Members: Wesley Becker, U. Oregon
David Born, U. Utah
Robert Hawkins, Eastern Michigan U.
Girard Hottleman, Massachusetts Teachers Assn.
K. Daniel O'Leary, SUNY at Stoney Brook, N.Y.
Beth Sulzer-Azaroff, U. Massachusetts
Carl Thoreson, Stanford U.
Doug Wilson, Mills Jr. H.S., Sacramento, Calif.
Advisory Members: Curt Braukmann, U. Kansas
Gilbert Hoffman, Bryan Elem. Sch., Washington, D.C.
Sec.: Judith Jenkins, U. Kansas

4. Teaching as Skill Performance

Chair: Richard Turner, Indiana U.
Members: Walter Borg, Utah State U.
Carl A. Grant, U. Wisconsin
Judy Henderson, Michigan State U.
Bruce Joyce, Stanford U.
Eugenia Kemble, UFT
Frederick McDonald, ETS
Bernard McKenna, NEA
Alan Purves, U. Illinois
Charles Stewart, Detroit Publ. Sch.
Beatrice Ward, Far West Laboratory for Educational R&D
Sec.: Mary Ella Brady, Indiana U.

5. Teaching as a Linguistic Process in a Cultural Setting

Chair: Courtney Cazden, Harvard U.
Members: Douglas Barnes, U. of Leeds, England
Arno Bellack, Columbia U.
Heidi Dulay, SUNY at Albany, N.Y.
Ian Forsyth, Center for Language in Primary Educ., London
John Gumperz, U. Calif. at Berkeley
William Hall, Rockefeller U.
Roger Shuy, Georgetown U.
B. O. Smith, U. of South Florida
Alan Tindall, SUNY at Buffalo, N.Y.
Sec.: Elsa Bartlett, Rockefeller U.

6. Teaching as Clinical Information Processing

Chair: Lee Shulman, Michigan State U.
Members: Thomas Good, U. Missouri
Edmund Gordon, Columbia U.
Philip Jackson, U. Chicago
Marilyn Johnson, San Jose Unified Sch. District, Calif.
Sara Lightfoot, Harvard U.
Greta Morine, Calif. State U. at Hayward
Ray Rist, Portland State U., Oregon
Paul Slovic, Oregon Research Institute
Bernard Weiner, U. Calif. at Los Angeles
Sec.: Ronald Marx, Stanford U.

7. Instructional Personnel Utilization

Chair: Robert Egbert, U. Nebraska
Members: Edward Barnes, NIE
George Brain, Washington State U.
Elizabeth Cohen, Stanford U.
Walter Hodges, Georgia State U.
Ruth Jones, Baskerville Sch., Rocky Mount, N.C.
Joseph Moren, Hibbing H.S., Minnesota
James O'Hanlon, U. Nebraska
John Prasch, Supt. of Schools, Lincoln, Neb.
Richard Schmuck, U. Oregon
Sec.: Linda Douglas, Lincoln Publ. Sch., Neb.

8. Personnel Roles in New Instructional Systems

Chair: Susan Meyer Markle, U. Illinois at Chicago Circle
Members: Eva Baker, U. Calif. at Los Angeles
Catherine Barrett, Syracuse Publ. Sch., N.Y.
Louis Bright, Baylor U.
Gerald Faust, Brigham Young U.
Robert Gagné, Florida State U.
Barbara Goleman, Miami/Dade Co. Publ. Sch., Fla.
Melvin Leasure, Oak Park Publ. Sch., Michigan
Gaea Leinhardt, U. Pittsburgh
Harold Mitzel, Pennsylvania State U.
Charles Santelli, N.Y. State United Teachers
S. Thiagarajan, Indiana U.
Advisory Member: Dean Jamison, ETS
Sec.: Linda Crnic, U. Illinois at Chicago Circle

9. Research MethodoTogy

Chair: Andrew Porter, Michigan State U.
Members: T. Anne Cleary, CEEB
Chester Harris, U. Calif. at Santa Barbara
Richard Light, Harvard U.
Donald L. Meyer, U. Pittsburgh
Barak Rosenshine, U. Illinois
Marshall Smith, Harvard U.
Susan Stodolsky, U. Chicago
Sec.: Linda Glendening, Michigan State U.

10. Theory Development

Chair: Richard Snow, Stanford U.
Members: David Berliner, Far West Laboratory for Educational R&D
William Charlesworth, U. Minnesota
Miles Meyers, Oakland H.S., Calif.
Jonas Soltis, Columbia U.
Sec.: Penelope Peterson, Stanford U.