ED 185 099 TM 800 104

AUTHOR Kim, Eui-Do; Coller, Alan R.
TITLE Research Methodologies Pertinent to the Study of Schooling Effects: A Synthesis.
INSTITUTION Research for Better Schools, Inc., Philadelphia, Pa.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE May 78
NOTE 115p.
AVAILABLE FROM Research for Better Schools, Inc., 444 North Third St., Philadelphia, PA 19123 ($7.00).

EDRS PRICE MF01/PC05 Plus Postage.
DESCRIPTORS Analysis of Covariance; Analysis of Variance; Correlation; *Educational Research; Elementary Secondary Education; Mathematical Models; Multiple Regression Analysis; Multivariate Analysis; Outcomes of Education; Quasiexperimental Design; *Research Methodology; *Sampling; School Role; *Statistical Analysis; Student Characteristics; Teacher Characteristics
IDENTIFIERS Coleman Report; Linear Models; Regression

ABSTRACT

A broad spectrum of methodologies pertinent to studies of schooling effects are reviewed. Methodological issues and problems are addressed according to a three-dimensional conceptual framework: (1) indicators of schooling effects; (2) study approaches; and (3) units of analysis. Problems and uses of status attainment and difference scores as indicators of schooling effects are discussed first. Study approaches to schooling effects are divided into two broad categories: experimental and nonexperimental. Methodological issues related to the experimental approaches are discussed in relation to two types of designs: experimental group only designs and control group designs. Problems related to the nonexperimental approaches are reviewed according to: partitioning of explained variance, comparison of regression coefficients, nonlinear regression methods, and causal models. Issues and problems related to the units of analysis are presented by contrasting two positions: that data should be analyzed at the individual student level, and that data should be analyzed at the classroom, school, or district level. A third position is reviewed: that multilevel analyses should be performed because schooling effects might result from many sources at many different levels. Finally, some methodological trends are identified and their implications for further schooling briefly considered. (Author/CTM)

RESEARCH METHODOLOGIES PERTINENT TO THE STUDY OF SCHOOLING EFFECTS: A SYNTHESIS

Eui-Do Rim and Alan R. Coller
Research for Better Schools, Inc.

Submitted to
The National Institute of Education
as part of a Final Report from
The Basic Skills Component
Research for Better Schools, Inc.
Suite 1700/1700 Market Street
Philadelphia, Pennsylvania 19103

May 1978

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Health, Education, and Welfare. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

Abstract

This paper reviews a broad spectrum of methodologies pertinent to studies of schooling effects. Methodological issues and problems are addressed according to a three-dimensional conceptual framework consisting of: (1) indicators of schooling effects, (2) study approaches, and (3) units of analysis. Problems and uses of status attainment and difference scores as indicators of schooling effects are discussed first. Study approaches to schooling effects are divided into two broad categories: experimental and nonexperimental. Methodological issues related to the experimental approaches are discussed in relation to two types of designs: experimental group only designs and control group designs. Problems related to the nonexperimental approaches are reviewed according to: partitioning of explained variance, comparison of regression coefficients, nonlinear regression methods, and causal models. Issues and problems related to the units of analysis are presented by contrasting two positions: that data should be analyzed at the individual student level, and that data should be analyzed at the classroom, school, or district level. A third position has emerged: that multilevel analyses should be performed because schooling effects might result from many sources at many different levels. Finally, some methodological trends are identified and their implications for further schooling effects studies are briefly considered.

TABLE OF CONTENTS

I. INTRODUCTION

Large-scale studies on student achievement (and secondary analyses thereof) concerned with the relative affects of schools, programs, and/or teachers have consistently yielded findings that challenge even our most cherished beliefs about the impact of education in America (Averch, Carroll, Donaldson, Kiesling, & Pincus, 1972; Circirelli, Cooper, & Granger, 1969; Coleman, Campbell, Hobson, McPartland, Mood, Weinfeld, & York, 1966; Heath & Nielson, 1974; Jencks, Smith, Acland, Bane, Cohen, Gintis, Heyns, & Michelson, 1972; and Mayeske, Wisler, Beaton, Weinfeld, Cohen, Okada, Proshek, & Tabler, 1969). However, critics of these studies, such as Cain and Watts (1970), Campbell and Erlebacher (1970), Guthrie (1973), and Hanushek and Kain (1972), ask: Are the results of these studies truly reflective of our schooling efforts, or are they at least partly artifacts of the research methodologies used by behavioral and/or social scientists as they study schooling effects?

It is instructive to examine what several educational researchers have had to say in response to such a question. Cain and Watts indicated that "the analytical part of the Coleman Report has such methodological shortcomings that it offers little policy guidance" (p. 228). In a scholarly critique of the Westinghouse/Ohio University study of compensatory education, Campbell and Erlebacher concluded: "It is tragic that the social experiment evaluation most cited by presidents, most influential in government decision making, should have contained such a misleading bias" (p. 203). As to the assignment of blame, they responded:

"In this instance, the failu ⁄ came from the inadequacies of the social science methodological community (including education, psychology, economics, and sociology) which as a population was not ready for this task" (p. 204). Herriott and Muse (1973) believe that "we currently lack both the conception and methodological tools essential for an unambiguous attribution of educational effects among competing explanatory variables" (p. 231). And Cronbach (1976), in a paper examining alternative ways of analyzing data, expressed his deeply felt concerns about methodology currently in use in educational research:

> 1. The majority of studies of educational effects--whether classroom experiments, or evaluation of programs, or surveys--have collected and analyzed data in ways that conceal more than they reveal. The established methods have generated false conclusions in many studies.
>
> 2. The traditional research strategy--pitting substantive hypotheses against a null hypothesis and requiring statistical significance of effects--can rarely be used in educational research. Samples large enough to detect strong but probabilistic effects are likely to be prohibitively costly. (p. 1)

Such critiques have stimulated researchers to consider carefully the advantages and disadvantages of employing one method over another, and have called attention to the need for methodologies that can be employed as alternatives to established practice. The use of "commonality analysis" (Mood, 1971) in the Instructional Dimensions

Study (Brady, Clinton, Sweeney, Peterson, & Poyner, 1977), of "path analysis" (Blalock, 1964; Duncan, 1966; Werts & Linn, 1970b; Wright, 1921) in the Beginning Teachers Evaluation Study, Phase II (McDonald & Elias, 1976), and of "polynomial regression analyses" (Cohen & Cohen, 1975; Fisher, 1925; Kerlinger & Pedhazur, 1973; Pearson, 1905) in the Follow Through Classroom Process Measurement and Pupil Growth Study (Soar & Soar, 1972) shows several of the more recent attempts to make use of more appropriate research methodologies in schooling effects studies. The Horst, Tallmadge, and Wood (1975) paper, developed i.. an attempt to improve the methodology used in the evaluation of educational programs in Title I, has had an impact well beyond its relatively limited goals, and is yet another example of the potential usefulness of endeavors such as this.

Purposes

The major purposes for this review and synthesis effort are: (1) to examine issues and problems associated with methodologies pertinent to research on schooling effects, (2) to call attention to recent developments in relevant research methodology, and (3) to describe general methodological trends in the study of schooling effects. It is anticipated that the knowledge bases established here will facilitate efforts to select and utilize methodologies that will be effectiv: and feasible for providing feedback to developers of technology designed to assist practitioners to identify and exploit opportunities for improving instruction and its outcomes.

Methods and Procedures

An Organizational Schema

The sheer number and richness of methods employed to study schooling effects represented both a blessing and curse to the reviewers; there was no lack of methodological areas to explore but seemingly no simple way by which discussions of those topics could be organized. In the process of seeking ways to develop an organizational framework for this review, it proved useful to assume that extant methodologies could be located in a multidimensional space; that is, they could be characterized by a limited set of dimensions or facets (Perkins, 1977; Willems, 1969). Facet design, as viewed by Runkel and McGrath (1972), "is a way of laying out a domain for research; it specifies the limits of the domain and the presumed ordering of its subparts" (p. 20).

The descriptive space shown in Figure 1 is designed to represent both the domain of interest (i.e., research methodologies pertinent to schooling effects studies) and the organizational schema by which the review is ordered and delimited. The three displayed facets, that is, indicators of schooling effects, study approaches, and units of analysis, answer, respectively, these questions: What? How? Which analytic unit? Section II of this review deals with issues and problems related to the indicators of schooling effects, that is, the "what?" dimension. Issues and problems related to study approaches in research on

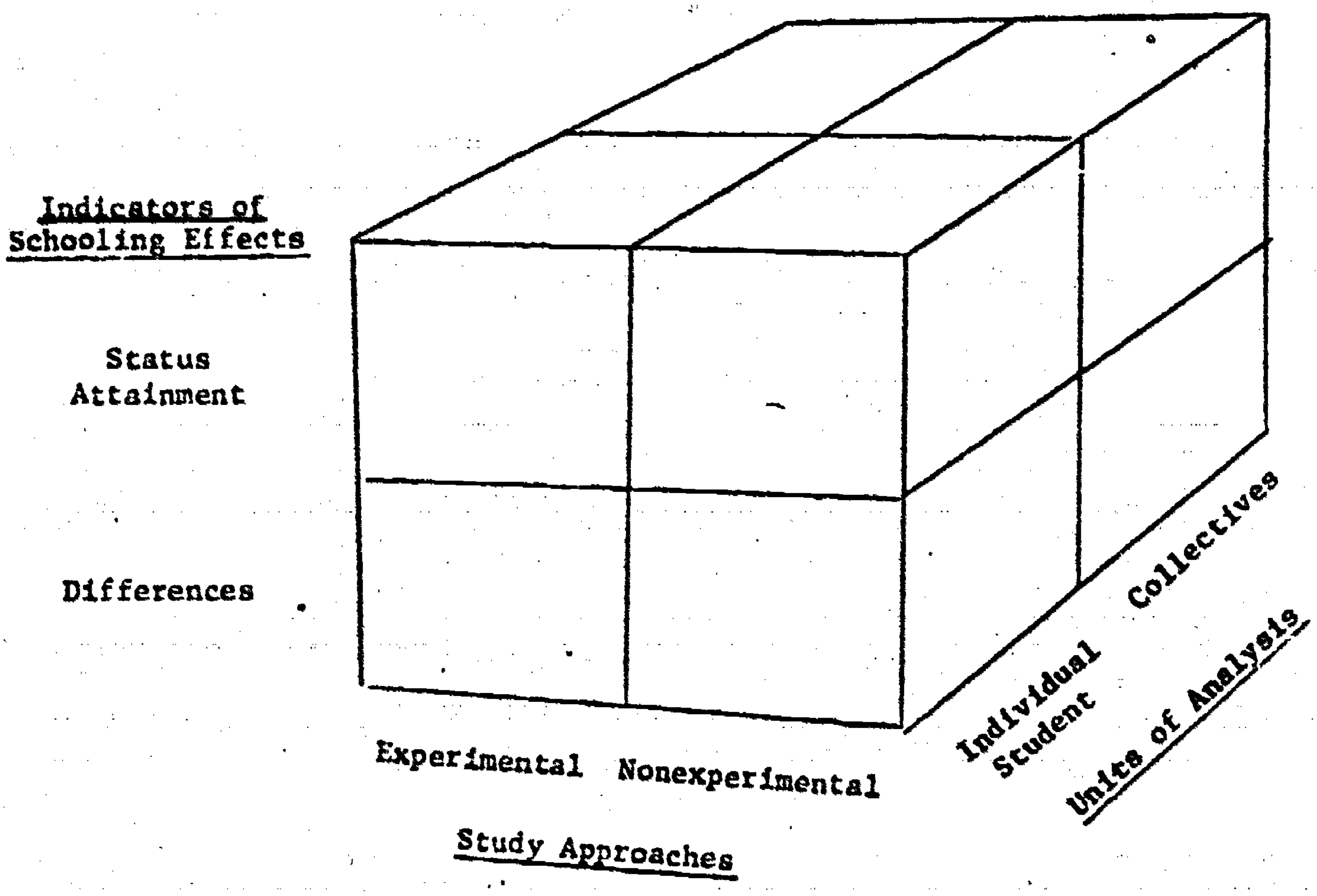

Figure 1. Organizational schema for reviewing research methodologies pertinent to the study of schooling effects.

schooling effects, or the "how?" dimension, are discussed in Section III. In Section IV are found discussions of issues and problems related to the un: of analysis and the analysis of multilevel data in studies of schooling effects, that is, the "which analytic unit?" dimension. A final section contains a review of methodological trends and a brief discussion of the implications that existing methodologies have for researchers in the design and conduct of schooling effects studies.

Search Delimitations

Porter and McDaniels (1974) have convincingly argued that design and measurement issues are equally as important to consider in schooling effects studies as they are in educational research in general. Despite this, a variety of practical constraints prevents the authors from dealing with both design and measurement issues in this paper. This is not meant to suggest that one area is more important than the other, but merely that the authors have chosen to focus on one and not the other. The present review, therefore, concentrates almost exclusively upon issues of design and statistical analysis.

In addition, the amount of literature related to research methodologies was so great that it became absolutely necessary not only to limit this review to the domain of interest (as defined by Figure 1), but also to be highly selective in its treatment. Some of the issues discussed in the later sections are briefly described in the paragraphs below and are accompanied by remarks of a delimiting nature.

Indicators of schooling effects. The question of what it is that is to be measured in studies concerned with schooling effects can be dealt with from at least two aspects. One might for example ask: What are schooling effects? It is the authors' intent to leave the resolution of such political and philosophical problems to others; nevertheless, it is necessary to point out that schooling effects studies have mainly been concerned with examining the effects of schooling on "immediate" student outcomes, such as student achievement in the basic skills areas. This historical concern is reflected in this paper. A second question that can be asked from a methodological standpoint is: What kinds of measurements should be used as indicators of schooling effects? This paper examines two major indicators of schooling effects: (1) status attainment or outcomes (i.e., effects measured at a single point in time), and (2) differences (i.e., effects resulting from differences between measurements occurring at two points in time or between observed and predicted outcomes).

Study approaches in research on schooling effects. Perhaps the most crucial methodological question is: How should schooling effects be studied? In responding to this issue we distinguish, as did Cronbach (1957), between two major study approaches: the experimental and the correlational (or nonexperimental). Discussed under the experimental category (which encompasses pre-experimental, true experimental, and quasi-experimental designs) are "experimental-group-only designs" and "control-group designs." A variety of correlational tvpe approaches to research is discussed under the nonexperimental category. Also discussed briefly are efforts to combine the experimental and nonexperimental approaches.

The descriptive case study approach will be excluded from this partic lar review since ethnographic methodology suited for the study of schooling effects is so different from the main thrust of this paper that its inclusion may be distractive more than beneficial.

Units of analysis. Methodological issues related to the units of analysis (and levels of data aggregation) facet are presented by contrasting two types of analytic units; that is, units at the individual student (or noncollective) level and units at the collective level (e.g., classrooms, schools, school districts). Additional discussion deals with the analysis of multilevel data. Not discussed are methodologies used for analyzing data from studies involving single students or groups (i.e., n = 1, or one-shot case studies).

Search Strategy

In keeping with the need to economize time and resources for this investigation, the literature search began with an examination of relevant articles appearing in recent volumes of the Review of Educational Research and the Review of Research in Education. This initial step resulted in a list of methodological topics from which an outline was generated. Key papers and reports related to each topical area were then identified, obtained, reviewed, and annotated. Treated as "key literature" were those papers or reports: (1) in which methods or ideas listed on the topical outline were originally proposed, (2) that were related to methods used in major and/or controversial studies, or (3) that suggested new directions and approaches. Additional references were examined depending upon the nature of their citation in the key literature and upon the recommendations of a panel of external reviewers. A final reorganization of the topical outline was helped by comments from this same panel of experts.

II. ISSUES AND PROBLEMS RELATED TO INDICATORS OF SCHOOLING EFFECTS

Research can be construed as the process of seeking relationships between variables (Gage, 1976a). This, of course, is also true of schooling effects research. All such studies, by definition, are concerned with the exploration of relationships between independent and dependent variables which represent, respectively, the "effectants" and the "effects" of schooling. From the measurement point of view, one can distinguish between two major types of dependent variables: "status attainment or outcomes" and "differences." Status attainment or outcomes indicators are measures of effect taken at a particular moment in time. Differences indicators or scores representing degrees of discrepancies between two measurements on the same scale can be subdivided into scores derived from differences between measures taken at two different points in time and scores representing differences between observed and predicted outcomes. Issues and problems related to these indicators of schooling effects are discussed below. A final section summarizes this knowledge base and reviews the advantages and disadvantages of both types of indicators.

Status Attainment or Outcomes as Indicators of Schooling Effects

All measurements taken on a student at a single point in time are to be regarded as indicators of status attainment or outcomes.

A major class of schooling effects studies that employs indicators of status attainment or outcomes is educational assessment programs. Programs or studies such as the IEA studies (e.g., Comber & Keeves, 1973; Purves, 1973; Thorndike, 1973), the National Assessment of Educational Progress program (e.g., NAEP, 1974), and the Educational Quality Assessment or EQA program (e.g., Pennsylvania Department of Education, 1973) use status attainment data.

Current efforts to establish minimal competency levels as a prerequisite to graduation (Madaus & Airasian, 1977; Pipho, 1977) represent another example of the use of status attainment indicators to assess schooling effects. A more complex example is the traditional dependenee on class standing, grade point average, and scores on entrance examination as the basis for admission to college.

Finally, studies of long-term schooling effects on such noncognitive variables as occupational status and income have also involved the collection of status attainment data (e.g., Fägerlind, 1975; Flanagan & Cooley, 1966; Flanagan, Dailey, Shaycoft, Gorham, Orr, & Goldberg, 1964; Jencks et al., 1972).

Difference Scores as Indicators of Schooling Effects

All scores representing degrees of discrepancies between two measurements on the same scale are to be regarded as indicators of differences. Two major types of difference indicators, that is, change or gain scores and residual scores, are discussed in this section.

Change as Indicators of Schooling Effects

In educational research, schooling effects frequently are evaluated on the basis of the amount of change in those observable student behaviors thought attributable to the school, program, teacher, or some combination thereof (e.g., McDonald & Elias, 1976). Almost all of the existing schooling effects studies have been designed to examine the *immediate* effects of schooling and, for purposes of analysis, have utilized adjusted or unadjusted change scores derived from calculating differences in pretest to posttest performance (Type A change scores). In contrast, the few studies that have researched *after-school* effects (or what Härnqvist (1977) refers to as the "enduring effects of schooling") have, again for purposes of analysis, utilized change scores but as derived from calculating differences in posttest-1 to posttest-2 performance (Type B change scores). Type A change scores measure learning (and/or "growth"); Type B change scores assess retention (and/or "conceptual modification").

Discussions below are concerned with issues and problems traditionally related to Type A change scores and will follow quite closely in topical coverage Linn and Slinde's (1977) comprehensive review article entitled, "The Determination of the Significance of Change Between Pre- and Posttesting Periods." The two types of change scores reviewed here include: first, "raw change" (or "difference") scores and, second, "estimated true change." These discussions are followed by a short section dealing with Type B change score issues.

Raw change scores. The simple raw change or difference score is "the most natural measure of change from one point in time to another" (Linn & Slinde, 1977, p. 122). Swimmers would be interested in assessing the difference between pretraining speed and speed after some amount of training; golfers, on the other hand, would be concerned with the number of strokes they were able to take off their game as a consequence of being coached by the club professional. The raw change score obviously is quite easy to calculate, but this simplicity belies the methodological complexities associated with its use.

There are several major areas of concern related to raw change scores. One serious problem with the use of raw change scores is that they typically are negatively correlated with the pretest (Bereiter, 1963; Linn & Slinde, 1977; Thorndike, 1966). This dependency relationship is more commonly referred to as the problem of regression toward the mean (Guilford, 1954; Herriott & Muse, 1973; Marcus, Keesling, Rose, & Trent, 1972). An implication of this problem is that students with low pretest scores are more likely to obtain large positive gains, while students with high pretest scores are less likely to do the same and perhaps show a loss.

Bereiter (1963) correctly indicates that "some kind of correction is called for" (p. 3). However, he also notes that attempts to correct for the regression toward the mean effect have led to what he calls the "under-correction/over-correction dilemma." He referred to Garside's (1956) article in which three methods of solving for the regression of gains on initial scores were studied. Garside's results were inconsistent; that is, with one method the regression estimate increased as the correlation between pretest and posttest increased, with another it decreased, and the third method was indifferent to this correlation. In another instance, Campbell and Erlebacher (1970) succinctly illustrated how biased adjustments could make the gains for one group look larger in relation to gains for other groups. Such results, they suggest, usually will occur when groups are constituted in such a way that the pretest scores for the groups are significantly different from one another, as is the case in many quasi-experimental studies.

It can be noted that researchers generally agree that none of the offered alternatives made to correct for biases resulting from regression effects provide a fully satisfactory solution to the problem (e.g.; Cronbach & Furby, 1970; Linn & Slinde, 1977).

Another problem with raw change scores is unreliability (Bereiter, 1963; Linn & Slinde, 1977; Lord, 1963). Linn and Slinde have illustrated vividly that the reliability of raw change scores is a function both of the reliability of the pretest and posttest and of their intercorrelation. Raw change score reliability increases as the reliability of the pretest and posttest increases, but decreases as their intercorrelation increases.

Linn and Slinde (1977) indicate that one implication of the unreliability of raw change or "difference" scores is that "it is quite risky to make any important decisions about individuals on the basis of gains from pre- to posttesting periods" (p. 124). To determine the trust one should have in raw change scores and, thereby, to reduce risks, Lord (1963) has recommended the computation of the correlation between observed change and true change, or between estimated true change and true change, or both. He indicated that this estimate should be calculated prior to analysis proper to be sure that the observed change scores are not simply the result of random fluctuations or so obscured by random fluctuations as not to be worthy of analysis.

Bereiter (1963), in an attempt to improve the reliability of the raw change score, introduced the "change item" concept and procedure. The change item was defined "as an item that is administered to the same person on two occasions and scored directly for direction and perhaps amount of change" (p. 10). The procedure produces, as a by-product, a lowered intercorrelation between the pretest and posttest while perhaps even raising the reliability of each. It follows that a possible outcome of using this procedure would be an increase in the reliability of the raw change score.

The very notion of increasing the reliability of the raw change score by decreasing the intercorrelation of the pretest and posttest raises another issue that Bereiter refers to as the "unreliability-invalidity dilemma." The dilemma posits that an increase in the reliability of the raw change score brought about by a decrease in the pretest-posttest intercorrelation also tends to lower the validity of the measure itself; that is, because of low intercorrelation, the same instrument administered as a pretest and posttest may be said to be measuring different things. Despite the above, Bereiter believes that the use of the change item practice is an admissible one for increasing the reliability of raw change scores.

Two other issues which are corollaries to the ıaw change score problems mentioned above seem worthy of note. First, the correlation of a raw change score with another variable that is in part a function of the pretest or posttest is, because the same errors of measurement are present in both quantities being correlated, usually considered <u>spurious</u> (Lord, 1963). When raw change scores are correlated with the pretest, a spurious negative correlation usually is obtained.

Second, unreliability has the effect of <u>attenuating</u> correlations (Lord, 1963). The implication of this is that correlations involving a raw change score having low reliability will tend to be quite low. Linn and Slinde (1977) noted that this is rather a discouraging implication for educational researchers interested in finding correlates of change.

<u>Estimated true change scores</u>. An alternative approach to that of the raw change score is to estimate "true" change, that is, the change that would obtain if there were no error of measurement (Linn & Slinde, 1977). As conceived of by Lord (1956, 1958, 1963) and by McNemar (1958), true change may be estimated by using multiple regression procedures based on estimates of reliabilities of the

pretest and posttest, their variances, and their covariance. The Lord-McNemar argument was extended by Cronbach and Furby (1970) in an attempt "to get a still better estimate" (p. 68). By distinguishing, as did Stanley (1967), between independent and linked measures (i.e., ones with correlated errors) and by suggesting the use of other available measures as predictors, Cronbach and Furby substantially advanced methodological theory in this area.

Cronbach and Furby suggested that a more precise estimate of the true score could be obtained by adding one or more available measures to the least squares estimation. In a study of this issue, Marks and Martin (1973) found that the precision of an extended pretest estimator of true change is an increasing function of the correlation between true change and the true score on the additional measures. More recently, Tatsuoka (1975) decomposed the squared multiple-R of the least-square estimate of true-score difference into the reliability of the difference score and the increment due to other predictors, which is always non-negative. Therefore, adding predictors increases the precision of estimation.

The distinction made by Cronbach and Furby between linked and independent measures led to the development of different formulas for estimating the reliabilities of raw change scores and true change. The formulas likewise require that a distinction be made between linked and independent pretest-posttest correlations. In a study of Cronbach and Furby's reasoning, Marks and Martin (1973) found that, as predicted, the magnitude of the correlations between true change and pretest true scores had a pronounced effect upon the precision of true change estimation. They also noted an analogue to Bereiter's (1963) unreliability-invalidity dilemma in respect to true change estimation. It was their suggestion that "as a general rule of thumb, the investigator computing true gain estimates should employ only test forms with reliabilities in excess of .85 and especially so if the true gain-initial true score correlation is expected or found to be .70 or less" (p. 190).

An estimated true change score has some advantages over a raw change score. For one, the reliability of the estimated true change score is as large as or larger than the reliability of a raw change score (Tatsuoka, 1975). In addition, Lord (1963) has empirically shown that when estimated true change scores are used in lieu of raw change scores, persons with relatively high pretest scores are more likely to be among those with large gains. The estimated true change scores, therefore, obviate the objection that raw change scores tend to favor persons with low pretest scores (i.e., the regression effect).

The Enduring Effects of Schooling

Only a small number of studies have been concerned with Type B change scores and even fewer with "after-schooling" effects in the cognitive domain (e.g., Dähllof, 1960; Härnqvist, 1968). Härnqvist (1977), while rightfully indicating that this is a neglected area in educational research, also cautions the researcher against the use of repeated measurements:

> 1. It is not easy to retrieve information even if it is there, somewhere in the long-term store. In a long-postponed measurement of retention, more and different types of cues are likely to be needed, and therefore a repeated measurement with the same instrument . . . directly after learning is not very informative or fair.
>
> 2. Since information is not just stored away until it is retrieved, but undergoes qualitative changes in the meantime, other things are likely to come out from the store than those originally put in, and such changes are not just distortions by a faulty memory but might very well be improvements also.
>
> 3. For both reasons a quantitative measurement of gains and losses over time is likely to be misleading. Only on a superficial operational level is there a difference between two comparable things. (p. 9)

Residuals as Indicators of Schooling Effects

The residual score, obtained by subtracting the predicted criterion score from the corresponding observed score (DuBois, 1957), has been widely used in recent schooling effect studies (e.g., McDonald & Elias, 1976; Soar, 1973). Residualizing removes from the criterion score the portion that could have been predicted linearly from predictors or covariates. The residual score, therefore, has a zero correlation with the covariate and consequently does not give an advantage to persons with certain values on the covariate measures (Linn & Slinde, 1977).

Residual scores. To avoid confusion, one should distinguish between two types of residual scores that differ according to the nature of the predictors used in computing predicted criterion scores. In one case, predictors are obtained from measures other than criterion measures, and in the other case, the same measures are repeatedly used in determining both the predictor and criterion measures (i.e., pretest-posttest). The latter type of score is often called the "residual gain score." Cronbach and Furby (1970) have the opinion that the residual score is not a corrected measure of gain. It is, they say, "primarily a way of singling out individuals who changed more (or less) than expected" (p. 24).

The first type of residual score was used in the schooling effects studies of Dyer (1970) and Astin and Panos (1966). In contrast, Soar (1973) and McDonald and Elias (1976) used the so-called "residmal gain score" in their investigations of process-product relations.

With residual scores the effects of covariates have been partialled out from the criterion variables, yet the residual score still has the same unreliability problem as does the raw change. Linn and Slinde (1977) showed that residual score reliability was a function of the reliability of pretest and posttest scores and of their intercorrelations. Although the reliability coefficients of residual scores are somewhat better than those of the corresponding raw change scores, they were still small whenever the correlation of pretest and posttest scores was large. The same cautions, therefore, that held for the unreliability of raw change scores must also apply to residual scores. And, since the problem of unreliability prevails with the residual scores, researchers are warned to correct for attenuation when computing partial correlations as well (Bereiter, 1963; Linn & Werts, 1973).

True residual gain scores. It has been noted that raw change or gain is to true gain as residual gain is to true residual gain. This relationship was used by Tucker, Damarin, and Messick (1966) in their attempt to draw attention to the "true residual gain" score which they referred to as a "basefree measure of change." They proposed to divide the true gain score into two components, one entirely dependant on the true score of the first, or baseline test, and one entirely independent of it (i.e., a true predicted gain and a true residual gain). Tucker et al. (1966) developed equations for estimating both of these components. However, Cronbach and Furby (1970) correctly criticized their proposals and in the process demonstrated a better way to estimate the true residual gain.

The Advantages and Disadvantages of Status Attainment and Differences as Indicators of Schooling Effects

Arguments for and against the use of status attainment or different indicators in schooling effects studies are many and varied. And, as is often the case, arguments for one are based on arguments against the other. For example, in opposing the use of change measures, Cronbach and Furby (1970) suggested that if one is testing the null hypothesis that two treatments have the same effect, the essential question is whether posttest average scores (i.e., status attainment or outcome scores) vary from group to group. They found no occasion in which the change score should be estimated in educational research and concluded: "It appears that investigators who ask questions regarding gain scores would ordinarily be better advised to frame their question in other ways" (p. 57). Linn and Slinde (1977) concurred: "The virtues in doing so [in measuring change] are hard to find. Major disadvantages in use of change scores are that they tend to conceal conceptual difficulties and they can give misleading results" (p. 147). Linn and Slinde, following Cronbach and Furby, further recommended that "where appropriate, regression analyses that treat the pretest no differently from other independent variables (or predictors), and the posttest as the dependent variable, avoid many of the difficulties that are introduced by gain scores" (p. 148).

In contrast, some methodologists hold the opinion that the task for the researcher is not to eliminate the use of differences as indicators of effect, but rather to find ways to minimize the problems their use creates. Bereiter (1963), for one, has described a number of ways by which problems associated with difference scores can be reduced. It is his argument that: (1) unreliability in pretest scores should be corrected before posttest scores are regressed on pretest scores, (2) the meaningfulness of change scores does not depend on a test's measuring "the same thing" on two occasions, and (3) measuring changes directly as subjective dimensions which do not necessarily have underlying physical continua is the only way that permits interpretable comparisons between changes on psychological dimensions for individuals with different initial standings.

In attempts to minimize further problems with difference indicators, researchers have developed better ways of estimating true change and true residual change (e.g., Cronbach & Furby, 1970; O'Connor, 1972) and have found that the use of group membership information treated as a dummy variable in a regression analysis improves the fairness of the estimators (Lord, 1963; McNemar, 1975).

In the opinion of the authors of this review, it is not yet possible to study the schooling effects that concern educators most without sometimes resorting to the analysis of differences. Despite their limitations, there are specifiable conditions when the analysis of difference scores is more appropriate than the analysis of status attainment or outcome scores. The remainder of this section is used to describe the conditions of use for the two major categories of indicators of effect and to discuss the relative advantages and disadvantages of each.

Status Attainment or Outcome Indicators of Effect

In schooling effects studies, measurements taken at a single point in time, usually after some intervention has taken place, are referred to as status attainment or outcome indicators.

Conditions of use. Status attainment or outcome measurements are appropriate indicators of schooling effects when: (1) initial student differences are expected to have little or no bearing on later status or outcomes (e.g., studies of long-term effects and mastery programs); (2) initial student differences on crucial variables have been controlled (for example, by random assignment of students to treatment conditions); and (3) there is no intent to attribute schooling effects specifically to schools, programs, and/or teachers (e.g., state assessment programs).

Relative advantages. Status attainment or outcome measurements generally are easier to collect, store, and process in respect to other types of measurements. If initial student differences on crucial variables are satisfactorily controlled, then the test of differences between treatment conditions on posttest indicators is straightforward and easy to interpret.

Relative disadvantages. In most educational settings, it is difficult to control for initial student differences, and as a consequence, it is not possible to attribute meaningful schooling effects specifically to schools, treatments, and/or teachers. This is an especially sensitive issue for those interested in short-term process-product research. Randomization, frankly, does not always work, and even when it does, there is no guarantee that selective attrition may not occur later on to bias the results.

Status attainment or outcome measurements are also affected adversely by test unreliability. The most effective corrective procedure is to select reliable measures in advance.

Difference Indicators of Effect

In schooling effects studies, adjusted or unadjusted scores representing differences between measurements taken at two points in time, with the interval usually being filled by some intervention strategy, are referred to as "change" or "gain" indicators. Adjusted or unadjusted scores representing differences between predicted and observed measurements are referred to as "residual" indicators. Change and residual indicators have been treated in this paper as subcategories of "differences" indicators and, in the discussion to follow, will be distinguished only when warranted.

Conditions of use. The analysis of differences is appropriate when the researcher anticipates that initial student differences cannot be fully controlled and will thus influence outcomes in ways that will prevent a clear interpretation of causality. The influence of initial student differences on such short-term effects as reading and mathematics achievement is usually considerable and serves as an example of the appropriate use of differences as indicators of effect. Difference indicators may also be appropriately used in relatively long-term studies. In this regard, Stallings and Kaskowitz (1974) report that the degree of the regression effects can drastically be reduced in a longitudinal study.

<u>Relative advantages</u>. The use of difference scores permits researchers to statistically control (as best they may) initial student differences.

<u>Relative disadvantages</u>. Because differences scores are derived from two potentially fallible measurements, they are usually unreliable. When differences between repeated measurements are calculated, they are likely to be affected by the phenomenon known as regression toward the mean. In attempting to correct for the effects of regression toward the mean, the researcher will usually be faced with the so-called over-correction/under-correction dilemma. This problem was discussed earlier.

In addition to the above, difference scores are more difficult and costly to come by. And, while the raw gain score is relatively easy to calculate, the remaining types of difference scores are much more difficult to derive.

<u>Summary</u>. The authors' review of the literature indicates that although there are considered objections to the use of differences as indicators of schooling effects, there are conditions under which they are more or less appropriate to analyze. It is also the case that recent developments in this area have tended to reduce the force of earlier objections.

In short, certain forms of adjusted difference scores seem to be appropriate indicators in the study of schooling effects, especially for the study of such relatively short-term effects as student achievement in reading and mathematics.

III. ISSUES AND PROBLEMS RELATED TO STUDY APPROACHES IN RESEARCH ON SCHOOLING EFFECTS

Cronbach (1957), in his presidential address to the American Psychological Association, indicated that there were two historic streams of method, thought, and affiliation in psychology: experimental and correlational psychology. These same study approaches have been evident in educational research over the years and still represent the predominant methodologies in investigations of schooling effects (Alwin, 1976; Gage, 1976a; Herriott & Muse, 1973). This section will review issues and problems related to study approaches in research on schooling effects; specifically, the "experimental" and "nonexperimental" approaches to the study of schooling effects will be discussed below. Also discussed in a subsequent subsection are two relatively new methodological developments in educational research; that is, aptitude-treatment interactions and meta analysis. A final subsection is devoted to a summary of review findings in this area.

Experimental Approaches

Experimental approaches to educational research are characterized by attempts to manipulate experimental variables while tightly controlling relevant situational variables. True experimental designs permit researchers to perform rigorous tests of hypotheses and to reject those hypotheses that are less tenable. In these designs, the random assignment of experimental units to treatment and control conditions is used as a mean of attaining initial group equivalency on crucial variables. However, experience suggests that it is almost impossible to assign randomly individual students to treatment and control conditions in most educational situations. Also, in the natural setting of the classroom or school the researcher seldom has full control over situational and/or experimental variables. Under such conditions, researchers may use alternative designs, that is, quasi-experimental designs.

Campbell and Stanley (1963) distinguish between three sets of experimental designs: pre-experimental, true experimental, and quasi-experimental. Identified as pre-experimental designs are the: (1) one-shot case study, (2) one-group pretest-posttest design, and (3) static-group comparision. True experimental designs include the: (1) pretest-posttest control group design, (2) Solomon four-group design, and (3) posttest-only control group design. The ten designs classified as quasi-experimental include the: (1) time series, (2) equivalent time samples design, (3) equivalent materials samples design, (4) non-equivalent control group design, (5) counter-balanced design, (6) separate-sample pretest-posttest design, (7) separate-sample pretest-posttest control group design, (8) multiple time series, (9) institutional cycle design, and (10) regression discontinuity.

With two exceptions, the above listed designs can be assigned to one of two major categories distinguished by the number of groups involved in the study, namely, one-group designs (i.e., experimental group only) or multiple group designs (i.e., control groups). Listed in the upper portion of Table 1 are designs classified according to this subdivision of the study approaches dimension and according to the indicators of schooling effects dimension (i.e., status attainment or differences). This latter dimension was depicted earlier in Figure 1.

Campbell and Stanley discussed the strengths and weaknesses of each of the sixteen designs in terms of internal validity (i.e., interpretability) and external validity (i.e., generalizability). The eight factors jeopardizing internal validity are: (1) history, (2) maturation, (3) testing, (4) instrumentation, (5) statistical regression, (6) selection, (7) mortality, and (8) selection-maturation interaction. The factors jeopardizing external validity are: (1) interaction effects of testing, (2) interaction effects of selection and treatment, (3) reacting effects of experimental arrangements, and (4) multiple-treatment interference. Readers are advised to refer to Campbell and Stanley (1963) for a full discussion of the strengths and weaknesses of each design. The strengths and weaknesses of the experimental group only designs and of control group designs are, however, reviewed briefly below.

Experimental Group Only Designs

Pre-experimental designs such as the one-shot case study and the one-group pretest-posttest design and quasi-experimental designs

Table 1

Research Designs and Analytical Models for Schooling Effects Studies

STUDY APPROACHES		INDICATORS OF SCHOOLING EFFECTS	
		STATUS ATTAINMENT OR OUTCOMES	DIFFERENCES
Experimental Approaches*	Control Group Design	.Static-Group Comparison (t-test) .Posttest-Only Control Group Design (t-test, ANOVA with blocking, ANCOVA) .Equivalent Materials Design (ANOVA)	.Pretest-Posttest Control Group Design (ANCOVA, Repeated Measures ANOVA) .Solomon Four-Group Design (2 x 2 ANOVA of posttests) .Nonequivalent Control Group Design (Regression-Discontinuity Model, Regression Projection Model, Generalized Multiple Regression Model) .Counterbalanced Designs (Latin-square ANOVA) .Separate-Sample Pretest-Posttest Control Group Design (ANOVA) .Multiple Time Series Design
	Experimental-Group-Only Design	.One-Shot Case Study (Common-knowledge Comparisons)	.One-Group Pretest-Posttest Designs (Norm-referenced Comparisons) .Time Series .Equivalent Time-Samples Designs (Nested design ANOVA) .Separate-Sample Pretest-Posttest Design (t-test)
Nonexperimental Approaches	Non-linear Regression Methods	.Polynomial regression analysis	.Polynomial regression analysis
	Partitioning of Explained variance	.Incremental Partitioning of Variance .Commonality Analysis	.Incremental Partitioning of Variance .Commonality Analysis .Partitioning Residual Criterion Variance
	Regression Coefficient Methods	.Multiple regression analysis ~ b and beta coefficients	.Multiple regression analysis - b and beta coefficients
	Causal Models	.Recursiva Model .Nonrecursive Models	.Recursive Models .Nonrecursive Models

*Fourteen out of the 16 designs in Campbell and Stanley (1963) are classified into this table. The Recurrent Institutional Cycle Design (A "Patch-Up" Design) is not included because it is of little utility. The Regression - Discontinuity Analysis is an analytical model rather than a design. It is listed as an analytical model for the Nonequivalent Control Group Design.

such as time series, equivalent time samples design, and equivalent materials samples design are classified here as experimental group only designs, simply because they lack a control group.

The two pre-experimental designs classified among experimental group only designs happen to be the weakest designs among the sixteen listed by Campbell and Stanley on both internal and external validity criteria. They lack control over almost all sources of invalidity. The one-group pretest-posttest design is, however, free from selection and mortality biases. If in a study involving a one-group pretest-posttest design a standardized test is administered around the norming date (Tallmadge & Horst, 1976), then the norm-group comparisons design, which controls for overall sources of internal invalidity, can be used to evaluate the school effects.

Three quasi-experimental designs classified as experimental group only designs(that is, time series, equivalent time samples design, and equivalent materials samples designs) have control over most sources of internal invalidity. However, these designs still lack control over sources of external invalidity.

Control Group Designs

Ten of the sixteen listed by Campbell and Stanley involve at least one control or comparison group and may be classified as control group designs. A control group in research designs tends to reduce (or control) confounding effects from such sources as history, maturation, testing, instrumentation, and regression. By using such complicated quasi-experimental designs as the separate-sample pretest-posttest design and the separate-sample pretest-posttest control group design, researchers may increa^ the external validity of findings.

Of some interest is Ary and Carlson's (1975) flowchart for selecting Campbell and Stanley designs, which takes into account threats both to internal and external validity. It is a helpful aid for the novice researcher in deciding upon the most appropriate strategy for a given research effort.

Experimental Approaches in Schooling Effects Studies

Inspection reveals that only five of thirteen possible study types identified by Stufflebeam and Webster (1978) employ experimental or quasi-experimental designs as methods for evaluating educational programs. The study types are: (1) public relations inspired studies, (2) experimental research studies, (3) policy studies, (4) decision-oriented studies, and (5) consumer-oriented studies. Two other experimental approaches may be added to this list: preplanned variation evaluation studies (Light & Smith, 1971) and a procedure proposed by Tatsuoka (1972) for evaluating nationwide intervention programs.

From an examination of Stufflebeam's (1971) CIPP Evaluation Model, one may infer that true-experimental designs have limited utility in educational evaluation. Difficulties in meeting assumptions of constancy of experimental treatment across both subjects and time and the inability to assign students randomly to experimental and control groups provide ample reasons for this view (Stufflebeam, 1969, 1971).

Tatsuoka (1972), however, has an altogether different viewpoint. Relative to the constancy requirement in experimental treatments, Tatsuoka observed:

> An educational program is, by its very nature, an entity that is in perpetual flux. This fluid, dynamic entity, with all its periodic modifications and refinements *is* the treatment. Nothing in experimental design forbids such types of treatment. (p. 3)

Tatsuoka admitted that under the present educational system random assignment of individual students to treatment and control conditions is difficult. Nevertheless, since students, in his view, are not appropriate units to study in large scale program evaluations, the problem is not a real one. He argued that "classes, schools, or even school districts are the proper units, and random assignment of these to the conditions is not nearly so infeasible as that of students" (p. 2).

Many researchers engaged in nonexperimental classroom process studies and teacher effectiveness studies (e.g., Brophy & Evertson, 1974) admit the need for experimental studies to test hypotheses generated via correlational studies. A general consensus among educational researchers is that hypotheses derived from educational theories and instructional models regarding relationships between student achievement and contextual and instructional process variables should be verified via experimental approaches.

Cronbach (1957) has noted that a distinctive characteristic of modern experimentation is the statistical comparison of effects. The early development of techniques in comparative experimentation is succinctly documented by Cochran (1976). Analytical methods are described in many reference sources such as Edwards (1972), Hays (1963), Kirk (1968), and Winer (1971), among others. Multivariate versions of statistical comparison are described in Anderson (1958), Bock (1975), Cooley and Lohnes (1971), Finn (1974), Tatsuoka (1971), Timm (1975), and elsewhere.

Tatsuoka and Tiedeman (1963) developed a schema for presenting statistical techniques in relation to educational research based on the role (i.e., dependent or independent), number (i.e., one or more than one), and scale-type (i.e., nominal, ordinal, or interval) of variables involved. Among the listed statistical techniques are multiple regression, analysis of variance and covariance, and such non-parametric statistics as the sign test, median test, Mann-Whitney's U test, Kruskal-Wallis one way-ANOVA, Friedman's two-way ANOVA, Chi-square test, Hotelling's T^2, McNemar's test for significance of changes, and Cochran's W test for several related proportions. These represent most of the methods that can be used in testing statistical hypotheses (usually null hypotheses) in an experimental approach. Their schema provides researchers with a reference point in selecting appropriate statistical analysis methods.

Another practical guide, advanced by Tallmadge and Horst (1975), listed five evaluation models named after appropriate analytical models: (1) posttest comparison with matched groups, (2) covariance analysis, (3) special regression, (4) generalized regression, and (5) norm-referenced. A decision tree constructed to aid in the selection of the most appropriate model for the conditions of the proposed evaluation is provided.

Tallmadge and Horst discussed the strengths and weaknesses of each model and provided an analytical method for testing statistical significance of the difference between experimental and control group mean scores. They also advanced the notion of educational significance, even though it remained a subjective criterion. These authors suggest that "if the observed posttest scores exceed the no-treatment expectation by one-third of a standard deviation, the treatment effect be considered educationally significant" (p. 69).

Nonexperimental, Correlational Approaches

According to Cronbach (1957), correlational approaches to educational research are intended for the study of natural relationships. While experimenters are interested only in the variation they themselves create, correlators are interested in already existing variation between individuals, social groups, and species. It is the correlators' mission to observe and organize data from nature's experiments and in the process to describe the ways by which variables covary. Thus, for example, researchers using statistical devices such as correlational coefficients can study the ways in which teacher behaviors are related to student outcomes on reading and mathematics tests of achievement. Such relationships may be found to be positive, neutral, or negative, and linear or nonlinear.

The correlator has access to a variety of correlational methods, and most of these have been described by Tatsuoka and Tiedeman (1963). A table listing these statistical techniques classified according to the role, scale type, and a number of variables involved has also been developed by these researchers. Listed on their table are methods ranging from the contingency coefficient "C" to canonical correlation.

The set of correlational techniques described in Table 1 of this section is not intended to cover all of the methods dealt with by Tatsuoka add Tiedeman. In fact, it is limited to regression techniques, associated with Pearson's product-moment correlation coefficient "r."

Correlational methods used in studies of educational effects are grouped into four categories in Table 1 and include: (1) partitioning of explained variance, (2) comparison of regression coefficients, (3) nonlinear regression methods, and (4) causal models. These four categories, which are discussed below, although not mutally exclusive, do differ in the method of correlational analysis used (usually regression analysis) and in their emphasis on different statistics obtained from the analysis.

Partitioning of Explained Variance

In regression models, the square of Pearson's product-moment correlation "r^2" is interpreted as the proportion of variance in the dependent variable that is accounted for by the independent variable. The analogue to r^2 in cases of multiple independent variables is R^2, the squared multiple correlation. When an R^2 is obtained in experimental research with balanced designs where predictors are independent from each other, the R^2 is equal to the sum of the squared zero-order correlations between each predictor and the criterion variable. Under such conditions, there is no ambiguity as to the amount of variance accounted for by a given predictor (Darlington, 1968).

In nonexperimental research, however, the predictors are almost always intercorrelated. The major sources of controversies with respect to studies of schooling effects include various attempts to partition variance and thereby to attribute specific portions of it to specific predictors.

Incremental partitioning of variance. One way of partitioning variance is to examine the increment in the proportion of variance accounted for by each predictor as it is entered into a regression analysis. This method was used in the Coleman Study (Coleman et al., 1966) and in a series of IEA studies (e.g., Comber & Keeves, 1973; Purves, 1973; Thorndike, 1973).

Coleman and his associates regressed student achievement scores on student background characteristics such as home SES and school resources. It is the case that when predictors are intercorrelated, the increment in variance attributed to a given predictor is determined, in part, by its order of entry in the analysis; in other words, the incremental variance is asymmetrical. In the Coleman study, the student background characteristics were entered into the analysis first and this accounted for a large amount of variance, leaving the effects of school factors negligible. In rationalizing this procedure, Coleman and his associates argued that since student background characteristics are "prior to school influence, and shape the child before he reaches school, they will be controlled when examining the effects of school factors" (p. 198). Pedhazur (1975), however, argues that it is not a sufficient justification to control one variable merely because it precedes another predictor.

Darlington (1968) has discussed the use of various general regression procedures, including the incremental partitioning of variance, indicating they are valid when predictors are mutually orthogonal but quite dubious otherwise. Creager (1971) has proposed the use of a complete orthogonal factor analysis for orthogonal decomposition of the regression system that would result in orthogonal components that are still interpretable in terms of the original variables.

Commonality analysis. As a solution for the asymmetry problem involved in the incremental partitioning of variance, commonality analysis, as developed by Mood (1969, 1971) and by Newton and Spurrell (1967), partitions the explained variance in the criterion variable that may be attributed uniquely to each of the predictors and the variance that is to be attributed to various combinations of predictors. The unique contribution of a given predictor is the increment in the proportion of variance in the dependent variable for which it accounts when entered last into the regression analysis. The unique contribution is the same as the squared part correlation a criterion variable with a predictor partialed on all other predictors in the regression equations. This method was extensively used in the reanalysis of the Coleman study data by Mayeske et al. (1969). In that reanalysis, the variance in the criterion variable was partitioned into the following three major portions: (1) that portion uniquely accounted for by student background factors, (2) that portion uniquely accounted for by school variables, and (3) that portion accounted for by the combination of student background and school variables.

Werts (1968) advocated the use of commonality analysis instead of the incremental partitioning of variance for studying schooling effects. According to Pedhazur (1975), it has a utility viewed from a predictive frame of reference. In other words, commonality analysis can be used to determine which variable may be deleted with a minimal reduction in the total proportion of variance. In fact, Newton and Spurrell (1967) recommended commonality analysis specifically for such a purpose. Despite the above, stepwise regression analysis represents a more effective way to reduce the number of predictors without affecting greatly the total proportion of variance.

Viewed from an explanatory frame of reference, commonality analysis has very limited value. Pedhazur has suggested that "it might even be argued that by its very nature it evades the problem of explanation, or, at the very least, fails to come to grips with it" (p. 254). Creager (1971), for one, called attention to difficulties in interpreting the variance accounted for by a combination of predictors. He indicated that two variables may be highly correlated because one of them is the cause of the other, or because they both share a common cause. Commonality analysis is unable to distinguish between the two. Thus, it is the case that the uniqueness and commonality elements are affected by the introduction of additional variables or by the deletion of variables, when the predictors are intercorrelated.

Another difficulty with commonaltiy analysis is that commonality elements may have negative signs. Ward (1969) has indicated that commonality elements may have negative signs when suppressor variables are involved and that as a consequence the sum of the unique contribution of the predictors may then exceed 100 percent. The former problem is not solved by arguing, as Mayeske et al. (1969) did, that: "Negative commonalities will be regarded as equivalent to zero" (p. 49). The solution for the latter problem should wait until the former is resolved.

It should be noted that a multiple dependent variable version of the partitioning variance method has been proposed by Lohnes and Cooley (1976).

Partitioning residual criterion variance. In the incremental partitioning of variance and also in computing the uniqueness of a predictor in commonality analysis, the effects of all predictors that precede it have been partialed out. Some researchers (e.g., Astin & Panos, 1966; Dyer, 1970) partition the residual criterion variance obtained by regressing the criterion or output variable (e.g., achievement) on the input variables (e.g., home SES, pretest scores). There is no difference in the prime analysis procedure between this method and the two variance partitioning methods discussed earlier. The difference is that criterion variables are first residualized on some predictors or input variables and then the resulting residual variance is used in partitioning.

In a series of college input studies, Astin (1970a, 1970b) and his associates used an input-output model which involved a two-step procedure for calculating a part correlation. In this procedure the input variation was used to residualize the output variable. The residualized student output variable was then correlated with the college environment variables.

In Dyer's (1970) student change model in an educational system, the performance indicator of a school system is derived from the residual output score of the system which was obtained from a regression analysis using the input and "hard to change" variables as predictors. After the performance indicators of educational systems are obtained, they are studied in relation to the "easy to change" surrounding conditions and the school process variables.

Among many problems related to the partitioning residual criterion variance, unreliability of change or "gain" scores, including residual scores, is the most serious one. Although Dyer's model uses school means rather than individual student scores, the reliability of the residuals still may be questionable. In Dyers, Linn, and Patton's (1969) cross-validation study, school residuals showed reasonable stability across subsamples. Marco's (1974) study also showed that the reliabilities of both individual and school residual scores were relatively stable in cross-validation. Forsyth (1973), however, reported that school residuals were unstable over time. Thus, it appears that the residuals may be relatively stable from one subsample of students to another within a single year, but relatively unstable from one year to the next.

Problems involved in the partitioning of variance when the predictors are intercorrelated are also relevant to this approach (Darlington, 1968).

Comparison of Regression Coefficients

The Coleman Study, which used a variance partitioning approach, was criticized not only for its validity but also for its usefulness as a guide for policy decisions (e.g., Bowles & Levin, 1968; Cain & Watts, 1970; Hanushek & Kain, 1972). These critics of the study argued that the proportions of variance accounted for by a given predictor and by certain combinations of predictors would not, in general, provide any guidance for policymakers to decide what course of action should be taken to increase student achievement. Consequently, they advocated comparing regression coefficients, a method whose purpose is to assess the effects of each predictor on the criterion variable. These same critics indicated that they preferred regression coefficients to percentages of explained variance as estimators of school effectiveness.

Unstandardized and standardized regression coefficients. In the following linear, additive model regression equation

$$Y' = a + b_1X_1 + b_2X_2 + \ldots + b_pX_p \quad ,$$

the "b" weights should be treated as partial regression coefficients. One interprets "b" as indicating the expected change in the criterion variable "Y" for a unit change in predictor "X" (with which it is associated), while holding all other predictors in the equation constant. Such an interpretation of the b weights is said, however, to be valid only in experimental research. Michelson (1970), for one, has indicated that it is incorrect to interpret a regression coefficient obtained from nonexperimental research as the expected change in the criterion variable resulting from a unit change in the predictor, while holding all other predictors constant. Mosteller and Moynihan (1972) noted that:

> We can estimate the difference in achievement between schools not having and those having a language laboratory, say. But we cannot tell whether actually adding or removing a language laboratory would produce nearly the same differences. (p. 35)

In the standardized expression of the regression equation,

$$Z'_y = \beta_1 Z_1 + \beta_2 Z_2 + \ldots + \beta_p Z_p ,$$

the standardized regression weights, betas, are scale-free indices and thus can be compared across different predictors. In spite of this advantage, some researchers (e.g., Cain & Watts, 1970; Linn, Werts, & Tucker, 1971) prefer unstandardized coefficients. The main reason for this seems to be that the β's are affected by the variability of the variables within a specific population being studied, while the b's remain fairly stable despite differences in the variability of the predictors in different samples (Blalock, 1964). There are problems, however, in interpreting unstandardized coefficients in schooling effects studies. For one, the magnitude of b's depends on the units used in the measurement of a given predictor (e.g., cents or dollars), and many of those measures are not interval variables (e.g., attitude). Smith (1972) reanalyzed the Coleman Study data and found numerous examples in which comparisons based on β's or b's led to contradictory conclusions in respect to the relative importance of the same predictors. Smith recommended that β's should be used when comparing coefficients for several predictors within a given sample, while b's should be used when coefficients associated with a given predictor are compared across samples.

Analysis of interaction effects. As a way of studying the interaction or joint influence of predictors on a criterion variable, researchers enter into a regression analysis the product terms for the predictors. Anderson (1970) used this technique in his study of the effects of classroom social climate on learning. He found that four out of fourteen subscores from the Learning Environment Inventory showed statistically significant interaction effects with intelligence for girls. In another instance, Cronbach (1968) reanalyzed some of the Wallach and Kogan (1965) data and questioned conclusions found in their article. Cronbach used the incremental variance partitioning method and reported a total of seven statistically significant increments added by the interaction of intelligence and creativity.

Pedhazur (1975) observed that the value of the concept of interaction (or the nonadditivity issue) is dubious. He noted that

> Attempts to interpret a regression coefficient for a cross-product vector in the conventional manner create an illogical situation in that one is led to state that the coefficient indicates the expected change in Y associated with a unit change in the cross-product vector while holding constant all other variables including those from which the cross-product vector was generated. (p. 265)

It is also to be noted that the coefficients for cross-product vectors are affected by, among other things, changes in the means of the predictors from which they are generated (Darlington & Rom, 1972).

Zero-Order Correlations and Nonlinear Regression Analyses

In correlational studies, it is traditional to investigate the linearity of the regression lines at the zero-order correlation level before conducting further analyses. If upon inspection the regression line appears to be nonlinear, an appropriate transformation is recommended in order that linear assumptions be met.

Recently many researchers, working in the field of teacher effectiveness studies and classroom instructional variables, have shown strong interest in the study of zero-order correlations (e.g., Brophy & Evertson, 1974; Rosenshine, 1976, Soar & Soar, 1973). Soar and Soar (1976) reported findings that were not only interesting to consider but were also consistent throughout four of their studies. One of these findings was that of a nonlinear relationship, most likely of an inverted "U" shape, between student gain in achievement and a measure of teacher behavior. In general terms, the inverted "U" suggests that student achievement is maximized with relatively moderate amounts of certain teacher behaviors and that extremes of the behavior, in either direction, tend to lead to reduction in student achievement. Another finding, that of the differentiated "U", suggests that different kinds of pupil learning varied in respect to teacher behaviors associated with greatest pupil gain. Brophy and Evertson (1974) and Brophy (1978) have also reported some nonlinear distributions.

The statistical procedure most widely used in detecting the nature of nonlinear relationships is the polynomial regression analysis in which powers of variables are introduced in the regression analysis. Cronbach (1976) made cautionary remarks against blind search for nonlinear relationships: "Nonlinearities may reasonably be explored, but unless there is a rationale for predicting nonlinearity, little credence can be given a nonlinear relationship the first time it turns up" (p. 3.11).

Polynomial regression analysis has problems, too. First of all, the nonlinear regression analysis can make a greater contribution in an explanatory framework. However, it is difficult to interpret the regression coefficients, even if they are unstandardized ones, in the prediction framework. In addition, it is not legitimate to test the significance of regression coefficients individually or to test intermediate regression coefficients for the purpose of deleting those that do not reach a prespecified level of significance. All a researcher can do is to test coefficients successively for the purpose of determining the pattern of the regression (Williams, 1959).

Causal Models

Another approach to the study of schooling effects is causal modeling in general and path analysis in particular. The technique of path analysis was developed by Wright (1921) more than a half century ago, but has not been widely used by educational researchers (Tatsuoka, 1973). Blalock (1964) and Duncan (1966) introduced this technique to sociology in the 1960s; Werts and Linn (1970b) were the researchers who introduced it in education.

Path analysis is an analytic tool for theory testing. In order to apply it, the researcher has to make explicit the theoretical framework within which he/she operates. In fact, the application of incremental partitioning of variance implicitly requires researchers to formulate a causal model for specifying relations among variables under study. In path analysis, causal models should be explicitly expressed, for example, in path diagrams.

There are many ways to formulate a causal model, particularly when the causes are unknown and/or unobserved. Wright has noted that "in cases in which the causal relations are uncertain, the method can be used to find the logical consequences of any particular hypothesis in regard to them" (p. 557). This suggests that researchers need to formulate not only one but many models and must test each one to determine if they are tenable or not. One of the main advantages of path analysis is that it enables the researcher to measure the direct and indirect effects one variable has upon another. In addition, it enables researchers to decompose the correlation between any two variables into a sum of simple and compound paths.

Two types of causal analysis models, that is, the recursive and nonrecursive models, can be distinguished. Issues and methods related to the analysis of data within the context of each of these models are examined next.

Recursive models. In recursive models, the hypothesized causal relations among variables are unidirectional; that is, if X is a cause of Y, then Y cannot be a cause of X. Simon's (1954) analysis, for example, started from the bivariate case, and then moved to a three-variable situation in which the basic concern was whether the observed correlation between two variables were spurious due to the presence of a third variable. Blalock (1964) has expanded the work of Simon and has developed a technique to test for the existence of linkages between variables in recursive models of any size.

Path coefficients in recursive models are usually obtained by ordinary regression techniques, which comply with regression assumptions and covariance restrictions that often lead to over-identification problems (Asher, 1976). Among the assumptions that underlie the application of the recursive path analysis, in addition to the one-way causal flow assumptions, are that: (1) the relations among the variables in the model must be linear, additive, and causal; (2) the residuals cannot be correlated among themselves, nor correlated with the variables in the system; and (3) the variables must be measured on an interval scale (Kerlinger & Pedhazur, 1973). Unmet assumptions might lead to sizable standard errors of the regression coefficients and the path coefficients. The problem of multicollinearity also arises with the use of the path analysis technique.

Comber and Keeves (1973) and Werts and Linn (1970b) used miniature recursive models of educational effects for illustrative purposes. It was McDonald and Elias (1976) who actually used the path analysis technique in their educational effects study. Anderson and Evans (1974) used the recursive models to reanalyze data from two studies that appeared in the literature. Magidson (1977) applied this approach to the Cicirelli et al. (1969) Head Start data and found small positive estimates of the program effects which were originally judged to be totally ineffective. Interestingly, Cicirelli et al. (1969) had earlier stated: "Results from the summer program are so negative that it is doubtful that any change in design would reverse the findings" (p. 245).

Nonrecursive models. In contrast to the recursive causal models, nonrecursive models involve interdependence, feedback, and reciprocal causation among at least some of the variables. The controversy between Jencks et al. (1972) and Smith (1972) regarding the causal flow between parents' expectations and student achievement in the Coleman study would have been settled had a nonrecursive model been employed.

One advantage of nonrecursive models is that they do not require the assumption that the residuals be uncorrelated. While this leads to a gain in realism, it brings about problems in the level of identification. When an equation is underidentified, there is no exact solution that gives satisfactory estimates. In those cases, the indirect least squares solution is to be used. In this procedure, either certain coefficients are assumed to be zero or other exogenous variables are introduced to the model. On the other hand, in a nonrecursive just-identified equation or a nonrecursive overidentified equation a two-stage least squares solution is usually used.

The use of nonrecursive models requires a high degree of theoretical and methodological conceptualization. In the field of educational research, the studies that used nonrecursive models are rare. Using the Coleman study data, Levin (1970) postulated a nonrecursive model and applied the two-stage least squares method in the estimation procedure. However, his attempt was primarily designed to serve as an illustration. Anderson (1978) also provided an empirical example of nonrecursive-type analysis using data from the Evans and Anderson (1973) study.

When all has been said and considered, path analysis has a potential to serve as a strong heuristic tool for the development of theories of education. Tatsuoka (1973) has recommended its greater uses in educational research.

Other Developments in Study Approaches

Two relatively new methodological developments in educational research are worthy of note. The first is the attempt to unify the experimental and correlational research traditions. The second is the effort to test results across studies for overall significance. Both of these developments are discussed below.

Aptitude Treatment Interaction

As part of his APA presidential address, Cronbach (1957) urged that the two major disciplines in psychology, that is, the experimental and correlational, be unified. In effect, he was proposing the study of "aptitude-treatment interactions." Almost twenty years later, Cronbach (1975) indicated "that hybrid discipline is now flourishing" (p. 116). At the same time he admitted that he and others had been thwarted by inconsistent findings from roughly similar inquiries. He indicated that it might be more fruitful to shift emphasis and study higher order interactions as well as first-order ones. Recently, Cronbach and Snow (1977), in a highly regarded book, synthesized research in this area.

Krus and Krus (1978) have observed a reluctance among psychologists to unify the discipline and remarked that:

> The present schism between experimentalists and correlationists seems to be due not to a different language, but to different levels of language in use with the general linear model. The

> experimentalists appear to stress computational operations on raw scores (e.g., sum of squares) while correlationists seem to prefer heuristic explanations at the standard score level (e.g., variance). (p. 120)

As Cronbach and Snow concluded, more time is needed to achieve a satisfactory level of unification or extention of the two disciplines. Krus and Krus observe that in recent years some experimentalists have gradually turned to regression methods, especially in cases in which they were pursuing interactions and more complex hypotheses. Based on these observations they concluded:

> When one considers the gulf separating these two disciplines a decade ago, the overall, integrative power and conceptual advantages of general regression theory seem to indicate that Cronbach's original vision of the unified discipline of scientific psychology is perhaps in the offing. (p. 123)

Meta Analysis

Another old but recently revitalized effort to synthesize the results of independent studies can be found among researchers in the fields of teaching and learning (e.g., Gage, 1976a; Glass, 1976; Rosenthal, 1978). The proposed method, referred to as a meta analysis, is described by Glass (1976) as "the analysis of analyses" (p. 3).

The need for the meta analysis of research seems to be obvious. In educational research, the findings vary in confusing irregularity across contexts, subject matters, and countless other factors. In order to design a study systematically on the basis of previous findings they first must be integrated in some fashion.

The origin of efforts to integrate findings from independent studies can be traced to Fisher (1948) and even to Pearson (1938). Since then, there has been a slow but steady increase in the amount of literature addressed to the question of how one may obtain an overall level of significance for results across studies. There have been some relatively recent attempts to integrate research findings by using a simple counting method; that is, by counting the number of studies reporting favorable or unfavorable outcomes (e.g., Bracht, 1970; Dunkin & Biddle, 1974; Jamison, Suppes & Wells, 1974; Light & Smith, 1971). However, this relatively simplistic method has not provided a satisfactory solution to the problem. More recently, Rosenthal (1978) has described methods for combining the probabilities obtained from two or more independent studies and provided a guide for selecting appropriate methods. Examples of the application of the meta analysis technique can be found in Gage (1976b), Glass (1976), and Smith and Glass (1977).

A Summary of Study Approaches

In line with existing divisions of research on schooling effects, the study approaches dimension was divided into two general categories: (1) the experimental (which includes pre-experimental, true-experimental, and quasi-experimental designs), and (2) the nonexperimental (which includes a variety of correlational techniques).

Experimental Approaches

For purposes of discussion the experimental category was subdivided into two components according to the presence/absence of control or comparison groups. The two sub-divisions were: (1) experimental group only designs, and (2) control or comparison group designs.

Experimental Group Only Designs. Experimental group only designs, which consist mostly of pre-experimental designs, suffer from the lack of internal validity (i.e., interpretability) and external validity (i.e., generalizability). Experimental group only designs of the quasi-experimental type, such as the time series, equivalent time samples design, and equivalent materials samples design, may provide more interpretable results than pre-experimental designs, but still lack external validity.

When standardized tests are appropriately administered on a pretest-posttest basis, a comparison with a norm group can be made even though there is no true control group.

Control or Comparison Group Designs. The most serious problem with control group designs is establishing equivalency between treatment and control groups on entry measures. Randomization is an essential (but not absolutely foolproof) manipulative procedure for establishing initial equivalency for both true and quasi-experimental designs. However, under most existing educational systems, it is extremely difficult for the researcher to arrange for the random assignment of individual students to experimental and control groups. It is less difficult to arrange for the random assignment of collectives (e.g., classrooms, schools, and school districts, etc.) to different treatment groups.

Even when the random assignment of collectives occurs, differences between treatment and control groups are not always completely eliminated. When initial differences are apparent, the use of statistical procedures that take initial differences into account are appropriate. Choosing, for example, between the multivariate analysis of covariance or the repeated measures design of the multivariate analysis of variance is a specific issue related to this problem area.

Nonexperimental Approaches

Again, for purposes of dicussion, the nonexperimental or correlational approaches were subdivided according to the nature of the coefficient calculated from a regression analyses. The four subdivisions were: (1) partitioning of explained variance, (2) comparison of regression coefficients, (3) causal models, and (4) nonlinaar regression methods. These subdivisions are not meant to be exhaustive, and, as they are based upon the same regression model, neither are they mutually exclusive.

Partitioning of explained variance. Two methods for the partitioning of explained variance (i.e., incremental analysis and commonality analysis) were discussed. In incremental analysis, the relative contribution of a given predictor is determined on the basis of the amount of increased variance accounted for when that predictor is entered into the regression equation. The results of incremental analysis are highly dependent on the order in which variables are entered into the equation. Consequently, when an underlying theory or hypothesis is controversial, the incremental analysis method cannot be employed to resolve the theoretical concerns.

Commonality analysis does offer a solution to problems associated with conflicting theories. In commonality analysis explained variance is partitioned into portions explained by each predictor and by combinations of predictors. Commonality analysis, therefore, provides results unaffected by the order by which variables are entered into the regression equation.

Comparison of regression coefficients. Many researchers regard the regression coefficient (either standardized or unstandardized) as more meaningful for policy-making than explained variance. Standardized coefficients are suitable for comparing the relative influence of each predictor within a sample, while unstandardized coefficients are useful for comparing the effect of a predictor across samples.

Causal models. Causal modeling, specifically path analysis, enables one to measure the direct and indirect effects that one variable has upon another. It also enables researchers to decompose the correlation between any two variables into a sum of simple and composed paths. It can be used, therefore, in testing theoretical hypotheses. Two types of causal models were considered: (1) recursive models in which the hypothesized causal relations among variables are unidirectional, and (2) nonrecursive models which include interdependence, feedback, and reciprocal causation among some of the variables. Nonrecursive models are more realistic than recursive models.

Nonlinear regression methods. Educational researchers interested in studying the relationships between classroom processes and student outcomes increasingly have attended to the issue of determining the true nature of the functions descriptive of the relationships. In particular, polynomial regression has been used in recent studies to identify nonlinear relationships.

Combined Study Approaches

A number of efforts to combine the experimental and correlational study approaches have been initiated. Cronbach was an early advocate for the unification of the experimental and correlational approaches to research. The aptitude x treatment interaction studies are what Cronbach has advocated. Some researchers see that these efforts have just been started. However, it seems to be fair to sav that a considerable progress is made in the area of aptitude x treatment interaction studies.

Meta Analysis

New methods for integrating results across studies have been developed and even utilized in a few studies. These methods will prove valuable in attempts to understand previous findings.

IV. ISSUES AND PROBLEMS RELATED TO THE UNIT OF ANALYSIS AND THE ANALYSIS OF MULTILEVEL DATA

Data collected at a given level, say, at the classroom level, may be aggregated to higher levels (e.g., the school or district level), or perhaps disaggregated to lower levels (i.e., the individual student level), or be retained and analyzed at the level at which it was originally collected. The analysis and interpretation of data that can be aggregated (or disaggregated) to different levels constitute an importtant methodological concern for the educational researcher who must select those "units of analysis" that are the most appropriate given the research question and other constraining factors. This selection problem is especially critical in large-scale studies of schooling effects, since multilevel data are collected in virtually all such investigations "simply because schools are, in part, aggregates of their teachers and pupils, and classrooms are aggregates of the processes and persons within them" (Burstein & Smith, 1977, p. 66).

When multilevel data have been collected, the researcher has the option either of aggregating data to higher levels or of disaggregating to lower levels or, more specifically, of using the collected data as proxies for lower level data. A seldom used third option is to engage in some form of multilevel analysis. Relative to this latter point, Burstein and Linn (1977) note:

> The effects of education exist in one form or another both between and within the unit at each level of educational systems. Yet the majority of studies of educational effects has restricted attention to either overall between-student, between-class, or between-school analyses. (p. 1)

One consequence of the above is a lack of consistency in findings across studies.

Sometimes, however, the researcher has a severely restricted set of options and must choose units of analysis at less than desirable aggregation levels. This is especially true when, say, data at the individual level may not be obtainable, or if obtainable, not identifiable for each individual (and, therefore, not relatable across data collected at different times or on different forms). Cost factors also may enter into decisions to select a particular unit for analysis and not others; it is normally cheaper to analyze data at higher levels of aggregation.

The selection of appropriate levels of analysis is not only an analytic concern; it relates also to the problems of interpretation, or more specifically of making inferences about relationships found at one level to relationships at other levels. This latter problem, known as the "fallacious inference" issue, is best understood by reviewing a study by Robinson (1950), who found that the size of correlations between illiteracy and race was a function of the units of analysis; .95 at the regional level, .77 at the state level, and .20 at the individual level. Had it been assumed that data aggregated to the regional level would provide the same information as data on the individual level, a fallacious inference surely would have been made (Alker, 1969).

Units of analysis issues are discussed in each of the following sections. The first deals with units of analysis as a general problem in educational research. The second section focuses on issues that must be considered when selecting appropriate units of analysis. A third section contains a discussion pertaining to the analysis of multilevel data in general. A brief summary is provided as a final section.

Units of Analysis Issues in Educational Research

The Coleman Report (Coleman et al., 1966), in which the school was treated as the unit of analysis, prompted educational researchers committed to the use of the student or classrooms as the unit of analysis to reexamine more closely issues and problems related to units of analysis. One such reexamination led Burstein and Linn (1977) to conclude: "efforts to identify the effects of education. . . on pupil performance have suffered from a lack of attention to the complications caused by the multilevel character of educational data" (p. 1). This is somewhat unsettling, since unit of analysis and data aggregation problems have long attracted the attention of other behavioral and social scientists. The study mentioned earlier by Robinson, a sociologist, is a case in point. In psychology, Estes (1956) argued that group learning curves said to show gradual learning may in reality be a composite of individual curves reflecting "sudden" learning. In economics, it was shown that the procedure of combining preferences or demand functions at the individual or family level was not useful in forecasting export and import demands at the national level (Scheuch, 1966).

The units of analysis issue was first publicly debated in education by Wiley, Bloom and Glaser (Wittrock & Wiley, 1970). It was Wiley's contention that "if the object of evaluation is a typical classroom instructional program where the instruction is received simultaneously by all students in the class, then the appropriate vehicle (or sampling unit) is the class and not the individual student" (p. 264). Bloom, and later

Glaser, argued that the unit of analysis should be individual students because it is students the school teaches and it is the effects on them that should be the focus of evaluation.

More recently, Brophy (1977) has argued that in schooling effects studies concerned with the nature of classroom interactions and the relationships between those interactions and student outcome measures, the student rather than the class mean should be the basic unit of analysis. He listed two reasons: (1) most teacher behavior directed to students is really directed at individuals rather than the whole class, and (2) even teacher behavior directed at the whole class interacts with individual student differences to determine outcomes.

There is agreement among some researchers that the end result of aggregation is a loss of information and the possible introduction of systematic error (e.g., Burstein, 1975; Hannan & Burstein, 1974; Haney, 1975). Haney (1974), using a small set of the Project Follow Through data, set about to demonstrate that the method by which data are aggregated (and, therefore, how units of analyses are formed) could affect analytic results. He reaggregated the Follow Through data into class and school-sized groupings using three different methods: (1) by random assignment, (2) by pretest scores, and (3) by posttest scores. In all three of these artificial groupings, correlations between pretest and posttest scores increased when data aggregated to the class and school levels were used. In contrast, correlations based on the aggregation of non-simulated data decreased. From these findings, Haney inferred that "when we aggregate data to the classroom level we confound all other causal variables of our outcome measure with classroom level effects" (p. 30). Haney issues the following warnings:

> The demonstration of its existence should make us highly wary of drawing inferences across different levels of analysis . . . just because variables have a particular relationship at the school level, is not sufficient reason to infer that the same relations hold at the class or individual level. Before we can make inferences across levels of analysis with any confidence, we must examine the aggregation relations and the potential manner in which they may artificially confound relationships between variables. (p. 31)

Grunfeld and Griliches (1960), however, suggest that aggregation in some cases may lead to a gain. And Hannan (1976) identified two special cases in which aggregation conceivably could lead to a gain: (1) aggregation that minimizes variation in confounding variables, and (2) aggregation by true scores.

Data from more recent studies of schooling effects are difficult to interpret because data collected mostly at the individual student level have been analyzed at higher levels of aggregation. For example, student data have been aggregated to the classroom level (e.g., Poynor, 1976; Soar, 1973; Walberg, 1969), the school level (e.g., Coleman, et al., 1966, Hanushek, 1968), the school district level (e.g., Kiesling, 1970; Bidwell & Kasarda, 1975), the state level (Walberg & Rasher, 1974), and even the national level (e.g., Bidwell, 1975; Comber & Keeves, 1973; Thorndike, 1973).

Selecting Units of Analysis

A crucial decision-point in the conduct of educational research is the selection of appropriate units of analysis. In making this selection "it is essential to have a clear picture of the spectrum of possible units, so that the choice based on the research problem may be a fruitful one" (Galtung, 1967, p. 37). In schooling effects studies, the spectrum of interest ranges from the individual student at the lowest level to perhaps the nation at the highest. In between these levels we find collectives (i.e., levels of aggregation) such as: small groups, classrooms, grades, schools, school districts, states, and regions. This section reviews issues that are to be considered in the process of selecting units of analysis that are the most appropriate given the purpose of the study and other constraining factors. Following Haney's (1974) format, major issues are discussed under the four following headings: (1) purpose of the study, (2) study design, (3) statistical considerations, and (4) practical considerations.

Study Purpose Considerations

To a great extent, the specific questions research seeks to answer dictate how a particular study will be conducted. It follows that the most basic consideration in the selection of a unit of analysis should be the purposes for which the study was undertaken. And while this is essentially true, it is also the case that "we cannot base our selection solely on the implications of the analysis questions" (Haney, 1976, p. 49). The determination of units of analysis, he says, is confounded with other issues, such as study design and data analysis issues, for example.

Studies differ in respect to the questions they attempt to answer and sometimes questions may properly be answered only by analyzing data at several different levels of aggregation. In the context of Project Follow Through, the questions that were to be answered reflected the need to analyze data at more than one level of aggregation. On the other hand, the questions posed by Brophy and Evertson (1977) could be answered best if the unit of analysis was the individual student.

Study Design Considerations

Study design, an issue overlapping with the purpose of the study, is another consideration when selecting a unit of analysis. According to Haney (1974), the three factors which give clues for selecting units of analysis are: (1) the units of treatment, (2) the independence of treatment units, and (3) the appropriate size for units.

Units of treatment. As a general rule, the unit of analysis should be the lowest level of aggregation at which units can receive different treatments or different replications of the same treatment (Cronbach, 1976; Glass & Stanley, 1970; Haney, 1974). At issue is how one determines the "unit of treatment." The sampling unit may be of some help in making such a determination (Burstein & Smith, 1977; Cline et al., 1974; Cronbach, 1976). For example, if randomization was used as a device for controlling initial differences, the units randomly assigned to treatment and control groups could be regarded as the unit of analysis (Haney, 1974). Sometimes, however, it is necessary to distinguish the sampling unit from the unit of treatment. Such a differentiation was needed in the Performance Contracting Experiment (Ray, 1972).

In this study, districts were sampled and one of only two schools selected from each wss assigned to the treatment condition. In this instance the district was the sampling unit, and the school was the unit of treatment and, hence, the appropriate unit of analysis.

Independence of treatment units. Another design consideration is statistical independence as addressed by Glass and Stanley (1970). According to the principle of independence, there should be no way in which the treatment applied to one unit should overlap or affect observations on another unit of treatment. If, for example, every student in the class correctly answers a question only because he/she earlier heard the teacher provide the answer to it when asked by student A, these responses are not independent; they are, in fact, dependent on the question asked by student A. Using simulated data, Glendening (1976) demonstrated that failure to meet assumptions of statistical independence of treatment when between-student analysis was performed could cause misleading results. In light of her findings, Glendening was forced to conclude:

> When dealing with educational data, in almost all cases, the group unit, such as classrooms, should be the unit of analysis. If, however, the data do happen to be independent of each other, it is clearly advantageous to use the individual unit as the unit of analysis. (p. 46)

On the other hand, Cronbach (1976) argues that "analysis at the level of the collective is likely to have no justification in science or policy studies unless the collective is in some real sense a carrier of an effect. He also indicates that "in educational research it does seem reasonable to think of classrooms and schools and districts as having real enough effects" (p. 1.19a).

It may be asked if it is feasible to choose a unit of analysis on the basis of a test for independence. In response, Glendening (1976) replied: "As a general rule of thumb a preliminary test of independence should not be used to choose a unit of analysis to test for treatment differences" (p. 48). The basis for choosing a unit of analysis should be the study design itself and careful observations to determine if the design was adhered to.

Appropriate size for units. A final design consideration with regard to choosing a unit of analysis concerns the "appropriate" size for the analysis unit. Given a limited amount of experimental material, the problem, which may not be solvable under actual research conditions, becomes one of determining the "unit size which will most reduce the variance of the estimated difference between two treatments" (Haney, 1974, p. 56).

Statistical Considerations

A variety of statistical issues must be considered in the process of selecting a unit of analysis. Not surprisingly, these considerations are related to the questions the analysis is intended to answer and its design. Haney (1974) has suggested three kinds of statistical considerations: (1) measurement reliability, (2) degrees of freedom, and (3) nonequivalence of treatment groups. Issues and problems pertinent to these topics are discussed in this section.

Measurement reliability. In Section II of this paper, the importance of measurement reliability was discussed in respect to measures of effects. If the research principle is accurate, the measures used to assess treatment effects must be reliable. A number of factors may affect the reliability of measures; among them is the level of aggregation. It has been generally known that measurement reliability increases as scores are aggregated to higher levels. Haney indicates, however, that this is true when the reliability coefficients are computed based on Shaycoft's (1962) model that does not account for group characteristics. When an alternative model proposed by Wiley (1970) was used to estimate the reliability of the same data, quite different coefficients were obtained. Haney's approach to the problem is somewhat unique, and is used also in identifying components of variance. He suggests that

> if a particular score has a relatively large component of variance between classes, then it makes sense to examine it using the class as a unit-of-analysis. Conversely, if there is little variance between classes ..., then it is less useful to perform a class level analysis. (p. 69)

Degrees of freedom. It is well known that as the degrees of freedom increase, the precision of an estimate improves, and equally important the "power" of the relevant statistical test increases. This understanding is an important statistical consideration in selecting a unit of analysis, as the degrees of freedom change from one level of analysis to another.

In this regard Emrick, Sorenson, and Stearns (1973) noted that "aggregating pupil level data to the classroom level . . . appears to shift evaluation focus from the individual child and to reduce statistical power and precision by decreasing observations" (p. A-5). Smith (1972), in contrast, employed the degrees of freedom argument to justify the use of classroom level data instead of the individual student data. "Had we used the child as the unit of analysis we would have been seriously overestimating the number of degrees of freedom available" (p. 108).

Haney (1974) believes that since the degrees of freedom issue has been used to argue for analysis at the lower and higher levels, this "cannot help but raise doubt about its validity" (p. 71). In respect to the issue of statistical significance, Haney concludes:

> The only time we ought to be concerned with degrees of freedom argument as it relates to statistical significance is when there are so few observations that we cannot distinguish statistical significance between effects estimates which are of a magnitude such that we would otherwise consider them educationally significant. (p. 72)

In respect to significance testing in nonexperimental situations, Haney concluded: "The degrees of freedom argument as a guide to the selection of a unit-of-analysis is . . . at best a heuristic one" (p. 74). When experimental designs are employed in schooling effects studies, the degrees of freedom argument should seriously be considered in deciding on units of analysis. This is the case because of known effects of units of analysis on the "power" of statistical tests.

Non-equivalence of treatment groups. Another statistical issue to be considered when deciding on units of analysis involves the method of adjusting for initial group differences and, by implication, disattenuation of the covariate. It may be remembered that Campbell and Erlebacher (1970) argued that the magnitude of adjustment bias will depend in part on the reliability of the covariate. Since covariate reliabilty is known to be affected by data aggregation, the relationship to the unit of analysis selection decisions becomes apparent. Cronbach (1976) has indicated that "the unit of analysis can make a difference in the estimate of a covariate-adjusted treatment mean, when persons or classes have not been assigned to treatments at random or when the number of independent assignments to treatment is small" (p. 1.3).

At issue is how the adjustments are to be made. Haney (1974) posits that "adjustment of posttest scores for pretest should be made at the pupil level prior to any aggregation to higher levels of analysis if the full effect of the adjustment is not to be lost" (p. 29). In an analysis of Follow-Through data, Abt Associates (Cline et al., 1974) adjusted for the fallibility of the pretest covariate at the individual student level only. They argued that classroom and school level data are much more "stable" and not in need of correction. Cronbach (1976), in contrast, holds the position that "group regressions may be just as fallible as individual ones" and, given this proposition, argues that "class and school analyses of covariance ought to be disattenuated when assignment is not random" (p. 1.8).

Practical Considerations

A number of practical considerations must also be given attention in selecting a unit of analysis. Haney (1974) assigned four issues to this category: (1) missing data, (2) policy research, (3) length of investigation, and (4) economy.

Missing data. In large-scale studies it is likely that some data will be missing; that students will have been absent in a particular data gathering period. There is, of course, no way that partial data can be used at the individual student level. With higher level units, partial data can be used. Dyer, Linn, and Patton's (1969) study indicates, however, that missing data may cause serious problems in obtaining discrepancy measures, even though data were analyzed at the school level.

Policy research. The purpose of policy research is that of improving policy rather than testing or improving theory. Given the above, Haney advocates the selection of a unit of analysis at a level "at which policy manipulable variables can best be studied" (p. 93).

Length of investigation. Evaluating an educational program over the course of years further complicates the unit of analysis issue. "The problem is that life of a classroom as a natural unit in most schools is only a single year" (Haney, 1974, p. 82). Under such conditions, it would be difficult to use the classroom as the unit of analysis in a multiyear analysis.

Economy. The final practical consideration is that of economy. Haney (1974) indicates:

> If a unit-of-analysis larger than the pupil is employed in an evaluation study, it is possible that a savings can be made by sampling only some of the sub-units within the desired units-of-analysis. (pp. 83-84)

In short, there is no simple way to select appropriate units of analysis. Indeed, some criteria discussed above may suggest directions that are in contradiction to one another. It is essential, therefore, that the researcher arrange these considerations in order of priority to optimize the selection of appropriate units of analysis.

Analyses of Multilevel Data

In schooling effects studies, it is not uncommon for researchers to have collected data at different levels or to have collected data that can be aggregated at different levels. This may represent an opportunity for researchers to analyze the data at multiple levels of aggregation (Burstein & Smith, 1977; Haney, 1974). Three different types of multiple level analyses can be discerned: (1) parallel analyses across levels of aggregation (Haney, 1976; Maw, 1976), (2) contextual analyses (Barton, 1970; Bowers, 1968), and (3) multilevel analyses (Burstein & Linn, 1976; Cronbach, 1976; Cronbach & Webb, 1975; Erlebacher, 1977; Keesling & Wiley, 1974).

Parallel Analyses

Multilevel data can be analyzed for each level of aggregation in such a way that only variables from the same level of aggregation enter into the analysis. When this type of single-level analysis is repeated at more than one level of aggregation, it is referred to as "parallel analysis." In the 1971-72 evaluation of Project Follow Through, Abt Associates employed single-level analyses at the student, class, and school levels; that is, they employed a parallel analysis strategy for data analysis. It is claimed that one advantage of parallel analysis is that it allows the researcher to study the consistency of results across levels of aggregation.

Contextual Analyses

A mixture of variables which represent a unit and those which represent the characteristics of its supra-unit can be used in an analysis, called contextual or compositional analysis, to study the effects of the supra-unit. For example, a mixture of student-level and school-level aggregates of student variables can be found in many schooling effects studies (e.g., Bowers, 1968; Coleman et al., 1966; Farkas, 1974). Coleman et al. (1966) found that certain contextual variables pertaining to characteristics of the student body explained additional variance in individual student achievement above and beyond that explained by the same characteristics at the individual level. Coleman and his associates argued that the academic climate of the school (i.e., contextual variables) has a direct influence on student performance.

Hauser (1970) opposed this kind of contextual interpretation and called it a contextual fallacy: "A not very distant cousin of the aggregative or ecological fallacy . . ., since both involve misinterpretation of the between group or ecological correlations" (p. 659). In the same article, he demonstrated a contrived contextual effect, namely, that educational aspiration of students rises as the proportion of males in a high school student body increases. He then demolished the claim for a contextual effect by reinterpreting the global sex ratio variable as a proxy for such variables as IQ and SES. The groups with high male-to-female ratios also were higher in the proportion of students with IQs and high SES.

Hauser's point is essentially concerned with "specification error." He noted:

> In a purely logical sense this objection can never be met because there are always "other" variables. From a practical standpoint, the objection means that one should be prepared to argue that his theory or relations among individual attributes is complete and correct, or at least defensible in relation to some explicit criterion, before speculating about residual group differences (p. 660).

Smith (1972), in a related study, included more background control variables in his reanalysis of the Coleman data and found no evidence "that characteristics of the student body have a strong independent influence on the verbal achievement of individual students" (p. 280). The results of Smith's reanalysis support Hauser's viewpoints.

Haney (1974) seems to be more cautious in dismissing the contextual effects. He notes, "Contextual effects may disappear when initial differences are fully controlled. Nevertheless, in a causal sense it is almost surely true that contextual effects are sometimes real" (p. 44). He continues:

> The obvious solution to this causal uncertainty is more powerful research designs than the non-experimental cross-sectional sort of design used in Project Follow Through or the Coleman study. Contextual analysis in non-experimental studies must be viewed with healthy skepticism. (p. 45)

Multilevel Analyses

The parallel analysis discussed earlier actually consists of two or more single-level analyses (e.g., between-student analysis, between-class analysis) with variables from a single level of aggregation involved. In the contextual analysis, variables from two or more different levels of aggregation are entered in a single analysis. Multilevel analyses are defined here as requiring analyses in at least two stages for at least two levels of units (Burstein & Linn, 1976). A few recently proposed models are reviewed below.

Between-group, pooled within-group analysis. Cronbach (1976) argues that overall between-student analyses are weighted averages of between-class and pooled within-class analyses and are rarely advisable in educational contexts. He notes that when heterogeneous within-class slopes that may reflect the teacher or treatment effects are present, the estimates based on the pooled within-class regression line probably are biased. Cronbach suggests the analysis of data at the classroom level (i.e., between-class analysis) and on the deviation from the class mean (i.e., within-class analysis). A pooled within group analysis on the deviation scores from mean is a feature of Cronbach's model, which distinguishes it from a parallel analysis.

Using this model, Cronbach and Webb (1975) reanalyzed Anderson's (1941) study, which reported finding an interaction of "drill vs. meaningful methods of arithmetic instruction" with student ability and achievement. In a reanalysis to separate between-class and within-class components of the outcome on an aptitude regression, the Aptitude x Treatment interaction(ATI) findings disappeared. Cronbach and Webb also applied the model to the Cooperative Reading Data (Bond & Dykstra, 1967) because of many reported instances of ATI. Again they found that conventional kinds of analyses (i.e., between students analyses) combine between-class and within-class effects in the sample and that some Aptitude x Treatment interactions disappeared when the effects were analyzed separately.

Using the same methodology, Rakow, Airasian, and Madaus (1978) reanalyzed data from American schools that had participated in the International Study of Achievement in Mathematics (Husen, 1967). Rakow et al. divided the within-school variation into two components, one associated with differences between mathematics teachers and the other with individual student differences. They found that "from thirty to forty percent of the within-school variation traditionally classified as individual student variance was associated with between-teacher performance differences within schools" (p. 19). These findings tend to support the further use of such types of analyses.

Regression analysis for hierarchical data. Keesling and Wiley (1974) argued that school-level indices, such as average daily attendance, do not convey independent information for each student within the school and thus should not be included in between-student analyses. At the same time, they indicated that the student-level data should be fitted at the level of the student within the school. The Keesling-Wiley analysis strategy includes: (1) a pooled within-school regression of outcomes on individual characteristics, (2) aggregation of predicted student outcomes over all students within a school, and (3) a between-school regression of school mean outcomes on school characteristics and school mean predicted outcomes.

Applying this method to the data from the Coleman Study, Keesling and Wiley showed that the estimation of the school input effects could be improved.

Analyses of slopes and intercepts. Burstein and Linn (1977) observed that "the variation of βj (specific within-class slope for class j) would become a potent source of information to researchers and policymakers, especially when such information is combined with the adjusted class effects" (p. 8). Their analytical strategy includes the estimation of specific within-class slopes and between-class regressions of class means and class slopes on teacher characteristics.

Using simulated data, Burstein and Linn studied the analytical consequences of heterogeneous, within-class regressions using different models, including their own, in education effects studies. A main conclusion was that neither student-level nor class-level analysis yielded correct estimates of teacher/class effects when there were systematic differences in within-class slopes that were determined by teacher quality.

Among the multilevel analysis models studied were Cronbach's between-class within-class analysis (Cronbach & Webb, 1975), the Keesling-Wiley analysis (Keesling & Wiley, 1974), and a slope-intercept analysis (Burstein, 1976; Burstein & Linn, 1976). These models yielded misleading estimates of the magnitude of teacher effects on mean class outcome; that is, all models tended to overestimate the direct effects of teacher quality when the "better" teachers had steep slopes, and tended to underestimate those effects when the "better" teachers had flat slopes. In addition, the Keesling-Wiley method showed an indication of bias in estimating teacher effects on mean class outcomes. All these results seem to justify Cronbach's (1976) caution about the possibility of developing a universally successful strategy.

At the conclusion of his review of the unit of analysis issue, Haney (1974) made the following recommendations:

> First, investigators ought to have a strong bias for studying various properties of the educational system at the level at which they occur.
>
> Second, variation in attributes of interest ought to be studied at those levels (or between those units) at which it does (or is expected to) occur. (p. 9)

Haney also advised researchers to make precise statements of the hypotheses to be tested (in terms of mathematical models, if possible), and to begin with strictly parallel analyses, if a researcher wants to conduct parallel analyses at different levels. Haney further urged researchers to treat classes and schools as units rather than as aggregates.

Summary

It may be said that there are two contrasting schools of thought relative to the units of analysis issue. One group of researchers holds the opinion that, in schooling effects studies, the appropriate unit of analysis is the individual student. This position is rationalized because actual learning occurs at the individual level. Another group argues that since educational treatments are normally administered at the system level, the collective (e.g., classrooms, schools, etc.) is the most appropriate unit of analysis.

A recently emerged position, held by a third group of researchers, suggests that, since student achievement can be influenced by factors existing at different levels of the educational system, data from schooling effects studies should be analyzed at multiple levels. The following three strategies for analyzing multilevel data were reviewed: (1) parallel analyses, (2) contextual analyses, and (3) multilevel analyses. An examination of these newly proposed techniques for the multilevel analyses revealed that they did not provide completely satisfactory results. Clearly, more research and development in this methodological area are required.

V. METHODOLOGICAL TRENDS AND THEIR IMPLICATIONS FOR RESEARCH ON SCHOOLING EFFECTS

In the process of examining research methodologies pertinent to studies of schooling effects, the authors noted that observations made by Dershimer and Iannaccone (1973), who earlier had examined social and political influences on educational research, overlapped with their own perceptions of trends in research methodology.

> A review of that literature points out that few scientific researchers, if any, select their problems at random. They are influenced by several factors, such as the "excitement of the chase," current scientific paradigms and theories, chance observations the scientists happen to have made, the dramatic nature of some phemonena, and the intellectual stimulation derived from work on complex tasks. Researchers are influenced by what their colleagues find important and vital; they respond to society's opinion of their work. They are sensitive to the interests of granting agencies or persons, and they are influenced by their institutions' support and provisions available for certain research tasks. (Dershimer & Iannaccone, 1973, p. 113)

From the authors' point of view, the single most important influence on trends in educational research methodology was, quite simply, federal dollars; it was not, however, the only influence. During the 1960s, events occurred that were to influence significantly the shape of American education and, to some extent, the methodologies used by educational researchers.

It would not now be in error to say, as does Mehan (1978), that "the most prevalent view in this country is that differences in scholastic and economic success are primarily the result of environmental influence rather than genetic endowment" (p. 33). Consequently, it must be difficult for some of us to comprehend why this view was not also prevalent in the very early 1960s. For example, Deutsch (1964) in a review of papers presented at a conference in the early 1960s on preschool enrichment observed:

> The overall impact of these papers and of their examination of the literature is to negate any concept of fixed intelligence [emphasis added] and to foster the belief that the human organism is highly malleable, particularly during its early years. (p. 208)

It probably was Hunt's (1964) book on Intelligence and Experience that first gave a measure of credence to this notion and in turn to the early intervention movement funded initially by private foundations such as Ford and Carnegie.

The fact that pupils in compensatory education programs made cognitive gains in excess of what was expected eventually got the attention of Congress. In the mid 1960s, Congress passed the Elementary and Secondary Education Act (ESEA) and in so doing brought to life first Headstart and later Follow Through. In its wisdom, Congress not only demanded that schools should be held accountable for the manner in which they spent monies, but also for the impact the school programs had on students. Pursuant to its accountability concern, Congress authorized a series of nationwide studies of programs funded by the federal government. The passing of ESEA legislation and the commissioning of a series of large-scale nationwide studies to assess the schooling effects of federally supported programs had a direct and irrepressible influence on educational research.

There was, however, one other important sociopolitical event that, in retrospect, influenced greatly research methodologies in the study of schooling effects. In 1964 the Civil Rights Act was passed and Congress commissioned James Coleman (Coleman et al., 1966) to document the suspected race-specific differences in the quality of public education (Shea, 1976).

In their attempt to respond to Congressional charges to study and evaluate the nation's schools and special programs, behavioral and social scientists came to realize that they lacked the methodological tools to carry out appropriately this important social task. This realization and the need to do something about it gave impetus to the use and refinement of methodologies seldom used for educational research and to the development of newer ones.

The following sections describe methodological trends in research on schooling effects, as perceived by the authors, in four topical areas: (1) study approaches, (2) independent variables, (3) indicators of effect, and (4) analysis of data. The implications of these trends for the conduct of future studies are also considered.

Trends in Study Approaches

Rosenshine and Furst (1973) introduced "a fairly complete paradigm for studying teaching" (p. 122), which corresponds fairly closely to the study approach dimension as presented in Figure 1 of this document. Their paradigm, which serves as a means of focusing the following discussion of trends, contains at least these elements:

1. development of procedures for describing teaching in a quantitative manner;

2. correlational studies in which the descriptive variables are related to measures of student growth;

3. experimental studies in which the significant variables obtained in the correlational studies are tested in a more controlled situation. (p. 122)

Prior to the 1960s, study approaches to research on schooling effects could be characterized as being almost exclusively limited in scope and oriented toward the comparison of two or more experimental units; that is, schooling effects research was essentially devoted to model building and hypothesis (null hypothesis) testing (Cronbach, 1975). During this period, true and quasi-experimental designs that were essentially univariate in character were used extensively in investigations (Campbell & Stanley, 1963). In terms of Rosenshine and Furst's (1973) descriptive-correlational-experimental loop, this period is demarcated by the "experimental" element.

The experimental approach to research on schooling effects continues and, without doubt, has been employed frequently since the beginnings of the 1960s. For example, a federal edict to ESEA Title I directors making them accountable for evaluating their programs actually led to an increase in the use of experimental type designs. However, most of the reports submitted were judged to be of inferior quality and as a result have contributed little to the schooling effects knowledge base. On the other hand, the work by Horst, Tallmadge, and Wood (1975) has improved methodology in this area. Of late, all levels of government are attempting to standardize, within relatively narrow limits, the experimental procedures that may be used in evaluating Title I programs (Tallmadge & Horst, 1975).

In the 1960s, the convergence of high-speed electronic data-processing equipment, advanced multivariate statistical software, and, perhaps, a "too rapid increase of federal support for research on education" (Howe, 1976, p. 46) led to a series of relatively large-scale, nonexperimental, multivariate studies. Some of these studies were initiated in response to the Congressional request for nationwide studies of federally funded educational programs. Among them were a series of studies on Follow Through (e.g., Soar, 1973; Stallings & Kaskowitz, 1974). Other studies initiated in response to the Civil Rights Act of 1964 included the study by Coleman et al. (1966) and its reanalyses by Jencks et al. (1972) and by Mayeske et al. (1969).

In addition, it is important to note that interest in such studies percolated down to state educational agencies, such as, California, which authorized, in conjunction with the National Institute of Education, several relatively large-scale nonexperimental studies as well (e.g., McDonald & Elias, 1976; Tikunoff, Berliner, & Rist, 1975). Other studies initiated at the state level include those of Brophy and Evertson (1975) and Soar and Soar (1973).

The large-scale nonexperimental (or correlational) approach to schooling effects has had an unprecedented effect on research methodology. The old adage suggesting that "necessity is the mother of invention" could never have been more true than during recent years. In attempting to answer pressing questions about schooling, nonexperimental study approaches have come of age. But since it is the expressed purpose of such studies to generate hypotheses for subsequent testing under experimental conditions, one may ask if large-scale experimental studies are far behind.

What about the descriptive element of the Rosenshine and Furst paradigm? Some interesting developments appear to be in the making. The Tikunoff et al. (1975) ethnographic study of a sample of teachers in the Beginning Teacher Evaluation Study (BTES) revealed, for example:

> that the methodology derived from sociology and anthropology . . . is promising for future research in teaching, particularly in identifying those classrooms where more effective teaching seems to be occurring. (p. 19)

Mehan (1978) in a discussion of nonexperimental methodology suggests that, "because it can address these problems, constitutive ethnography provides a rigorous methodological alternative to large-scale surveys as a means of guiding educational reform" (p. 62).

It would appear then that all three of the elements in the Rosenshine and Furst (1973) paradigm are actively employed and will become increasingly important.

Trends in Studying Independent Variables

One of the earliest large-scale input-output studies of schooling effects (Coleman et al., 1966) almost exclusively included independent variables far removed from classroom events (e.g., family background, age of school building, etc.). The major finding of the Coleman Report was that family background was a more important "determinant" of student achievement than such inputs as the quality of schooling. In effect, per pupil expenditures and school facilities were found to have little relationship to student achievement (Shea, 1976). In a study using simulated data modeled after the Coleman study (Mayeske et al., 1969) and Project Talent (Flanagan et al., 1964; Jencks & Brown, 1975), Luecke and McGinn (1975) indicated that their results;

> suggest that studies which find little or no relationship between educational inputs and achievement may be highly misleading. Our findings suggest that the combination of data and statistical technique [emphasis added] most often used is unlikely to reveal such relationships even when they exist. (p. 34)

They also observed that "researchers who conceive of education mechanistically, and use research designs which ignore the actions of individuals in schools, will find results which confirm their assumptions" (p. 348). Luecke and McGinn argued for a different category of input-type variables in schooling effects studies:

> For us, advancement will come through an improved understanding of what actually takes place in schools and classrooms. Studies using educational production functions must attend more to variables pertinent to the educational production process, and less to exogenous factors like family background. . . . this strategy may make it possible to discern the kinds of inputs that can make schools more effective institutions. We need to look more closely at what teachers, principals and superintendents do as they assign resources to students, teachers and schools, and to pay more attention to the direct effects of their actions. Perhaps research will then be more useful to those decision makers. (p. 348).

The Coleman Report and its offshoots (Jencks et al., 1972; Mayeske et al., 1969; Mosteller & Moynihan, 1972) also "tended to minimize" the role of the teacher in accounting for educational outcomes (Berliner, 1976). This finding stimulated a host of large-scale classroom process studies or process-product research (e.g., Brophy & Evertson, 1974; McDonald & Elias, 1976; Soar, 1973; Stallings & Kaskowitz, 1974; Tikunoff et al., 1975).

Brunswik (1956) presented a classification schema in which psychological variables were classified according to their remoteness from the central processes of the behaving organism. This schema is useful in understanding trends in selecting independent variables for schooling effects studies. Brunswik used the terms "central," "proximal," and "distal" to distinguish three broad regions of reference; "central" here refers to events within the organism, "proximal" refers to events at the interface between the organism and the environment, and "distal" suggests events with which the organism is not in direct contact, or over which the organism does not exercise immediate control (Snow, 1968).

Using this schema, trends in the selection of independent variables for large-scale studies appear to be moving from distal (e.g., Coleman et al., 1966) to essentially proximal-central variables (e.g., Brophy & Evertson, 1974; Soar, 1973; Tikunoff, Berliner, & Kist, 1975). The Stallings and Kaskowitz (1973) and the McDonald and Elias (1975) studies examined variables in all three regions (i.e., distal, proximal, and central variables).

From the perspective of the authors, it would appear that schooling effects studies are more and more using proximal-central variables, but not necessarily at the expense of distal ones. Within the central region, there is some indication of a shift toward a more detailed examination of the student behavior (McDonald & Elias, 1976; Tikunoff et al., 1975). In this latter respect, ethnographic techniques such as those used in the Tikunoff et al. (1975) study may prove quite useful.

Trends in Indicators of Effects

Since the 1960s, an increased use of all types of indicators of schooling effects is evident. Status attainment or outcome data were collected and analyzed for the Coleman Report (Coleman et al., 1966), for the National Assessment of Educational Progress (NAEP, 1974), and in a host of statewide assessment programs (e.g., Pennsylvania Department of Education, 1973). The continued use of status attainment data is expected, and its use should even increase as schools begin to establish minimum competency levels as the basis for granting certain diplomas.

Most large-scale short-term schooling effects studies employed some form of difference scores for analysis. For example, unadjusted change or "gain" scores were used in the McDonald and Elias (1976) study, and residual scores were used in studies by Soar (1973) and Stallings and Kaskowitz (1974).

Educational practitioners interested in determining the relationships between educational improvement efforts and short-term student achievement will find the residual score to be of use where initial student differences cannot be controlled.

Trends in Data Analysis

With the advent of modern electronic data processing systems and the development of increasingly sophisticated statistical software packages, there has been a clear tendency to employ multivariate analyses (Cooley, 1965; Tatsuoka, 1973). At the same time, with the realization that the relationship between certain classroom process variables and outcome variables may be nonlinear, there has been an increase in the examination of both linear and nonlinear bivariate relations or regressions (e.g., Brophy & Evertson, 1974; Soar, 1973).

Another important trend in schooling effects studies is the increasing tendency to analyze data at the individual student level. Perhaps, more important is the trend to employ multilevel analyses (e.g., Burstein, 1976; Cronbach & Webb, 1975).

The search for differentiated effects or interactions across different students, teachers, schools, and or programs is on the upswing (e.g., Brophy, 1977; Cronbach & Snow, 1977; Soar & Soar, 1975). However, in spite of more than a decade of research, there still are no consistent findings resulting from aptitude-treatment interaction studies (Cronbach, 1975). This seems to imply that further research is needed in the areas of higher-order interactions and/or differentiated nonlinear relationships. Another implication is that researchers need to conceptualize schemas by which the findings across studies can be synthesized (e.g., Medley, 1977) and areas that require further investigation identified.

Another important new trend is that of synthesizing the findings across studies using meta analysis techniques (e.g., Gage, 1976b; Class, 1976) so as to arrive at an overall index of, for example, program effectiveness. Meta analysis and the conceptual schema mentioned above represent extremely important methodological developments for researchers in their attempts to build comprehensive knowledge bases and construct new theories.

Summary

Prior to the 1960s, educational research on schooling effects could be characterized generally as limited in scope, devoted to model building and hypothesis testing (Cronbach, 1975), rarely including formal observations of the behavior of teachers when they taught or of pupils when they learned (Medley & Mitzel, 1963), univariate in approach (Kerlinger & Pedhazur, 1973), and dedicated to the quest for nomothetic theory (Cronbach, 1975). In short, it was an era during which the predominant methodological approach to the study of schooling effects was the small-scale nonprocess-oriented, essentially univariate experiment concerned with the discovery of universally applicable laws.

The 1960s represented a turning point in research on schooling effects. Spurred on particularly by the Coleman Report (Coleman et al., 1966) and by Congressional authorization to study Head Start and Follow Through on a nationwide basis, educational researchers reexamined closely their research methodologies. Since the late 1960s, the research on schooling effects receiving the most attention has been large-scale, multiregional (i.e., distal-proximal-central), multivariate, and non-experimental in focus. During this period, the unit of analysis has shifted from the school district to the classroom and individual student level, and, more importantly, to multilevel units.

This statement of trends should not be taken to imply that methodologies used to study schooling effects prior to the 1960s are no longer being employed; indeed, almost without exception, they exist side by side with current methodological innovations. On the whole, the authors were hard pressed to find examples in which established research methodologies were totally discarded in lieu of "innovative" procedures. Nor were many "new" methodologies discerned. However, methoaologies have changed; they have become more refined. Shulman's (1970) observations are relevant:

> The present era is one of significant methodological progress in the behavioral sciences and education. The development of new techniques, especially in the multivariate domain, proceeds at a rate which dazzles the non-specialist, even though in the eyes of the educational statistician, most of the "new developments" are merely variations on a few major themes. (p. 390)

Whether these methodological trends are regarded as a methodological advancement or as mere variations of existing methods depends upon one's point of view. Methodological trends, regardless of whether they are methodological advancements or refinements, seem to provide educational researchers with better perspectives on educational development.

REFERENCES

Alker, H. R., Jr. A typology of ecological fallacies. In M. Dogan & S. Rokkan (Eds.), Quantitative ecological analysis in the social sciences. Cambridge, Mass.: MIT Press, 1969.

Alwin, D. F. Assessing school effects: Some identities. Sociology of Education, 1976, 49, 294-303.

Anderson, G. J. Effects of classroom social climate on individual learning. American Educational Research Journal, 1970, 7, 135-152.

Anderson, G. L. A comparison of the outcomes of instruction under two theories of learning. Unpublished doctoral dissertation, University of Minnesota, 1941.

Anderson, J. G. Causal models in educational research: Nonrecursive models. American Educational Research Journal, 1978, 15, 81-97.

Anderson, J. G., & Evans, F. B. Causal models in educational research: Recursive models. American Educational Research Journal, 1974, 11, 29-39.

Anderson, T. W. An introduction to multivariate statistical analysis. New York: Wiley, 1958.

Ary, D., & Carlson, J. A. A flowchart for Campbell and Stanley designs. CEDR Quarterly, 1975, 8, 3-5.

Asher, H. B. Causal modeling. Beverly Hills, Calif.: Sage Publications, 1976.

Astin, A. W. The methodology of research on college impact, I. Sociology of Education, 1970, 43, 223-254. (a)

Astin, A. W. The methodology of research on college impact, II. Sociology of Education, 1970, 43, 437-450. (b)

Astin, A. W., & Panos, R. J. A national research data bank for higher education. Educational Records, 1966, 47, 5-17.

Averch, H., Carroll, S. J., Donaldson, T. S., Kiesling, H. J., & Pincus, J. How effective is schooling? A critical review and synthesis of research findings. Santa Monica, Calif.: Rand Corporation, 1972.

Barton, A. H. Comments on Hauser's "Context and Consex." American Journal of Sociology, 1970, 76, 514-517.

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

Berliner, D. C. A status report on the study of teacher effectiveness. Journal of Research in Science Teaching, 1976, 13, 369-382.

Bidwell, C. Nations, school districts and schools: Are there schooling effects anywhere? Vice-Presidential address, Division G, American Educational Research Association, Washington, D.C., 1975.

Bidwell, C. E., & Kasarda, J. D. School district organization and student achievement. American Sociological Review, 1975, 40, 55-70.

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

Bock, R. D. Multivariate statistical methods in behavioral research. New York: McCraw-Hill, 1975.

Bond, G., & Dykstra, R. The Cooperative Research Program in first grade reading instruction. Reading Research Quarterly, 1967, 2, 5-10.

Bowers, W. Normative constraints on deviant behavior in the college context. Sociometry, 1968, 63, 58-69.

Bowles, S., & Levin, H. M. The determinants of scholastic achievement: An appraisal of some recent evidence. Journal of Human Resources, 1968, 3, 1-24.

Bracht, G. H. Experimental factors related to aptitude treatment interactions. Review of Educational Research, 1970, 40, 627-645.

Brady, M. E., Clinton, C., Sweeney, J. M., Peterson, M., & Poynor, H. Instructional dimensions study. Washington, D.C.: Kirschner Associates, Inc., 1977.

Brophy, J. E. The student as the unit of analysis (Research Report No. 75-12). Austin: University of Texas at Austin, 1975.

Brophy, J. E. Training teachers in experiments: Considerations relating to nonlinearity and context effects. Paper presented at the annual meeting of the American Educational Research Association, New York, 1977.

Brophy, J. E. Research on effective instruction in first grade reading groups. Paper presented at the annual meeting of the American Association of Colleges for Teacher Education, Chicago, 1978.

Brophy, J. E., & Evertson, C. M. Learning from teaching: A developmental perspective. Boston: Allyn & Bacon, 1974.

Brophy, J. E., & Evertson, C. M. Teacher behavior and student learning in second and third grades. In G. D. Borich & K. S. Fenton (Eds.), The appraisal of teaching: Concepts and process. Reading, Mass.: Addison-Wesley Publishing Co., 1977.

Brunswik, E. Perception and the representative design of psychological experiments. Berkeley, Calif.: University of California Press, 1956.

Burstein, L. Issues concerning inferences from grouped observations. Paper presented at the annual meeting of the American Educational Research Association, Chicago, 1974.

Burstein, L. Data aggregation in educational research: Applications. Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C., 1975.

Burstein, L. The choice of unit of analysis in the investigation of school effects: IEA in New Zealand. New Zealand Journal of Educational Studies, 1976, 2, 11-24.

Burstein, L., & Linn, R. L. The effects of education in the analysis of multi-level data: The problem of heterogeneous within-class regressions. Paper presented at a conference on Methodology for Aggregating Data in Educational Research, Stanford, Calif., 1976.

Burstein, L., & Linn, R. L. The identification of teacher effects in presence of heterogeneous within-class relations of input to outcome. Paper presented at the annual meeting of the American Educational Research Association, New York, 1977.

Burstein, L., & Smith, I. D. Choosing the appropriate unit for investigating school effects. Australian Journal of Education, 1977, 21, 65-79.

Cain, G. C., & Watts, H. W. Problems in making policy inferences from the Coleman Report. American Sociological Review, 1970, 35, 228-242.

Campbell, W. T., & Erlebacher, A. How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. Heilmuth (Ed.), Compensatory education. A national debate (Disadvantaged Child, Vol. 3). New York: Brunner/Magel, Inc., 1970.

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.

Cicirelli, V. E., Cooper, W. H., & Granger, R. L. The impact of Head Start: An evaluation of the effects of Head Start on children's cognitive and affective development (OEO Contract No. B89-4536). Washington, D.C.: Office of Economic Opportunity, 1969.

Cline, M. G., Ames, N., Anderson, R., Bales, R., Ferb, T., Joshi, M., Kane, M., Larson, J., Park, D., Proper, E., Stebbins, L., & Stern, C. Final report. Education as experimentation: Evaluation of the Follow Through Planned Variation Model (Vols. IA, IB). Cambridge, Mass.: Abt Associates, 1974.

Cline, M. G., Ames, N., Anderson, R., Conte M., Fert, T., Hall, C., Joshi, M,, Larson, J., Park, D., Proper, E., Stebbins, L., & Stern, C. Final report. Education as experimentation: Evaluation of the Follow Through Planned Variation Model (Vols. IIA, IIB). Cambridge, Mass.: Abt Associates, 1975.

Cochran, W. G. Early development of techniques in comparative experimentation. In D. B. Owen (Ed.), On the history of statistics and probability. New York: Marcel Dekker, Inc., 1976.

Cohen, J., & Cohen, P. Applied multiple regression/correlation analysis for the behavioral sciences. New York: John Wiley, 1975.

Coleman, J. S. Methods and results in the IEA studies of effects of school on learning. Review of Educational Research, 1975, 45, 335-386.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, T. D., & York, R. L. Equality of educational opportunity (OE-38001). Washington, D.C.: U.S. Office of Education, 1966.

Comber, L. C., & Keeves, J. P. Science education in nineteen countries (International Studies in Evaluation, Vol. I). Stockholm, Sweden: Almqvist & Wiksell. 1973.

Cooley, W. W. Canonical correlation. Paper presented at the annual meeting of the American Psychological Association, Chicago, 1965.

Cooley, W. W., & Lohnes, P. R. Multivariate data analysis. New York: Wiley, 1971.

Creager, J. A. The interpretation of multiple regression via overlapping rings. American Educational Research Journal, 1969, 6, 706-709.

Creager, J. A. Orthogonal and nonorthogonal methods of partitioning regression variance. American Educational Research Journal, 1971, 8, 671-676.

Cronbach, L. J. The two disciplines of scientific psychology. American Psychologist, 1957, 12, 671-684.

Cronbach, L. J. Intelligence? Creativity? A parsimonious reinterpretation of the Wallach-Kogan data. American Educational Research Journal, 1968, 5, 491-511.

Cronbach, L. J. Beyond the two disciplines of scientific psychology. American Psychologist, 1975, 30, 116-127.

Cronbach, L. J. Research on classrooms and schools: Formulation of questions, designs and analysis. Stanford, Calif.: Stanford University, 1976.

Cronbach, L. J., & Furby, L. How do we measure "change"--or should we? Psychological Bulletin, 1970, 74, 68-80.

Cronbach, L. J., & Snow, R. E. Aptitudes and instructional methods: A handbook for research on interactions. New York: Irvington Publishers, Inc., 1977.

Cronbach, L. J., & Webb, N. Between-class and within-class effects in a reported aptitude X treatment interaction: Reanalysis of a study by G. L. Anderson. Journal of Educational Psychology, 1975, 67, 717-727.

Dähllof, U. Kursplaneunersökningar i matematik och modersmålet. Stockholm, SOU, 1960. Reported in K. Härnqvist, Enduring effects of schooling: A neglected area in educational research. Educational Researcher, 1977, 6 (10), 5-11.

Darlington, R. B. Multiple regression in psychological research and practice. Psychological Bulletin, 1968, 69, 161-182.

Darlington, R. B., & Rom, J. F. Assessing the importance of independent valuables in nonlinear causal laws. American Educational Research Journal, 1972, 9, 449-462.

Dershimer, R. A., & Iannaccone, L. Social and political influences on educational research. In R. M. W. Travers (Ed.), Second handbook of research on teaching. Chicago: Rand McNally & Co., 1973.

Deutsch, M. Facilitating development in the pre-school child: Social and psychological perspectives. Merrill-Palmer Quarterly of Behavior and Development, 1964, 10, 249-263.

DuBois, P. H. *Multivariate correlational analysis.* New York: Harper & Row, 1957.

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

Dunkin, M., & Biddle, B. *The study of teaching.* New York: Holt, Rinehart & Winston, 1974.

Dyer, H. S. Toward objective criteria of professional accountability in the schools of New York City. *Phi Delta Kappan*, 1970, *52*, 206-211.

Dyer, H. S., Linn, R. L., & Patton, M. J. A comparison of four methods of obtaining discrepancy measures based on observed and predicted school system means on achievement tests. *American Educational Research Journal*, 1969, *6*, 591-606.

Edwards, A. E. *Experimental design in psychological research* (4th ed.). New York: Holt, Rinehart & Winston, 1972.

Emrick, J. A., Sorenson, P., & Stearns, M. S. *Interim evaluation of the national Follow Through programs 1969-1971* (A technical report). Menlo Park, Calif.: Stanford Research Institute, 1973.

Erlebacher, A. Design and analysis of experiments contrasting the within and between-subjects manipulation of the independent variables. *Psychological Bulletin*, 1977, *84*, 212-219.

Estes, W. K. The problem of inference from curves based on group data. *Psychological Bulletin*, 1956, *53*, 134-140.

Evans, F. B., & Anderson, J. G. The psychocultural origins of achievement motivating the Mexican-American family. *Sociology of Education*, 1973, *46*, 396-416.

Fägerlind, I. *Formal education and adult earnings.* Stockholm: Almqvist & Wiksel, 1975. Reported in K. Härnqvist, Enduring effects of schooling: A neglected area in educational research. *Educational Researcher*, 1977, *6* (10), 5-11.

Farkas, G. Specification, residuals, and contextual effects. *Sociological Methods and Research*, 1974, *2*, 333-363.

Feldman, K. A. Some methods for assessing college impacts. *Sociology of Education*, 1970, *44*, 135-150.

Finn, J. D. *A general model for multivariate analysis.* New York: Holt, Rinehart & Winston, 1974.

Fisher, R. A. Statistical methods for research workers. Edinburgh: Oliver & Boyd, 1925.

Fisher, R. A. Combining independent tests of significance. American Statistician, 1948, 2, 30.

Flanagan, J. C. Changes in school levels of achievement: Project TALENT ten and fifteen year retests. Educational Researcher, 1976, 5 (8), 9-12.

Flanagan, J. C., & Cooley, W. W. Project TALENT: One year follow-up studies. Pittsburgh: University of Pittsburgh, 1966.

Flanagan, J. C., Dailey, J. T., Shaycoft, M. F., Orr, D. B., & Goldberg, I. Studies of the American high school. Washington, D.C.: Project TALENT Office, University of Pittsburgh, 1964.

Flanagan, J. C., & Jung, S. M. Progress in education: A sample survey (1960 - 1970). Palo Alto: American Institute for Research, 1971.

Forsyth, R. A. Some empirical results related to the stability of performance indicators in Dyer's student change model of an educational system. Journal of Educational Measurement, 1973, 10, 7-12.

Gage, N. L. Models for research on teaching (Occasional Paper No. 9). Stanford, Calif.: Stanford Center for Research and Development in Teaching, Stanford University, 1976. (a)

Gage, N. L. Issues in the field of research on teaching. Invited address to the SIG/Research in Mathematics Education at the annual meeting of the American Educational Research Association, San Francisco, 1976. (b)

Galtung, J. Theory and methods of social research. New York: Columbia University Press, 1967.

Garside, R. F. The regression of gains upon initial scores. Psychometrika, 1956, 21, 67-77.

Glaser, R. Evaluation of instruction and changing educational models. In M. D. Wittrock & D. E. Wiley (Eds.), The evaluation of instruction: Issues and problems. New York: Holt, 1970.

Glass, G. V. Primary, secondary, and meta-analysis of research. Educational Researcher, 1976, 5 (10), 3-8.

Glass, G. V., & Stanley, J. C. Statistical methods in education and psychology. Englewood Cliffs, N.J.: Prentice-Hall, 1970.

Glendening, L. *The effects of correlated units of analysis: Choosing the appropriate unit.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1976.

Grunfeld, Y., & Griliches, Z. Is aggregation necessarily bad? *Review of Economics and Statistics*, 1960, *42*, 1-13.

Guilford, J. P. *Psychometric methods*. New York: McGraw-Hill, 1954.

Guthrie, J. W. The new skeptics have gone too far. In R. J. Solomon (Chairman), *Proceedings of the conferences on improving school effectiveness*. Princeton, N.J. Educational Testing Service, 1973.

Haney, W. *Units of analysis issues in the evaluation of Project Follow Through*. Unpublished report, Huron Institute, Cambridge, 1974.

Haney, W. *Analysis of interim Follow Through evaluation reports* (Sponsored by OE/DHEW). Cambridge, Mass.: Huron Institute, 1976.

Hannan, M. T. *Aggregation gain reconsidered*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1976.

Hannan, M. T., & Burstein, L. Estimation from grouped observations. *American Sociological Review*, 1974, *39*, 374-392.

Hanushek, E. *The education of negroes and whites*. Unpublished doctoral dissertation, Massachusetts Institute of Technology, 1968.

Hanushek, E. A., & Kain, J. F. On the value of Equality of Educational Opportunity as a guide to public policy. In F. Mosteller & D. F. Moynihan (Eds.), *On equality of educational opportunity*. New York: Vintage Books, 1972.

Härnqvist, K. Changes in intelligence from 13 to 8. *Scandinavian Journal of Psychology*, 1968, *9*, 50-82.

Härnqvist, K. Enduring effects of schooling-a neglected area in educational research. *Educational Researcher*, 1977, *6*, 5-11.

Hauser, R. M. Context and consex: A cautionary tale. *American Journal of Sociology*, 1970, *75*, 645-664.

Hays, W. L. Statistics for psychologists. New York: Holt, Rinehart & Winston, 1963.

Heath, R. W., & Nielson, M. A. The research basis for performance-based teacher education. Review of Educational Research, 1974, 44, 463-484.

Herriott, R. E., & Muse, D. N. Methodological issues in the studies of school effects. In F. N. Kerlinger (Ed.), Review of Research in Education 1. Itasca, Ill.: F. E. Peacock Publisher, 1973.

Horst, D. P., Tallmadge, G. K., & Wood, C. T. A practical guide to measuring project impact on student achievement (Report 573-586). Washington, D.C.: U.S. Government Printing Office, 1975.

Howe, H. The trouble with research in education. Change, 1976, 8(7), 46-47.

Hunt, J. M. The psychological basis for using pre-school enrichment as an antidote for cultural deprivation. Merrill-Palmer Quarterly of Behavior and Development, 1964, 10, 209-248.

Husen, T. (Ed.). International study of achievement in mathematics (Vols. 1 and 2). New York: Wiley, 1967.

Hyman, H. H., Wright, C. R., & Reed, J. S. The enduring effects of education. Chicago: The University of Chicago Press, 1975.

Jamison, D., Suppes, P., & Wells, S. The effectiveness of alternative instructional media: A survey. Review of Educational Research, 1974, 44, 1-67.

Jencks, C., & Brown M. Effects of high schools on their students. Harvard Educational Review, 1975, 45, 273-324.

Jencks, C. S., Smith, M., Acland, H., Bane, M. J., Cohen, D., Gintis, H., Heyns, B., & Michelson, S. Inequality: A reassessment of the effects of family and schooling in America. New York: Basic Books, 1972.

Kain, J. F., & Hanushek, E. A. On tne value of equality of educational opportunity as a guide to policy (Discussion Paper No. 26). Cambridge, Mass.: Program on Regional and Urban Economics, Harvard University, 1968.

Keesling, J. W., & Wiley, D. E. Regression models of hierarchical data. Paper presented at the annual meeting of the Psychometric Society, Stanford, Calif.: 1974.

Kerlinger, F. N., & Pedhazur, E. J. Multiple regression in behavioral research. New York: Holt, 1973.

Kiesling, H. J. A study of cost and quality of New York school districts. Washington, D.C.: U.S. Department of Health, Education and Welfare, 1970.

Kirk, R. E. Experimental design: Procedure for the behavioral sciences. Belmont, Calif.: Brooks/Cole, 1968.

Krus, D. J., & Krus, P. H. The partitioning of variance habits of correlational and experimental psychologists: The two disciplines of scientific psychology revisited. Educational and Psychological Measurement, 1978, 38, 119-124.

Levin, H. M. A new model of school effectiveness. In R. M. Mood (Ed.), Do teachers make a difference? (OE 58042). Washington, D.C.: U.S. Department of Health, Education and Welfare, 1970.

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

Linn, R. L., & Slinde, J. A. The determination of significance of change between pre- and posttesting periods. Review of Educational Research, 1977, 47, 121-150.

Linn, R. L., & Werts, C. E. Assumptions in making causal inferences from part correlations, partial correlations, and partial regression coefficients. Psychological Bulletin, 1969, 72, 307-310.

Linn, R. L., Werts, C. E., & Tucker, L. R. The interpretation of regression coefficient in a school effects model. Educational and Psychological Measurement, 1971, 31, 85-93.

Lohnes, P. R., & Cooley, W. W. Partitioning redundancy among predictor domains in reduced-rank models for school effects. Paper presented at the annual meeting of American Educational Research Association, San Francisco, 1976.

Lord, F. M. The measurement of growth. Educational and Psychological Measurement, 1956, 16, 421-437.

Lord, F. M. Further problems in the measurement of change. Educational and Psychological Measurement, 1958, 18, 437-454.

Lord, F. M. Elementary models for measuring change. In C. W. Harris (Ed.), Problems in measuring change. Madison: University of Wisconsin Press, 1963.

Luecke, D. F., & McGinn, N. F. Regression analyses and education production functions: Can they be trusted? Harvard Educational Review, 1975, 45, 325-330.

Madaus, G. F., & Airasian, P. W. Issues in evaluating student outcomes in competency based graduation programs. Journal of Research and Development in Education, 1977, 10, 79-91.

Magidson, J. Toward a causal model approach for adjusting for preexisting differences in the nonequivalent control group situation: A general alternative to ANCOVA. Evaluation Quarterly, 1977, 1, 399-420.

Marco, G. L. A comparison of selected school effectiveness measures based on longitudinal data. Journal of Educational Measurement, 1974, 11, 225-234.

Marcus, A. C., Keesling, J. W., Rose, C., & Trent, J. W. An analytical review of longitudinal and related studies as they apply to the educational process. Volume III. Methodological foundations for the study of school effects. Los Angeles: Center for the Study of Evaluation, University of California at Los Angeles, 1972. (ERIC Document Reproduction Service No. ED 079 848)

Marks, E., & Martin, C. G. Further comments relative to the measurement of change. American Educational Research Journal, 1973, 10, 179-191.

Maw, C. E. The problem of data aggregation and cross-level reference with categorical data. Paper presented at the annual meeting of The American Educational Research Association, San Francisco, 1976.

Mayeske, G. W., Wisler, C. E., Beaton, A. E., Weinfeld, F. D., Cohen, W. M., Okada, T., Proshek, J. M., & Tabler, K. A. A study of our nation's schools. Washington, D.C.: Department of Health, Education and Welfare, 1969.

McDonald, F. J. Report on Phase II of the Beginning Teacher Evaluation Study. Journal of Teacher Education, 1976, 27, 39-42.

McDonald, F. J., & Elias, P. J. The effects of teaching performances on pupil learning, Beginning Teacher Evaluation Study Phase II (Final Report). Princeton, N. J.: Education Testing Service, 1976.

McNemar, Q. On growth measurement. *Educational and Psychological Measurement*, 1958, *18*, 47-65.

McNemar, Q. On so-called test bias. *American Psychologist*, 1975, *30*, 848-851.

Medley, D. M. *Teacher competence and teacher effectiveness: A review of process-product research*. Washington, D.C.: American Association of Colleges for Teacher Education, 1977.

Medley, D. M., & Mitzel, H. E. Measuring classroom behavic' by systematic observation. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand McNally & Co., 1963.

Mehan, H. Structuring school structure. *Harvard Educational Review*, 1978, *48*, 32-64.

Michelson, S. The association of teacher resourceness with children's characteristics. In *How do teachers make a difference?* (OE-58042). Washington, D.C.: U.S. Department of Health, Education and Welfare, Office of Education, 1970.

Mood, A. M. Macro-analysis of the American educational system. *Operations Research*, 1969, *17*, 770-784.

Mood, A. M. Partitioning variance in multiple regression analyses as a tool for developing learning models. *American Educational Research Journal*, 1971, *8*, 191-202.

Mosteller, F., & Moynihan, D. P. (Eds.) *On equality of educational opportunity*. New York: Vintage, 1972.

National Assessment of Educational Progress. *National Assessment of Educational Progress: General information yearbook* (Report No. 03/04-GIY). Washington, D.C.: U.S. Government Printing Office, 1974.

Newton, R. G., & Spurrell, D. J. A development of multiple regression for the analysis of routine data. *Applied Statistics*, 1967, *16*, 51-64.

O'Connor, E. F., Sr. Extending classical test theory to the measurement of change. *Review of Educational Research*, 1972, *42*, 73-97.

Pearson, E. S. The probability integral transformation for testing goodness of fit and combining independent tests of significance. *Biometrika*, 1938, *30*, 134-148.

Pearson, K. Mathematical contributions to the theory of evolution. XIV. On the general theory of skew correlation and non-linear regression. In *Draper's Company Research Memoirs. Biometric Series, II*. Cambridge: Cambridge University Press, 1905.

Pedhazur, E. J. Analytic methods in studies of educational effects. In F. N. Kerlinger (Ed.), *Review of Research in Education 3*. Itasca, Ill.: F. E. Peacock Publishers, 1975.

Pennsylvania Department of Education. *Educational quality assessment in Pennsylvania: The first six years*. Harrisburg, Pa.: Pennsylvania Department of Education, 1973.

Perkins, D. N. T. Evaluative social interventions: A conceptual schema. *Evaluation Quarterly*, 1977, *1*, 639-656.

Pipho, C. Minimal competency testing: A look at state standards. *Educational Leadership*, 1977, *34*, 516-520.

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

Poynor, H. Selecting units of analysis. In G. Borich (Ed.), *Evaluating educational programs and products*. Englewood Cliffs, N.J.: Educational Technology Press, 1974.

Poynor, H. *Spurious aggregation and the units of analysis*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1976.

Purves, A. C. *Literature education in ten countries* (International Studies in Evaluation, Vol. 2). Stockholm: Almqvist & Wiksell, 1973.

Rakow, E. A., Airasian, P. W., & Madaus, G. F. Assessing school and program effectiveness: Estimating teacher level effects. *Journal of Educational Measurement*, 1978, *15*, 15-21.

Ray, H. W. *Final report on the Office of Economic Opportunity experiment in educational performance contracting*. Unpublished report, Battelle Laboratories, Columbus, Ohio, 1972.

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

Rosenshine, B. *Teaching behaviors and student achievement*. London: National Foundation for Educational Research in England and Wales, 1971.

Rosenshine, B. Classroom instruction. In N. L. Gage (Ed.), *The psychology of teaching methods* (Seventy-fifth yearbook of the National Society for the Study of Education). Chicago: University of Chicago Press, 1976.

Rosenshine, B., & Furst, N. The use of direct observation to study teaching. In R. M. W. Travers (Ed.), *Second handbook of research on teaching*. Chicago: Rand, McNally & Co., 1973.

Rosenthal, R. Combining results of independent studies. *Psychological Bulletin*, 1978, *85*, 185-193.

Runkel, P. J., & McGrath, J. E. *Research on human behavior*. New York: Holt, Rinehart & Winston, 1972.

Scheuch, E. K. Cross-national comparisons using aggregate data: Some substantive and methodological problems. In R. L. Merritt and S. Rokkan (Eds.), *Comparing nations: The use of quantitative data in cross-national research*. New Haven: Yale University Press, 1966.

Shaycoft, M. The statistical characteristics of school means. In J. C. Flanagan et al., *Studies of the American high school*. Pittsburgh: University of Pittsburgh, 1962.

Shea, B. M. Schooling and its antecedents: Substantive and methodological issues in the status attainment process. *Review of Educational Research*, 1976, *46*, 463-526.

Shively, W. P. "Ecological" inferences: The use of aggregate data to study individuals. *American Political Science Review*, 1969, *63*, 1183-1196.

Shulman, L. S. Reconstruction of educational research. *Review of Educational Research*, 1970, *40*, 371-396.

Simon, H. A. Spurious correlations: A causal interpretation. *Journal of the American Statistical Association*, 1954, *49*, 467-479.

Smith, M. L., & Glass, G. V. Meta-analysis of psychotherapy outcome studies. *American Psychologist*, 1977, *32*, 752-760.

Smith, M. S. Equality of educational opportunity: The basic findings reconsidered. In F. Mosteller and D. P. Moynihan (Eds.), *On equality of educational opportunity*. New York: Random House, 1972.

Snow, R. E. Brunswikian approaches to research on teaching. *American Educational Research Journal*, 1968, *5*, 475-485.

Snow, R. E. Representative and quasi-representative designs for research on teaching. *Review of Educational Research*, 1974, *44*, 265-291.

Soar, R. S. *Final report. Follow Through classroom process measurement and pupil growth (1970-1971)*. Gainesville, Fla.: College of Education, University of Florida, 1973.

Soar, R. S., & Soar, R. M. An empirical analysis of selected Follow Through programs: An example of a process approach to evaluation. In I. J. Gordon (Ed.), *Early Childhood Education*. Chicago: National Society for the Study of Education, 1972.

Soar, R. S., & Soar, R. M. *Classroom behavior, pupil characteristics, and pupil growth for the school year and for the summer*. Gainesville: Institute for Development of Human Resources, University of Florida, 1973.

Soar, R. S., & Soar, R. M. An attempt to identify measures of teacher effectiveness from four studies. *Journal of Research in Teacher Education*, 1976, *27*, 261-267.

Stallings, J. A., & Kaskowitz, D. H. *Follow Through classroom observation evaluation 1972-1973*. Menlo Park, Calif.: Stanford Research Institute, 1974.

Stanley, J. C. General and special formulas for reliability of differences. *Journal of Educational Measurement*, 1967, *4*, 249-252.

Stufflebeam, D. L. Evaluation as enlightenment for decision making. In W. H. Beatty (Ed.), *Improving educational assessment as an inventory for measures of affective behavior*. Washington, D.C.: Association for Supervision and Curriculum Development, National Education Association, 1969.

Stufflebeam, D. L. The use of experimental design in educational evaluation. *Journal of Educational Measurement*, 1971, *8*, 267-274.

Stufflebeam, D. L., & Webster, W. J. *The state of theory and practice in educational evaluation*. State-of-the-Art lecture (Division H) presented at the annual meeting of American Educational Research Association, Toronto, Canada, 1978.

Tallmadge, G. K., & Horst, D. P. *A practical guide to measure project impact on student achievement* (Number 1 in a series of monographs on evaluation in education). Washington, D.C.: U.S. Government Printing Office, 1975.

Tallmadge, G. K., & Horst, D. P. *A procedural guide for validating achievement gains in educational projects*. Washington, D.C.: U.S. Government Printing Office, 1976.

Tatsuoka, K. K. Vector-geometric and Hilbert-space reformulations of classical test theory. *Dissertation Abstracts International*, 1975, *36*, 246-A.

Tatsuoka, M. M. *Multivariate analysis: Techniques for educational and psychological research*. New York: Wiley, 1971.

Tatsuoka, M. M. *Nationwide evaluation and experimental designs*. Paper presented at the annual meeting of American Educational Research Association, 1972.

Tatsuoka, M. M. Multivariate analysis in educational research. In F. N. Kerlinger (Ed.), *Review of Research in Education 1*. Itasca, Ill.: F. E. Peacock Publishers, 1973.

Tatsuoka, M. M., & Tiedeman, D. V. Statistics as an aspect of scientific method in research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand, McNally & Co., 1963.

Thorndike, R. L. Intellectual status and intellectual growth. *Journal of Educational Psychology*, 1966, *57*, 121-127.

Thorndike, R. L. *Reading comprehension education in fifteen countries* (International studies in evaluation, Vol. 3). Stockholm: Almqvist & Wiksell, 1973.

Tikunoff, W. J., Berliner, D. C., & Rist, R. C. *Abstract from special study A: An ethnographic study of the forty classrooms of the Beginning Teacher Evaluation Study known sample*. San Francisco: Far West Laboratory for Educational Research and Development, 1975.

Timm, N. H. *Multivariate analysis with application in education and psychology*. Belmont, Calif.: Wadsworth Publishing Co., Inc., 1975.

Tucker, L. R., Damarin, F., & Messick, S. A base-free measure of change. *Psychometrika*, 1966, *31*, 457-473.

Tukey, J. W. Causation, regression, and path analyses. In O. Kempthorne, T. A. Bancroft, J. W. Gowen, & J. L. Lush (Eds.), *Statistics and mathematics in biology*. Ames: Iowa State College Press, 1954.

Walberg, H. J. Social environment as a mediator of classroom learning. Journal of Educational Psychology, 1969, 69, 443-448.

Walberg, H. J., & Rasher, S. P. Public school effectiveness and equality: New evidence and its implications. Phi Delta Kappan, 1974, 56, 3-9.

Wallach, M., & Kogan, N. Modes of thinking in young children. New York: Holt, Rinehart & Winston, 1965.

Ward, J. H. Partitioning variance and contribution or importance of a variable: A visit to a graduate seminar. American Educational Research Journal, 1969, 6, 467-474.

Werts, C. E. The partitioning of variance in school effects studies. American Educational Research Journal, 1968, 5, 311-318.

Werts, C. E., & Linn, R. L. Analyzing school effects: How to use the same data to support different hypotheses. American Educational Research Journal, 1969, 6, 439-447.

Werts, C. E., & Linn, R. L. A general linear model for studying growth. Psychological Bulletin, 1970, 73, 17-22. (a)

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

Wiley, D. T. Design and analysis of evaluation studies. In M. D. Wittrock & D. E. Wiley (Eds.), The evaluation of instruction: Issues and problems. New York: Holt, Rinehart & Winston, 1970.

Willems, E. P. Planning a rationale for naturalistic research methods. In E. P. Willems & H. L. Raush (Eds.), Naturalistic viewpoints in psychological research. New York: Holt, Rinenart & Winston, 1969.

Williams, E. J. Regression analysis. New York: Wiley, 1959.

Winer, B. J. Statistical principles in experimental design (2nd ed.). New York: McGraw-Hill, 1971.

Wittrock, M. D., & Wiley, D. E. (Eds.). The evaluation of instruction: Issues and problems. New York: Holt, 1970.

Wright, S. Correlation and causation. Journal of Agriculture Research, 1921, 20, 557-585.