DOCUMENT RESUME

ED 221 554 TM 820 573

AUTHOR Burstein, Leigh
TITLE State of the Art Methodology for the Design and Analysis of Future Large Scale Evaluations: A Selective Examination.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
REPORT NO CSE-R-177
PUB DATE 81
NOTE 58p.

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Data Analysis; *Program Evaluation; *Quasiexperimental Design; *Research Design; *Research Methodology; State of the Art Reviews
IDENTIFIERS *Large Scale Programs; LISREL Computer Program; *Structural Equation Models

ABSTRACT

Two specific methods of analysis in large-scale evaluations are considered: structural equation modeling and selection modeling/analysis of non-equivalent control group designs. Their utility in large-scale educational program evaluation is discussed. The examination of these methodological developments indicates how people (evaluators, methodologists, agency staff) involved in the design and conduct of large-scale program evaluation might approach decisions concerning appropriate methodology and its proper use. Evaluation activities and the range of methodological issues considered are: field-based investigations of large-scale programs; evaluation of on-going programs and of various forms of social experiments; and both well-defined and broad-based educational programs. Both analytical procedures employ explicit models of the phenomena believed to be responsible for the difficulties in estimating program effects. Both are also adaptable to situations where there are no specific comparison or control groups and where panel data exists on program participants. A discussion of the LISREL model and its limitations as an analytical approach to estimation in structural equation modeling with latent variables is included. (PN)

STATE OF THE ART METHODOLOGY FOR THE DESIGN AND ANALYSIS OF FUTURE LARGE SCALE EVALUATIONS: A SELECTIVE EXAMINATION

Leigh Burstein

CSE Report No. 177
1981

Center for the Study of Evaluation
Graduate School of Education, UCLA
Los Angeles, California 90024

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education Department of Education. 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.

Introduction

The following report selectively examines recent developments in quantitative methodology and considers their possible utility in large-scale program evaluations in education. At the outset we limit attention to two specific categories of analytical methods: structural equation modeling and selection modeling and related issues in analysis of quasi-experimental data (non-equivalent control group designs). While these topics, by no means, cover the full range of recent advances in the technology for analyzing quantitative data in large-scale program evaluations, they are representative of the methodological concerns that arise in such investigations, the means analysts propose to deal with the concerns, and the strengths and limitations of primarily technical approaches to resolving ambiguity in evaluation results. As such our examination of these methodological developments is intended to suggest how persons (evaluators, methodologists, agency staff) involved in the design and conduct of large-scale program evaluation might approach decisions about appropriate methodology and its proper use.

Delineation of Relevant Program Evaluations

We further delineate the purview of this investigation by stating the types of evaluation activities and the range of methodological issues to be considered. We are concerned with field-based investigations of large-scale programs typically approved by legislative actions and implemented (or to be implemented) by governmental agencies. Both evaluations of on-going programs (e.g., Title I) and of various forms of social experiments (e.g., Negative Income Tax experiments) are relevant to the present

discussion (Cook (1981) restricts his attention to the former). The domain also encompasses both well-defined programs (i.e., those with a discrete number of specific program alternatives such as the various models in operation in Planned Variation Follow Through) and broad-based educational reforms as represented by Title I, the Emergency School Aid Act (ESAA), and bilingual education. (A related paper (Burstein, 1981) focussed strictly on evaluations of well-defined programs).

Types of Evaluation Questions

The limits placed on the evaluation activities of interest are in the kinds of questions one seeks to answer and the form of data collection in the evaluation. Cook (1981) discusses six types of questions that evaluators try to answer:

1) Who are the clientele and service providers and to what extent are target groups among the clients? (Demography)
2) What are the delivered services and the contexts in which services are received? (Implementation)
3) How do program services affect clients in both expected and unexpected ways? (Effectiveness)
4) How are other elements (teachers, schools, families, etc.) of the educational system affected by the program services? (Impact)
5) Why do program services affect outcomes in the way they do? (Causation)
6) What are the costs of the services and how cost-effective are different ways of achieving a particular result? (Economic costs)

The questions about effectiveness, impact, and causation are central to our examination. To be comprehensive, investigations of these types of questions require information about the characteristics of the program,

its clients and participants and the context in which it is implemented, the educational and social processes (intended and actual) occurring within program sites, and the outcomes of programs at various levels (student, teacher, classroom, school, community, etc.) of the educational system. Conceptual and analytical machinery are then employed to elucidate the linkages and connections among the various sources of information.

Types of Data Collections

In the past, most large-scale field evaluations of educational programs collected mainly "quantitative" measures of program characteristics and outcomes largely derived from survey questionnaires completed by clients and other relevant program participants (e.g., teachers, principals, parents), limited interviews with program personnel and observations of program activities (e.g., Stallings and Kaskowitz, 1974), and paper-and-pencil measures of cognitive and affective outcomes. Data were collected from multiple sites for each variant of the program to achieve a given degree of information about program variation and a sufficient number of observations for statistically powerful tests of program effects.

Recently, however, data collection in even large-scale program evaluations has taken on an increasingly "qualitative" character. Extended case studies were conducted in either a subset or all sites in a number of recent large-scale evaluations (e.g., Title I Parent Involvement Study conducted by SDC; Study of the Longitudinal Effects of the California Early Childhood Education Program conducted by CSE, the Rand Study of Federal Programs Supporting Educational Changes, the evaluation of Curriculum Development Projects in Science Education conducted by CIRCE). At the least, the inclusion of case studies in these evaluations provide a richer picture of program process

than was obtainable from strictly questionnaire information. And, as methods for synthesizing multiple case studies and integrating qualitative and quantitative information improve, qualitative methods will play an increasingly more prominent role in the reportoire of evaluation activities previously concentrated on less dense forms of data collection.

Despite the increasing role of qualitative methods and our positive attitude about their central role in future evaluations, the remainder of the paper will restrict attention to developments in quantitative methods from multi-site investigations using questionnaire, interview, test and perhaps small-scale observational data. We impose this restriction for two reasons. First, the analytical developments considered are appropriate primarily for the more traditional kinds of quantitatively oriented studies. Second, others (e.g., Daillak & Alkin, 1981) are more capable at this point of stating the case for qualitative methods.

Overview of the Report

The remainder of the report will proceed as follows. First, a general overview of current perspectives on the design and conduct of large-scale program evaluations is presented. The intent is to explain why the climate for future large-scale evaluations is conducive to the introduction of improved methods of analysis. Second, two specific categories of analytical methods (structural equation modeling, and selection modeling/analysis of non-equivalent control group designs) are considered. The basic conceptual and analytical foundations for each method are described, issues that motivate its use in program evaluations are delineated, and specific strengths and weaknesses of each method in program evaluation contexts are identified.

Current Perspectives on Design and Analysis in Large-scale Program Evaluations

There are strong signs that large-scale educational evaluation has witnessed the end of an era. From the late '60's and throughout the 1970's, the federal government, under legislative mandate, mounted major evaluations of just about every conceivable educational program. Wargo (1977) points to 110 major evaluations of federal educational programs funded by the Office of Planning, Budgeting, and Evaluation of the Office of Education at a cost of over $80 million during the 1971-1979 period. The figure does not even include all the major evaluations done by the Office of Education, much less NIE and other branches of HEW.

Many of these large-scale multiyear studies have been highly visible in the educational community though their direct influence on legislative action is less clear (Barnes & Ginsberg, 1979; Cohen & Garet, 1975; Cross, 1979 Wisler & Anderson, 1979). In most cases, the debates about the quality and merits of these evaluations have been heated. This has especially been the case for evaluations of compensatory programs such as Head Start (e.g., Cicirelli et al., 1969, 1971; Smith & Bissell, 1971), Project Follow Through (Anderson, 1976; Cline et al., 1974; Haney, 1977a, 1977b; House, Glass, McLean, & Walker, 1978; Stebbins et al., 1977), and Bilingual Education (AIR, 1979; Center for Applied Linguistics, 1979). The literature on evaluations of these programs is replete with critiques, reanalyses, and secondary analyses, not to mention the often self-serving attacks from program advocates and critics.

Signs of Change.

Emphasis. There are clear signs, however, that the large-scale evaluations of the 1980's may well be different. First, recent scholarly (e.g., Cook, 1981; Cronbach, 1978; Cronbach & Associates, 1980; House, 1977, 1978; Raizen & Rossi, 1981) and policy (e.g., Boruch & Cordray, 1980) contributions provide well-reasoned accounts of the complexity of program evaluations in highly politicized contexts and persuasive arguments for different views of evaluation's role in the formation of social policy. These writings urge that less emphasis be placed on the traditional social science/experimental design paradigm for impact evaluation while more effort be devoted to describing and explaining the processes of educational programs and their consequences over a broad range of outcomes. The overly simplistic overall program impact question (i.e., does program A affect pupil outcomes?) that guided so many of the OPBE funded studies (e.g., ESAA (Coulson et al., 1977); Follow Through (Stebbins et al., 1977); and Bilingual Education (AIR, 1979)) appears to be on the decline.

Instead, recent evaluations involve more direct efforts to investigate and describe the consequences (intended and otherwise) of educational programs. This "information" as characterized by Cronbach et. al. (1980) involves a "move away from stand-alone evaluations of programs and toward a more synoptic view of the numerous programs that address the same social programs" (p. 72-73) and they urge that evaluations employ multiple studies using different strategies to investigate subquestions and that the evaluation plan evolve as individual studies expose uncertainties more clearly. The NIE Compensatory Education Study (NIE, 1977) and the evaluations of services to handicapped children under Public Law 94-142 (Bureau of

Education for the Handicapped, 1978) are clear examples of this type of evaluation.

This shift in evaluation emphasis is a logical response to the findings that variation in implementation within a program is generally greater than between programs (Stebbins, et al., 1977), new program "treatments" are quickly diffused to non-participating groups (schools, etc.) (Coulson, 1978) and that the effects that are discerned depend on the characteristics of the program processes "as implemented" rather than on the ascribed program characteristics (Cook, 1981; Cronbach, 1978; Cronbach & Associates, 1980; Rogosa, 1978). Under such conditions, only those evaluation activities that delve beneath the surface descriptions of programs can be expected to generate quality information for policy formation.

Methodological improvements. Clearly, the impetus for change in the conduct of large-scale educational evaluation exists. The philosophical, theoretical, and political bases for the changes have been and are being articulated. Under such conditions, the climate for evaluation in the 1980's is quite open to new designs and strategies for evaluating the effects of educational programs. The task of defining these designs and strategies and illustrating their worth remains.

Fortunately, it is unnecessary to begin from scratch in the design of large-scale evaluations for the 1980's. While actual educational evaluations over the past decade, for the most part, utilized pre-1970's technology (quantitative methodology, psychometric methods), investments of resources in basic research on methodology and measurement during the 1970's led to substantial improvements in the state of the art.

The relatively unsophisticated applications of experimental, quasi-experimental, and non-experimental methods that led to the findings of the Coleman report (Coleman et al., 1966) and of early Head Start and Follow Through evaluations need not be repeated. Better, more sensitive quantitative methodology is now available and is more suited to the shift in emphasis in large-scale evaluations.

The same can be said for the measurement of program outcomes and processes. Approaches for developing program sensitive test instruments as well as a broader view of the range of program outcomes are currently on the evaluation agenda. The investment of resources to obtain more intensive and descriptive measures of program implementation and processes appears to be a standard feature of recent large-scale evaluations (e.g., the Title I Parent Involvement Evaluation conducted by Systems Development Corporation). These measurement strategies should facilitate more useful evaluations. Better methods of knowledge and data synthesis (e.g., recent work by Glass and Light) should also contribute to better evaluations.

Basis for Methodological Improvements

The special issue of the Journal of Educational Statistics on the Emergency School Assistance Act (ESAA) Evaluation (JES, 1978; see especially Rogosa, 1978) and Cronbach's report on designing educational evaluation (1978) provide documentation of key evaluation methodology issues and help to motivate our general concerns. The basis for our investigation into evaluation methodology is in part the following set of general premises:

(1) Evaluation is inevitably an empirical enterprise, "examining events in sites where the program is tried and the reactions and subsequent performance of the persons served .... (as such it) is typically identified with the application of social science methods: observation, measurement, and/or use of informants." (Cronbach, 1978, pp. 25-26).

(2) "The success of an evaluation effort should be measured by its social usefulness or utility .... Technical decisions should not be made independently of the political and social context of an evaluation. The central question is: How can we design, analyze and report evaluations so as to make them maximally useful?" (Rogosa, 1978, p.80, emphasis added).

(3) "Evaluators are unwise to collect data only on pretest and posttest achievement measures or conduct analyses that only determine the statistical significance of the overall treatment effect. Additional data on process, and on program realization, are essential for adequate descriptions of programs operating in complex settings." (Rogosa, 1980, p. 81).

(4) The analytical strategies in program evaluations should be adapted to the substantive problems under investigation rather than adapting the evaluation of program impact to fit the analytical methods. Natural designs and analysis should evolve from the structure and function of the program. (Burstein, 1980).

(5) Program evaluation is typically carried out within a multilevel educational context. Program activities occur in the groups (classrooms, schools, etc.) to which an individual belongs. These groups

influence the thoughts, behaviors, and feelings of their members. (Burstein, 1980).

(6) Educational interventions are typically implemented within on-going programs. They vary in "fit" with existing activities and predilections and vary in duration. Interventions in social settings are inherently dynamic activities.

There are more specific methodological corollaries to these general premises:

(1) "No one level is uniquely responsible for the delivery of and response to educational programs ... confining substantive questions to any one level of analysis is unlikely to be a productive research strategy" (Rogosa, 1978, p. 83). Thus, attempts to answer questions about the effects of educational programs require analyses at and within the levels of the educational hierarchy (Burstein, 1980).

(2) Even when one starts with a controlled experiment with random assignment, features of the experimental design break down through processes of attrition, contamination, and differential penetration of the treatment. Under such conditions, quasi-experimental forms of adjustment and control are inevitably necessary and thus should be anticipated as part of the evaluation design.

(3) In the course of an educational program, students are members of multiple groups (e.g., classes). The features of these group contexts and the consistency of student's educational experiences within them over time warrant consideration for dynamic modeling of program experiences (Burstein, 1981; Tuma, Hannan, & Groenfeld, 1978; Rogosa, 1980).

(4) In field experiments with well-defined treatments, the variation in the fidelity of program practices with teacher (school, etc.) predilections and skills leads to a continuous range of program processes. Under these conditions, modeling the intervention as a dichotomous rather than a continuous event is an insufficient approach for investigating program effects (Burstein, 1981; Cronbach, 1978; Rogosa, 1978).

(5) Even when random assignment occurs at some aggregate level (e.g., school), the variation in the treatment effects for students within aggregates needs to be investigated, especially in terms of its consequences for the equalization of educational opportunity.

(6) Programs have multiple effects. Multiple measurement is needed to encompass intended and unintended effects (desirable or undesirable), (Cronbach, 1978, p. 26).

Fortunately, one can point to specific bodies of methodological work that are responsive to both the general perspectives and the accompanying methodological corollaries. In the following sections we will elaborate the connections for a selected set of methodological strategies.

Examination of Specific Analytical Developments

The analytical methods to be examined represent broad areas of methodological concerns that first developed within social science research in general. To understand why this is both an obvious and proper starting point, one need only consider the criteria used to delineate our relevant universe of large-scale program evaluation. In particular we are interested in design and analytical problems in evaluations that fit the following description:

(1) The evaluation should have been conducted on a distinct funded educational program(s) rather than be a general shift in the behaviors of an educational system. There must have been some form of intervention, innovation, or change in the ongoing educational program.

(2) The evaluation must have involved multiple sites of each presumably distinct program type.

(3) The program must have been implemented (i.e., the main program activities must operate) at the level of the school or lower.

(4) Both outcome and program process data must have been collected during the course of the evaluation.

(5) Outcome data must be available over multiple time points.

(6) Good documentation of the original evaluation must exist.

The above delimiters eliminate evaluations which are short-term efforts, have a limited number of sites, or are of programs presumably constant over all schools in a district. These criteria include evaluations of well-defined program interventions such as in provided by a specific Head Start or Follow Through model, interventions that are less specific in program prescription but nonetheless are assigned to "sites" in a systematic manner such as by random assignment (e.g., the ESAA Evaluation) and more pervasive social interventions where participants are essentially all persons with a prescribed set of characteristics (e.g., Title I, Bilingual Education).

To gain a better perspective on the kind of study situation evnisioned consider the following modified version of the conceptual framework for investigating the impact of educational reforms outlined in Burstein (1981). One starts by identifying the specific elements of educational and social systems in which programs are introduced and the processes and outcomes that result. The elements are the characteristics and attributes of individual

students, families, groups of students, teachers, classes, groups of teachers, schools, and communities. The processes are developmental, instructional, curricular, psychological, interpersonal, and social. Both elements and processes can take on either static or dynamic properties though the latter are more likely in school settings, especially those with large numbers of poor children participating in school reform programs.

A general model containing the essential elements and processes of the conceptual framework is as follows. The interrelations among five distinct classes of variables are incorporated in the model: program instruction, schooling context (class, school, community, etc.), student entering characteristics, and student performance. Each class may represent many distinct variables (or sets of variables). For example, "instruction" refers to the various characteristics of the instruction a student receives in a specific classroom or school. Particular teacher attributes (e.g., warmth, enthusiasm, clarity of presentation) and instructional processes (e.g., structure, grouping, pacing, types of reinforcements, teachers' questioning behavior, quality and variety of instructional materials) both fit under the instruction rubric. Certain aspects of the instructional practices also provide evidence about the degree of program implementation. Nonetheless, any measure of program implementation would still fall within the "instruction" category for present purposes.

The term student "performance" is meant in the broad sense; the full range of educational, social, and psychological outcomes fit under this general rubric. The restriction to student outcomes could be broadened to include other units (teachers, classes, schools), but not without making the task of generating the framework even more unwieldy than it will appear here.

The role of schooling context in the model is multifaceted. Its most proximal manifestations are in the classroom where the program is implemented. For example, the overall level and heterogeneity of ability in a class places constraints on instructional content, organization, and management. The consequences of these constraints vary for different reform programs. Class heterogeneity places a strain on time and resources in individually prescribed educational programs. Decisions about the pacing of instruction become more difficult in programs emphasizing large group instruction.

The student's role within the classroom is also directly influenced by its composition (Burstein, 1980b; Firebaugh, 1980; Webb, 1980). There is obviously a complicated balance between having classmates compatible in ability and temperament versus having peers that are more or less able and/or have contrasting personalities. Either combination might foster intellectual, social, and psychological growth under the "right" conditions. Here, again, programs with different emphases and organization might interact differentially with class composition, making a given student's role more comfortable or stressful.

There are also other elements of context provided by the class, school and community environment for the program. Sirotnik and Oakes (1981) provide a particularly comprehensive discussion of the possible components of schooling context.

The pattern of relationships depicted in Figure 1 include the following:

(1) Students are eligible for the program and are selected on the basis of entering characteristics.

(2) Student entering characteristics (ability, "preferred learning style", motivation to learn, "preparation for learning") affect performance at any point in time.

(3) Entering characteristics interact with program characteristics to give certain students relative advantages in certain programs (e.g., low ability students benefit from relatively higher levels of teacher control and direction for language and mathematics mechanics).

(4) Programs interact with school personnel characteristics (preferred style, personality, authority relationships, cohesiveness).

(5) Schooling context (ability distribution, personality, presence/absence of demanding/disruptive students, orderliness at class or school level) affects instruction (emphasis, amount of material covered, organization, program delivery).

(6) Students' shared educational and social experiences in classrooms and schools depend on student entering characteristics, instruction, schooling context and program characteristics.

(7) Students from same class in year 1 may be assigned to different classes in year 2 or may leave the school.

(8) Students not present in year 1 may enter school (and thus program classes) during year 2.

(9) Implementation of programs may differ for year 2 from year 1.

(10) Instructional (program) characteristics e.g., teacher "style", organization) may differ from year 1 to year 2 and effect of instruction (program) year 1 followed by instruction (program) year 2 is not necessarily additive.

(11) Contextual characteristics may differ from year 1 to year 2.

(12) Conditions (1) - (5) hold for year 2 in similar fashion as for year 1.

(13) Program differs from "normal" standard instruction and may interact. Though instruction of Type A may be better than instruction of Type B, instruction of Type B might be better for students following participation in the program than Type A would be.

The two areas of analytical developments to be discussed below become relevant in a program of the type described above for several reasons. First, eligibility for program participation typically depends on specific ascribed characteristics (e.g., poverty, bilingualism, ethnicity). Even in nominally "experimental" investigations, selection for participation may have non-random aspects at some level as in the case where the program is randomly assigned to a sample of schools from a pool of volunteers. A further complication is the non-stable participant sample; students enter and leave classrooms, teachers and schools drop out of programs for various reasons.

A second feature requiring analytical attention is the sheer number of elements that potentially enter a comprehensive picture of program processes and outcomes, the complexity of their interrelation, and the inherent problems in measuring key variables by the kinds of questionnaire, interview, observation and test data typically used. All of the elements of model specification from a clear understanding of the question of interest through identification and operationalization to appropriate analyses and interpretation have a bearing on the fidelity of the evaluation conclusions to the program's actual consequences.

To a certain degree, these features align with the two analytical developments to be considered below.

Non-Equivalent Control Group Designs/Selection Modeling

From the inception of the large-scale educational evaluation efforts of the 1960's, evaluators have tried to employ the paradigm for experimentation in the field investigations. With rare exception, however (see Boruch, 1974), investigators quickly found themselves in the midst of non-experimental or at best quasi-experimental studies wherein all the best intentions about random assignment went unfulfilled.

From a methodological perspectice, consciousness about the inadequacy of analytical methods in these investigations can be traced back to Campbell and Erlebacher's (1970) lament (perhaps complaint is the better term) that regression artifacts in quasi-experimental evaluations were causing compensatory education to look harmful. While certain aspects of the original Campbell-Erlebacher critique have been found to be less generally applicable than originally believed, the design constraints that bothered them remain at the center of current analytical concerns.

Basic analytical issues. Reichardt's (1979) and Barnow, Cain and Goldberger's (1980) discussions of the problems in analyzing non-equivalent control group designs are a particularly helpful starting point for our examination. As Reichardt points out, the main issue is the effect of uncontrolled selection on the estimation of program effects. When subjects are randomly assigned to programs (or non-program), groups can be considered initially equivalent though the equivalence can be vitiated if there is differential attrition. Without random assignment program groups would not be expected to equal even in the absence of a program effect. Thus, in order to "equate" non-equivalent groups, it is necessary to adjust or control for initial differences.

The analyst this juncture invariably recognizes that the task at hand is to (a) identify the selection process underlying group membership (program, non-program) and (b) include the variables that determine selection in the analysis of program effects. Ideally, this analytical strategy would control for the effects of initial differences.

Until recently the statistical method typically employed by analysts in quasi-experiments was the analysis of covariance (ANCOVA), which is essentially a linear regression of program outcomes, Y on program status Z, (e.g., 1 = in program, 0 not in program) and pre-program true ability W^2. Thus the "ideal" analytical model is represented by (1) below:

$$Y = \alpha Z + W + \varepsilon \quad , \qquad (1)$$

where α is the estimate of program effect, Z, and W is the covariance adjustment for true initial differences.

But as is well-known, W is unobservable. Under these conditions Barnow, Cain and Goldberger (1980) ask "How may the evaluator persuade an interested audience that the measured effect of Z on Y is free of any contamination from a correlation between Z and W, given that W is not available as an explanatory variable?" (p. 47). Their answer to their own question is that "unbiasedness is attainable when the variables that determine treatment assignment are known, quantified and included in the equation." (Barnow, et. al., 1980, p. 47. See also Barnow, 1975; Cain, 1975; and Goldberger, 1972). Thus if one has an observed variable t that was used to determine group assignment (in general t will be a score based on a composite of variables, some of which may be correlates of W), then t may be used to replace W as the explanatory variable in (1):

$$Y = V_1 Z + \beta_2 t + \varepsilon^* \qquad (2)$$

Under conditions to be specified, β_1, in equation (2) would be an unbiased

estimate of the program effect α. Thus either W or t will remove the contamination which leads to "selectivity bias".

But the question arises about whether the selection process can be known precisely (i.e., one is unable to quantify t). In this case, investigators have settled for a set of variables, X, that serve as proxies for W. The X's may also include variables which enter t. The equation to be estimated is then

$$Y = \gamma_1 Z + \gamma_2 X + \varepsilon^{**} \quad . \qquad (3)$$

Equation (3) is essentially the standard ANCOVA model as employed in the analysis of quasi-experimental data. Unfortunately, an estimate of γ_1 will in general be a biased estimate of the true program effect α. Statistically, this bias depends on the covariance of Z and W conditional on X. Moreover, contrary to Campbell and Erlebacher's (1970) assertion, the bias may be either positive or negative. Investigations by Goldberger (1972); Barnow (1973), Cain (1975), Cronbach, Rogosa, Floden, and Price (1977) and Bryk and Weisberg (1977) clearly demonstrate this property.

To better understand the ramifications of the inability to observe true preprogram ability (W) and/or to accurately quantify the selection process (t), we consider the sources of biases in estimation of program effects when the ANCOVA model is employed with nonequivalent groups. Reichardt (1980) discusses seven sources, most of which are pertinent to this inquiry.

The problems due to errors in measuring the covariates (the X's in equation (3)) are the most frequently examined source of bias. Even when measurement errors are random, they lead to attenuated estimates of covariate effects and thus result in an underadjustment for pre-existing differences between different programs. The errors in the covariate cause the treatment

effect estimate from ANCOVA to converge toward estimates from an ANOVA which completely ignore pre-existing group differences.

The second source of bias in ANCOVA is the possibility of differential growth rates among identifiable subpopulations under conditions where subpopulation membership is related to program assignment. Though individuals from different subpopulations may be the same initially, their later differences may be attributed to differences in maturation. In this case, growth invalidates ANCOVA because within-group growth does not completely account for between-group differences in growth.

According to Reichardt, related sources of bias due to changes between the time of program entry and measurement of program outcomes which are irrelevant to the treatment are trait instability and the changing structure of behavior. Trait instability refers to differential variability (fluctuation) in scores over time as opposed to average mean differences. The changing structure of behavior refers to the possibility that the processes that account for given naturally occurring behaviors vary over time with different characteristics and processes becoming disproportionately important at various times. (Cronbach et al (1977) discuss this source in some detail.)

Other complications identified by Reichardt include (a) operationally unique pretests and posttest (i.e., even though the measure of initial status and final performance is nominally the same, they are operationally distinct as different abilities and skills are tapped at different points in time); (b) non-linear regression lines (not properly incorporated in the model) and non-parallel regression lines (due to treatment interaction effects, floor and ceiling effects, differential growth between groups, or between group differences in the reliability of the covariates).

Reichardt (1980) describes four approaches for ruling out selection differences as a rival explanation for program effects. The first three (namely, developing a causal model of the posttest, developing a causal model of the assignment process, the Cronbach et. al. (1977) combination of the two approaches) are basically elaborations on the identification of W, t, or both as described earlier. One essentially adopts a broader, theoretically grounded and empirically estimated model of how posttest behavior is expected to vary in the absence of the program (modeling the posttest; Cronbach et. al. call this identifying the "ideal covariate"), how individuals are assigned to "treatment" groups (modeling the assignment process; or identifying the "complete discriminant" in Cronbach et al.'s terminology) or do both. After determining a specific approach, there are still questions about appropriate analytical machinery to adjust for measurement errors and estimate W and t appropriately. The sheer complexity of the adjustment has led some investigators to recommend the use of procedures derived from the work of Joreskog (1970, 1973, 1974, 1977, Joreskog and Sorbom, 1976, 1978) for the analysis of covariance structures. These methods attempt to simultaneously correct for the effects of measurement error and irrelevance in multiple covariates. We withhold further discussion of these techniques to the next major section of our report.

Value-added analysis. The fourth approach discussed by Reichardt (1980) is the modeling of change or growth. Promising work on this topic has been carried out by Bryk and Weisberg (Bryk, 1977; Bryk and Weisberg, 1976; Bryk, Strenio , and Weisberg, 1980; Strenio, 1977; Weisberg 1978). They introduced a variety of analytical methods for estimating the "value-added" by program participation. Their value-added analysis is built upon the notion that educational programs are dynamic interventions in natural growth processes. Thus Bryk and Weisberg first modeled natural growth processes and then assessed program impact on the processes.

The basic idea underlying Bryk-Weisberg value-added procedure is to compare average observed growth between pre- and post-test with an estiamte of the amount expected in the absence of an intervention.

To employ their techniques, one needs to have pretest (Y_{1i}) and post-test data (Y_{2i}) on a sample of individuals as well as the time (calendar dates t_1 and t_2) at which observations were obtained and the age (a_{i1}, a_{i2}) of each individual at these times. In the more general case, one would also obtain information on other background variables (X_i). Their methods also seem to be applicable whether treatment is represented by a discrete group membership variable (treatment A vs. treatment B) or by a set of variables describing program and instructional differences (e.g., explicit charicteristics of instruction, schooling, context, and program implementation).

Bryk and Weisbergs's general model can then be expressed as

$$Y_i(t) = G_i(t) + R_i(t) \qquad (4)$$

$$G_i(t) = \pi_i a_i(t) + \delta_i \qquad (5)$$

$$\pi_i = \theta_0 + \sum_{j=1}^{J} \theta_j X_{ij} + \varepsilon_i \qquad (6)$$

In (4) above, $G_i(t)$ and $R_i(t)$ represent systematic growth and random components respectively. π_i and δ_i are slopes and intercepts of individual growth curves, $a_i(t)$ is the age of individual i at time t. The X_{ij} are the values of the jth background variable for subject i, θ_j are the corresponding coefficients and ε_i are unmeasured determinants of individual growth rates. Given one of several choices of assumptions about error structure (e.g., $E[R_i(t)] = 0$; $Var_r(R_i(t)) = \sigma_r^2$, constant over all subjects and times; R_i independent of t, π_i, δ_i, and any R_i; $E(\varepsilon_i|X_i) = 0$; $Var(\varepsilon_i|\underset{\sim}{X}_i) = \sigma_\varepsilon^2$ and $Cov(\varepsilon_i, \underset{\sim}{X}_i) = \underset{\sim}{0}$), one then estimates the value-added by first regressing pretest on age and its interactions with background variables to determine estimates of individual growth rates (π_i^*) and then calculates a value-added for an individual using the expression

$$v_i = Y_i(t_2) - Y_i(t_1) - \pi_i^* \Delta_i, \qquad (7)$$

where Δ_i represents the time interval between pretest and posttest. The average of the individual value added,

$$V = \frac{\sum_{i=1}^{n} v_i}{n} \qquad (8)$$

is then an estimate of program impact.

Byrk and Weisberg's procedures appear seductively simple and broadly applicable. One models the growth process as best one can from relevant background variables and the time span over which the program measurements are obtained then attributes the remaining average increment in performance to the program. In their most recent article (Bryk et al., 1980), extensions of the basic value-added analysis model to cases where errors in

regression models are heteroscedastic, growth is non-linear, comparison group data are available, when programs are administered to non-randomly formed groups of individuals, and when aptitude-treatment interactions are believed to exist are discussed.

Important limitations of the value-added procedure are also indicated by Bryk et al. (1980). The problem of a shifting metric for measuring growth over time cannot be alleviated through value-added procedures. Whether it is simply a matter of the restandardization of scores at different age and grade levels or the more serious (analytically, at least) concern that the component skills accentuated at different ages vary, the basic complication falls outside the purview of a modeling procedure of this type.

Another limitation is the inability of the lone value-added model to deal with the lack of monotonicity of growth that occurs in schooling data with multiple years of schooling separated by summer vacations. In our companion report (Miller, 1981), a rudimentary example of this non-monotonicity arises in the Beginning Teacher Evaluation Study (BTES) data. Maddahian (1981) showed that this occurred for other BTES measures and others (e.g., Klibanoff & Haggart, 1980) have uncovered similar examples in other evaluation studies. It is not inherently impossible to apply the value-added approach to more complex growth models; it is just unclear at present how one converges substantively on an adequate model for these more complex dynamic processes.

There is no mention in the Bryk-Weisberg work of how the investigator is to alleviate the problem of measurement errors in explanatory variables.

While the concentration on a single group model (no comparison group) seemingly removes the concerns about differential attenuation of estimates the two-stage estimation process (estimate growth from pretest and predict growth increments to subtract from posttest) would appear to place greater demands for precise estimation not likely to be met by the current value-added approach. In principle the model should work best during periods when individuals are experiencing substantial observed growth which suggests that the technique is most suitable for the study of programs for younger children. But outcome measures are notoriously less reliable and stable during the preschool years and early grades of formal schooling than in later years.

Similarly, from a modern perspective, it is advantageous to be able to model program processes and examine their effects directly rather than rely simply on program participation as the indicator of program effects. As Bryk et al. (1980) demonstrate, the value-added approach can be used to estimate the effects of program characteristics on program outcomes (i.e., the value-added for a given site). Yet here, too, the errors in measuring program process characteristics as opposed to, say, ascribed individual and program characteristics are likely to inadequately reflect the true state of affairs.

Finally, there is no provision in the current literature on the value-added approach to deal with multiple measures of growth. Presumably, analysts must choose some means of arriving at a single growth measure (e.g. some form of composite) before proceeding with the value-added analysis. The alternative is to generate a series of value-added estimates, one for each combination of pre- and posttests. Our sense is that the former will typically be less than satisfactory because of the changing

character of the ideal composite over time. The latter quickly becomes unwieldy unless a reasonable scheme of interpreting the pattern of effects can be determined (e.g., see Weisberg, 1978).

In conclusion we judge the value-added approach to be a useful addition to the complement of analytical strategies for evaluating program consequences. However, the biases associated with measurement errors, changing metrics and the changing structure of behavior linger and may, in certain respects, be exacerbated. Nor is the multiple measures of outcome programs adequately considered. Nonetheless, if investigators do choose to employ the multiple analysis strategies perspective advocated here, the value-added approach will be a wise choice for inclusion in a broad range of evaluation situations.

Selection modeling. Another recently developed set of analytical approaches for dealing with selection bias can be traced to evaluations of social experiments on welfare reform (Rossi & Lyall, 1976; Stromsdorfer & Farkas, 1980). Economists working on these evaluations developed methods for adjusting for selection effects in estimating the effects of interventions. Volume 5 of the Evaluation Studies Review Annual (Stromsdorfer & Farkas, 1980) is the most comprehensive published source on selection modeling methods. Representative papers from several of the major contributors (e.g., Hausman, Heckman, Goldberger) are included along with useful discussions of the issues by the editors (Stromsdorfer & Farkas, 1980), and by Barnow, Cain, and Goldberger (1980). However, this work is rapidly developing and even recent synthetic reviews by Muthen (Muthen, 1981; Muthen & Joreskog, 1981) cannot keep up with the latest technical nuances. In addition a whole set of seemingly related techniques developed by sociologists (e.g., Tuma & Hannan, 1978; Tuma, Hannan, & Groenveld, 1978) for dynamic modeling with panel data are not even considered by the economists.

We will not attempt to describe all the particular analytical developments in our discussion of selection modeling. Instead, we try to indicate the ways in which the methods are designed to alleviate specific problems in the analysis of quasi-experimental data, point out the broad categories of analytical approaches that are currently available, and attempt to pinpoint the set of problems left unresolved by these methods. And, although we find the methods of Tuma and Hannan potentially valuable for longitudinal evaluations of social programs, the discussion will concentrate on the econometric work.[3]

The general problem that motivates the selection modeling work is the selectivity bias that results when individuals (or, for that matter,

aggregates of individuals such as schools) are self-selected (non-randomly selected) into experimental and control groups (or into different program types) or when data on the study sample are non-randomly missing (see our earlier discussion of work by psychologists on this topic (i.e., work reviewed by Reichardt , 1979). According to Stromsdorfer and Farkas (1980), "the realization that the difficulties associated with self-selection, censored samples (where some variables are unmeasured for certain individuals in the sample), truncated samples (where all variables are unmeasured for certain individuals who should be in the sample), and limited dependent variables (variables restricted to some subset of values: for example, weeks worked, which must be zero or above or the probability of being employed, which must lie between zero and one) all have a common foundation" (p. 14) was perhaps the most important statistical development in social science methodology during the 1970's. This realization led investigators to develop methods for incorporating analytical procedures for handling self-selection, censored and truncated samples, and for limited dependent variables within the general analytical model for estimating program effects.

The general analytical procedures involved in econometric selection-modeling can be sketched as follows. (This discussion draws heavily from Barnow, Cain, and Goldberger (1980), Goldberger (1979), and Muthen and Joreskog (1981).) Because of non-random assignment to program it is necessary to incorporate information about the selection process into the equation for estimating program effects. Thus, equation (3) for program outcomes,

$$Y = \gamma_1 Z + \gamma_2 X + \varepsilon^{**} \qquad (3)$$

(remember Z represents program; Z=1 for program participated and Z=0 for

group comparison) needs to be supplemented by an equation for selection into the program. A selection equation with Z as the dependent variable is specified and restrictions are placed on it to remove pre-existing differences between program and comparison gorups from the estimates of the treatment effect (γ_1 in (3)). The restrictions on the selection equation appear to be of two types. First, there must be variables that determine selection that do not affect outcome. Thus, there must be variables necessary to account for Z that are not among the X's from equation (3). Second, the functional form of the relation between X and W (true ability as identified in equation (1)) and a non-linear relation between Z and X are specified. This leads to a non-linear functional form of X in the outcome equation that is necessary to control for any relationship between Z and W that is not controlled by X.

In more formal terms we begin with three observable variables (Y, X, Z), two unobservable variables (W and t, the true selection variable; these two are anaologous in many respects to Cronbach et al.'s ideal covariate and complete discriminant) and various disturbances for the equations. Then

$$Z = \begin{cases} 1, & \text{if } t > 0 \\ 0, & \text{if } t \leq 0 \end{cases} \qquad (9)$$

and, as stated earlier program outcomes are determined by

$$Y = W + \alpha Z + \varepsilon_0 \qquad (1)$$

where ε_0 (ε in original version of equation (1)) is normally distributed, independent of W and Z, and has expectation zero and standard deviation σ_0. the relations among X, W, and t prior to selection and program participation are given by

$$W = \theta_1' X + \varepsilon_1 \tag{10}$$

$$t = \theta_2' X + \varepsilon_2 \ , \tag{11}$$

where θ_1' and θ_2' are coefficients relating X to W and t, and disturbances ε_1 and ε_2 are bivariate-normal, uncorrelated with X and ε, have standard deviations σ_1 and σ_2 and covariance σ_{12}. Thus, W and t may be related via X or through correlated disturbances. Substituting from (10) into (1) yields

$$Y = \theta_1' X + \alpha Z + \varepsilon_3 \tag{12}$$

where $\varepsilon_3 = \varepsilon_1 + \varepsilon_0$ and ε_3 and ε_2 are bivariate normal, etc., with covariance $\sigma_{23} = \sigma_{12}$. (Note that equations (12) and (3) are the same except for assumptions about ε_3.) Turning next to the selection equation, we see that $Z = 1$ is equivalent to $\theta_2' X + \varepsilon_2 > 0$ which in turn implies $\varepsilon_2 > -\theta_2' X$ and $\varepsilon_2/\sigma_2 > -\theta' X$ where $\theta' = \theta_2'/\sigma_2$. But (ε_2/σ_2) is a standard normal variable independent of X. And since Z is binary it follows that

$$E(Z|X) = \text{Prob}(Z=1|X) = 1 - F(-\theta' X) = F(\theta' X) \ , \tag{13}$$

where $F(\cdot)$ is the standard normal cummulative distribution function. Furthermore,

$$E((\varepsilon_2/\sigma_2)|X,Z=1) = f(\theta' X)/F(\theta' X) \tag{14a}$$

and

$$E((\varepsilon_2/\sigma_2)|X,Z=0) = f(\theta' X)/(1 - F(\theta' X)) \ , \tag{14b}$$

where $f(\cdot)$ denotes the standard normal density function. Equations (14a) and (14b) can be rewritten in combined form and rearranged to give

$$E((\varepsilon_2/\sigma_2) = \frac{f(\theta' X)(Z - F(\theta' X)}{(1 - F(\theta' X))F(\theta' X)}$$

$$= h(X,Z;\theta) \tag{15}$$

or, equivalently,

$$E(\varepsilon_2|X,Z) = \sigma_2 h(X,Z;\theta) \quad .$$

Also,

$$E(\varepsilon_3|X,Z) = (\sigma_{12}/\sigma_2^2)E(\varepsilon_2|X,Z) = (\sigma_{12}/\sigma_2)h(X,Z,\theta) \quad . \qquad (16)$$

Given (16), the expectation of (12) conditional on X and Z is then

$$E(Y|X,Z) = \theta_1'X + \alpha Z + (\sigma_{12}/\sigma_2)h(X,Z;\theta) \quad . \qquad (17)$$

Equation (17) is the conditional expectation function relating observable values and its parameters (θ_1, α_1, σ_{12}/σ_2, $\theta = \theta_2/\sigma_2$) can be estimated by non-linear least squares. The crucial feature of this expression is the inclusion of $h(X,Z;\theta)$ which takes the conditional relationship between X and Z into account, thus removing a source of bias (omission of a variable) in estimating α, the program effect.

In practice (17) is estimated by a two-step procedure (Heckman, 1976) whereby θ ($=\theta_2/\sigma_2$) is estimated by maximum-likelihood probit analysis of Z on X, these estimates are inserted in (15) to estimate $\hat{h} = h(X,Z;\theta)$ for each observation, and then θ_1, α, and (σ_{12}/σ_2) are estimated by linear least-squares regression of Y on X, Z, and $\hat{h}$. There is an alternative estimation procedure attributed to Maddala and Lee (1976) that operates in a similar fashion.

The essential feature of the Heckman-Maddala-Lee procedures is that they resolve the problem of selectivity bias by modifying the outcome equation for presumed selection process effects. As in simple ANCOVA, the adjustment is only necessary in those conditions where treatment selection (Z) and true ability (W) are related after controlling for the observed covariates (X). Thus, if there is no relationship between ε_1 and ε_2 ($\sigma_{12} = 0$), then no bias is introduced through selection, and the more complicated selection modeling adjustments are unnecessary.

In their review, Barnow et al. (1980) cite a number of problems with the selection modeling that require further attention:

(1) which consistent estimation procedure is best,

(2) how to deal with severe collinearity in the second-step regression,

(3) the effect of non-normal disturbances on the robustness of estimators,

(4) misspecification of the original model, and

(5) multiple selection rules.

Several of these problems have since been addressed to some degree (e.g., see Goldberger, 1980; Heckman, 1980; and Olsen, 1979 on the effects of the departures from normality).

Our reading of the current view (Muthen (1981) is the most recent and comprehensive we have seen) is that the consequences are quite serious (i.e., the procedures fail to remove the selectivity bias) when errors in the regression relation depart from normality and/or homoscedasticity (e.g., Goldberger, 1980; Hurd, 1979; Olsen, 1979) and when the functional form of the selection and/or outcome relations are misspecified. The latter can take several forms. For example, it may be that the true relationship of program and ability to outcome is nonlinear though the specification includes only linear effects. Such a situation might suggest the need for adjustments via selection modeling when a more appropriate modification requires a shift to a new functional form for the relationships.

The second form of specification problem that is likely to occur quite frequently is when relevant variables are omitted from the selectivity bias adjustment. In the Heckman procedures, this problem is manifested by leaving out variables that should be incorporated in the probit step. Again, the consequence is the failure to properly adjust estimates in

the outcome equation (Muthen, 1981 reviewing work (not currently available for citation) by Cronbach and Goldberger).

Two other concerns raised earlier about other approaches to analysis of quasi-experimental data warrant mention here. First, virtually all of the econometric discussions of selection modeling foucs on a single outcome measure. Second, the possibility of measurement errors associated with any of the observable variables (either Y's or X's) is not discussed.

Surely one would want to be able to deal with multiple outcomes and with latent exogeneous (explanatory) variables. At the least it would be helpful to state the expressions for selection and outcome modeling in terms of latent, rather than fallible observed variables. Work by Muthen, Joreskog, and Sorbom (Muthen & Joreskog, 1981; Sorbom, 1978, 1981; Sorbom & Joreskog, 1981) represent initial attempts at selection modeling with latent exogenous variables. Essentially one first estimates latent variables via LISREL and then applies the Heckman procedures using the latent variables rather than the observed set of X's. Unfortunately, these methods of estimating latent variables are currently restricted to models with strictly continuous X variables because of their reliance on maximum likelihood procedures that require multivariate normality.

The above concerns notwithstanding, the selection modeling procedures developed by economists clearly offer improvements over the ANCOVA methods described earlier. Though the demands for careful thinking about selection mechanisms are severe, the rewards of such efforts are often substantial, both analytically and substantively.

Summary. We have described in some detail both the basis for concerns about bias in quasi-experimental studies and two sets of analytical developments (the value-added approach and selection modeling) intended to remove

or adjust for bias. Both procedures are improvements over the past mainly because they employ explicit models of the phenomena believed to be responsible for the difficulties in estimating program effects. Both approaches are also adaptable to situations where there are no specific comparison or control groups (instead the effects of specific program features are to be estimated) and where panel data exists on program participants.

Neither approach directly addresses such concerns as measurement errors in the explanatory variables, changes in the scales of measurement over time and changes in the structure of behavior over time. Multiple measures of both exogenous and endogeneous variables with known scale properties are needed to gain a better grip on these problems. If these problems can be alleviated, selection and growth modeling can become even more widely useful.

Structural Equation Modeling

At various points in the discussions of improvements in analyses of non-equivalent control group designs, we encountered lingering concerns about the nature of the model specification for both selection processes and outcomes, fallible measurements, the handling of multiple indicators, changing scales of measurement and changes in the structure of behavior over time. Resolution of the first of these concerns is never complete; one progresses through obtaining better understanding of the phenomena under investigation (both its elements (constructs) and their interrelations). "Better" theories are the only answer. The combination of improvements in the accumulated wisdom on given phenomena (i.e., better thinking about how a program works and about its possible consequences) and better operationalization of the elements of one's theoretical model (i.e., more comprehensive and valid measurement of its constructs) are a necessary foundation for positive increments in the quality of investigations of social programs. Analytical methods for handling the remaining concerns cited in the opening

paragraph of this section (namely fallible measurements, multiple indicator, changing scales of measurement and structure of behavior over time) would seem to be useful to ensure that better thinking and operationalization is reflected in better data analysis and interpretation. Such analytical advances would seem to be particularly pertinent to the broad conception of large-scale program evaluation advocated here.

In theory, the techniques of structural equation modeling with latent variables (see Bentler, 1980; Bentler and Woodward, 1979; Bilby and Hauser, 1979; Goldberger and Duncan, 1973; Joreskog, 1980, 1973, 1974, 1977; Joreskog and Sorbom, 1976, 1978; Sorbom and Joreskog, 1981; Wiley, 1973) appear to be particularly well-suited for resolving several of the remaining methodological problems cited above. These techniques are designed to estimate the unknown coefficients in specified "causal" structures among latent (unobservable) variables.[4] The references cited above provide extensive discussions of the current state of work on structural equation modeling including indications of the kinds of substantive and methodological problems for which these techniques are applicable. Most of the literature addresses mainstream social research issues. However, there have been several applications in educational research contexts (see Lomax (1981) for partial bibliography of educational research applications; however, one of the most comprehensive and carefully documented applications of these methods to educational questions (namely, Munck, 1979) and recent applications with hierarchical data (Keesling, 1978; Wisenbaker, 1980; Wisenbaker and Schmidt, 1978) are not cited).

Existing applications in large-scale educational evaluations are even more limited. The best known is the exchange between Magidson (1977, 1978) and Bentler and Woodward (1978, 1979) on the effects of Head Start. Abt and

Madigson (1980) also use structural equation modeling in their evaluation of a specific school reform. Sorbom and Joreskog (1981) discuss how these techniques can be applied in evaluation research. Finally, structural equation modeling of latent variables is the primary analytical method in the longitudinal examinations of the effects of the characteristics of the educational process and students' background on academic achievement during elementary school years [conducted as part of System Development Corporation's (SDC) Sustaining Effects Study; see Wingard, 1980] and was one of the analytical methods used in SDC's cross-sectional study of the effects of instruction on the achievement growth of compensatory-education students (Wang, et. al., 1981). Given the prominence (and cost) of the Sustaining Effects Study among the set of recent large-scale evaluations in education, we are likely to see additional attempts to apply these methods, assuming of course the continuation of large-scale qualitatively oriented evaluations.

We will not attempt to recount in detail the various analytical nuances of structural equations modeling with latent variables. Instead the general strategy employed by Joreskog and his associates in their LISREL (Linear Structural, Relations) modeling will be described. We then provide a partial accounting of the specific analytical problems in program evaluations that can be addressed, at least in part, by these methods. As with the analytical developments considered earlier, we conclude with a discussion of what we perceive to be the main limitations of structural equation modeling in evaluation contexts.

Basic approach. In currently available variants of structural equation modeling, one begins with a theoretical model about the structural (perhaps causal) relations among a set of pertinent latent (unobservable) constructs (e.g., student background and ability, program and instructional quality,

schooling context, student performance). One attempts to operationalize these constructs through the collection of information on observable indicators of each construct (say, measures of aptitudes and some quality at time of program entry; measures of program and instructional characteristics (e.g., emphasis, intensity); measures of environmental characteristics (ability, composition, perceived climates); measures of cognitive, affective, and social outcomes).

The information from these indicators has an observed covariance structure (i.e., each variable yields observed estimates of variance as well as exhibiting covariation with other observed variables). One then estimates the relationships among latent variables and of latent variables to observed variables via statistical means and attempts to reconstruct the observed variance-covariance structure (matrix of variances and covariances) from the estimated variances and covariances implied by the theoretical specification. At this point one judges the acceptability of the fit of the estimated structure to the observed structure, and depending on one's perspective (there is lots of debate about what to do next), either stops or goes through another iteration of the specification-estimation process if the results are unsatisfactory.

LISREL. As we said earlier, the LISREL model developed by Joreskog and associates associates (Joreskog, 1973, 1974, 1977; Joreskog and Sorbom, 1978) is the most widely used analytical approach to estimation in structural equation modeling. This method handles a set of linear structural relations. "The variables in the equations system may be latent variables and there may be multiple indicators or causes of each latent variable...the method allows for both errors in equations (residuals, disturbances) and errors in the observed variables (errors of measurement, observational errors)...yields

estimates of the residual covariance matrix and the measurement error covariance matrix as well as estimates of the unknown coefficients in the structural equations, provided that all these parameters are known (Joreskog, 1980, p. 106)"

There are two submodels in the LISREL estimation of structural relations among latent variables. There is a structural model which specifies the relationship among latent variables. In addition, there is a measurement model which specifies the relationships of the measured variables to the unobserved constructs. Typically, there are multiple indicators of each latent construct. The interrelationships among the observed indicators of the same construct are then used to separate the presumed underlying true constructs from the irrelevant and error components of each measure.

The analyst starts with a specification of the structural model and the measurement model. If the unknown parameters in both parts of the model are identified (i.e., there are at least as many observed variances and covariances as parameters to estimate) and if the measured variables have a multivariate normal distribution, maximum-likelihood estimates for the parameters are provided along with accompanying standard errors. There are also procedures for testing lack of fit for all or part of the model (e.g., Bentler and Bonnett, 1981). More formally, the LISREL model can be specified as follows. Let $\eta = (\eta_1, \eta_2, \ldots \eta_m)$ and $\xi = (\xi_1, \xi_2 \ldots \xi_m)$ be random vectors of latent dependent (endogenous) variables and independent (exogeneous) variables. In a simple input-process-outcome model of program impact with non-experimental data, the latent variables in ξ might be socioeconomic background (ξ_1) quality of the home (ξ_2) and student ability (ξ_2). The latent dependent variables would be program quality (η_1; program quality is treated as endogenous because it is viewed as determined in part by the

specific input characteristics of students) and program outcomes such as cognitive (η_2) and social (η_3) functioning. The system of linear structural relations is given by

$$B\eta = \Gamma\xi + \zeta \,, \tag{18}$$

where $\underset{\sim}{B}$ and $\underset{\sim}{\Gamma}$ are coefficient matrices for the relations among endogenous variables (e.g., between η_1 and η_2) and of the exogeneous variables to the endogeneous varaiable (e.g., ξ_2 to η_2) and $\underset{\sim}{\zeta}$ is a random vector of residuals (errors in equation, random disturbance terms).

The vectors $\underset{\sim}{\eta}$ and $\underset{\sim}{\xi}$ are not observed. Instead we observe vectors $\underset{\sim}{Y} = (Y_1 \ldots, Y_p)$ and $\underset{\sim}{X} = (X_1 \ldots X_q)$ which are indicators of the latent endogeneous and exogeneous variables, respectively. For example, program quality (η_1) might be measured by the opportunity to learn relevant curriculum (Y_1) and the quality of the presentation of the material (Y_2). Cognitive functioning (η_2) might be measured by reading (Y_3) and mathematics achievement tests (Y_4) and social functioning by sociometric measures of friendship networks (Y_5), and teacher ratings of social functioning (Y_6). Observed indicators of the latent exogeneous variables might be family income (X_1) and mother and father's education (X_2 and X_3) for socioeconomic background (ξ_1); availability of learning resources (X_4) and parental aspirations for their child (X_5) for quality of the home (ξ_2), and pretests on reading (X_6) and mathematical skills (X_7) for student ability (ξ_3). The system of equations expressing the measurement model can be written as

$$\begin{aligned} \underset{\sim}{y} &= \underset{\sim}{\Lambda}_y \underset{\sim}{\eta} + \underset{\sim}{\varepsilon} \,, \\ \underset{\sim}{x} &= \underset{\sim}{\Lambda}_x \underset{\sim}{\xi} + \underset{\sim}{\delta} \,, \end{aligned} \tag{19}$$

where $\underset{\sim}{\Lambda}_y$ and $\underset{\sim}{\Lambda}_x$ are matrices of regression coefficients relating $\underset{\sim}{\eta}$ to $\underset{\sim}{y}$ and $\underset{\sim}{\xi}$ to $\underset{\sim}{x}$, respectively and $\underset{\sim}{\varepsilon}$ and $\underset{\sim}{\delta}$ are vectors of errors of measurement in $\underset{\sim}{y}$ and $\underset{\sim}{x}$, respectively.

(3) Measuring changes in the scaling of variables over time (e.g., Joreskog, 1979, Sorbom, 1979a).

(4) Detecting changes in the structure of behavior over time (Joreskog, 1979; Shavelson, Bolus and Keesling, 1981).

(5) Detecting differences in the structural relations across groups (e.g., Bentler and Woodward, 1978; Sorbom, 1979b, 1979c).

The first four applications select contributions targeted toward specific concerns that arise in quasi-experimental and non-experimental evaluation studies. The last application allows analysts to compare specific program alternatives (e.g., participation in Title I vs. Follow Through or High Scope vs. Direct Instruction Follow Through Models, etc.,) in a more sensitive, comprehensive, and, we believe, sensible way.

Limitations. Unfortunately, as with most analytical advances, there are important practical limitations in applying structural equation modeling in general and LISREL, specifically. The most serious and endemic problem is that the adequacy of the methods is inherently dependent on the quality of the model specification--both the limits of current theory (which constructs are pertinent) and of current operationalization through the measures one collects. Bad theory and bad data are no less bad simply because we analyze them in a sophisticated and complicated fashion. It is unclear whether the consequences of these shortcomings are more severe in structural equation models though the appearance of sophistication whenever parsimonious and simple examinations are flawed would seem to be a dangerous attribute of any analytical technique.

Another potentially serious limitation is the question of robustness of LISREL to violation of multivariate normality assumptions. Current versions of LISREL are not well-suited for such complications of discrete

If Σ represents the population covariance matrix among the p and q measured variables (13 in our hypothetical example, 6 indicators of endogeneous variables and 7 of exogeneous variables), the elements of this matrix can be expressed as functions of the elements of the four matrices of regression parametrics ($\underset{\sim}{\Lambda}_y, \underset{\sim}{\Lambda}_x, \underset{\sim}{B}, \underset{\sim}{\Gamma}$), the covariance matrix among the exogeneous latent variables ε (typically denoted by $\underset{\sim}{\phi}$), and the covariance matrices of the errors in the struvtural ($\underset{\sim}{\psi}$) and measurement ($\underset{\sim}{\theta}_\varepsilon$ and $\underset{\sim}{\theta}_\delta$) models. In application some of these elements are fixed (assigned given values), others are constrained (unknown but equal to one or more other parameters) and the remainder are free parameters to be estimated by the procedures.

Areas of application in evaluation contexts. In most practical applications of LISREL, one focusses on estimating the regression parameter matrices ($\underset{\sim}{B}, \underset{\sim}{\Gamma}, \underset{\sim}{\Lambda}_y$ and $\underset{\sim}{\Lambda}_x$). The ultimate intent is obviously to represent the true structural relationships. The specific analytical problems in program evaluation that LISREL can handle are those that arise in many social research settings. LISREL may be used to deal with a number of problems simultaneously (e.g., Madidson, 1977, Bentler and Woodward, 1978) or may be restricted to handling a single problem (e.g., perhaps obtaining estimates of latent variables for use in selection modeling, or for estimating the factor structure among observable indicators).

Particular applications include:

(1) Correcting for the effects of measurement error (e.g., Keesling and Wiley) in quasi-experiments.

(2) Taking both irrelevance (specific factors unrelated to the construct of interest but present in measured variables) and measurement errors into account (e.g., Linn and Werts, 1977).

measures of independent and dependent variables (except for the multiple group comparison application). Muthen (1979) has worked out procedures for handling certain structural models involving dichotomous variables (e.g., factor analysis of dichotomous variables) but they are not nearly as comprehensive as LISREL. Some researchers have turned to a related set of methods, partial least-squares (PLS), developed by Wold (see McGarvey and Bentler, 1980) because they do not require the multivariate normality. However, in the few empirical examples currently available, the estimates from LISREL and PLS are not very different and the rationale for PLS remains more obscure.

Despite some initial forays by Schmidt and others (Keesling, 1978; Schmidt, 1969; Wisenbaker, 1980; Wisenbaker and Schmidt, 1978), structural equation models for analyzing the hierarchical data frequently encountered in evaluations remain underdeveloped. It is simply too early to tell how to proceed in the area.

Finally, even though the primary reason many investigators turn to LISREL is its ability to estimate complex models with multiple latent constructs and multiple measurements, the practical reality is that LISREL estimation is often overwhelmed by the sheer size and complexity of such models. There are too many ways to go wrong. With large data sets with lots of parameters, practically inconsequential differences in parameters cause statistical fit indices to be significant (necessitating modification of the model). Though LISREL is capable of simultaneously estimating measurement and structural models, in practice researchers with a large number of varaibles often have to estimate these models in separate stages. And the analyses are very expensive by current standards for cost of alternative, though simplified, analytical methods. In his analyses of the SES study

of longitudinal data (Wingard (personal communication)) estimates that his typical computer run involving roughly 8 latent eonstructs with 3 to 10 indicators each costs roughly $250 and often may not even converge to within acceptable limits for the maximum-likelihood estimation.

So, again, we find ourselves with an obvious improvement in analytical methods that is applicable in large-scale program evaluation but is flawed in important respects. Clearly, structural equation modeling is a tool worth having but also one that must be used cautiously.

Concluding Remarks

In our examination of two general classes of analytical methods we have attempted to highlight why they might be considered, how they can be applied, and the limitations on their application. We could have taken each major area of analytical improvements in the past few years and treated them similarly (see, for example, the excellent review of Traub and Wolfe (in press) of the promise and problems in latent trait models for educational measurement).

But this is as it should be. Empirical investigations, be they randomized experiments or simply "passive observational studies", have their imperfections and special shortcomings. Thus, it is not surprising that there is no handy-dandy analytical method that solves all problems. The design and analysis perspective advocated here and presumably shared by Cook (1974, 1981) and Cronbach et. al. (1980), (see also Burstein (1981)) does not require that any one method be without flaws. Instead, it is the weight of the evidence from multiple analyses (and reanalyses) on perhaps overlapping but separatable questions and sets of data that should guide interpretation.

One last caveat. After beginning our work on analytical advances, we quickly became convinced that there were more fundamental problems in the area of data collection in program evaluations that greatly limit the payoff from analytical developments. In fact, we view data collection as the "Achilles Heel" of program evaluation, especially in the way it vitiates the validity of data analysis and interpretation. Elsewhere we (Burstein, Freeman and Sirotnik, 1981) have outlined our reasons for concerns about data collection. At some point, methodologists working in the area of program evaluation will devote greater attention to data collection problems. If not, the next generation of evlauation studies are destined to suffer the fate of the last generation's despite their enhanced analytical power.

Footnotes

1. We simply do not subscribe to the conspirational view of the shift in emphasis (essentially, if you can't find significant effects, change the question) as characterized in several recent accounts of the political history of the evaluation of social programs. Certainly, social programs develop a political constituency (often labeled Stakeholders) consisting of legislators, bureaucrats, service providers, program participants, members of the public as well as evaluators that have a stake in maintaining program activities. These programs also develop enemies (political and ideological) and suffer through internal bickering and lack of common perspective. Yet the interplay of competing forces surrounding any societal activity that has political, economic, and social consequences is the norm rather than the unusual. Moreover, this interplay introduces its own set of dynamics that affect the activity in complex and often unknown ways. Over time a more refined articulation of activities (expected and actual) and their consequences (expected and actual) evolve. It is only natural, then, that the search for better understanding also shifts to more sophisticated and sensitive methods for explicitly linking activities with their consequences.

2. This part of the presentation draws heavily from Barnow et. al. (1980).

3. Tuma and Hannan's work (Tuma and Hannan, 1978; Tuma, Hannan, and Groenveld, 1978) grounds the analysis of changes over time on a categorical dependent variable in a continuous-time stochastic model. They start with a continuous-time Markov model, extend it to deal with population heterogeneity (e.g., differences in background and program characteristics) and time dependence, and develop a maximum-likelihood estimation procedure for estimating the model

from what they call "event-histories" (data giving the number, timing and sequence of changes for a categorical dependent variable). These methods seem to be responsive to certain concerns addressed in the Bryk and Weisberg value-added analysis (i.e., dynamic models of change processes) as well as the econometric selection modeling (dealing with various selection problems such as attrition and systematic selection). However, the techniques are currently restricted to discrete outcome variables (e.g., decision to attend college or not; or college dropout decision) while the present review in restricted to evaluation studies in which the outcomes are viewed as essentially continuous dimensions.

4. We have chosen to use the term "structural equation" modeling rather than the label "causal" modeling more widely used in educational and psychological applications. In our view, the latter term attracts too much criticism about whether phenomena are truly "causal" as opposed to simply relational. This criticism detracts from the analytical potential inherent in these statistical aspects of the models. No one denies that practice in less than ideal (i.e., we never really know the causes in non-experimental studies (or experimental ones for that matter) and this misspecification is an inherent property of empirical social research. Misspecification, in turn, inevitably leads to flawed estimation. Nonetheless, one can conceive of a continuum of better vs. worse empirical approximations to reality. We contend that structural equation modeling with latent variables can potentially yield results that approach the "better" role of the continuum and thus should not be excluded because they are flawed (some philosopher might judge them "wrong".)

REFERENCES

Abt, W.P., & Magidson, J. Reforming schools: Problems in program implementation and evaluation. Beverly Hills, CA: Sage Publications, Inc., 1980.

Anderson, R.B. Follow Through: Testing one model of evaluation and several models of compensation. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA, April 1976 (ERIC #120-238).

Barnes, R.E., & Ginsburg, A.L. Relevance of the RMC models for Title I policy concerns. Educational Evaluation and Policy Analysis, 1979, 1(2), 7-14.

Barnow, B.S. The effects of Head Start and socioeconomic status on cognitive development of disadvantaged children. Unpublished doctoral dissertation, University of Wisconsin-Madison, 1975.

Barnow, B.S., Cain, G.G., & Goldberger, A.S. Issues in the analysis of selection bias. In E.W. Stromsdorfer and G. Farkas (Eds.), Evaluation studies review annual, volume 5. Beverly Hills, CA: Sage, 1980.

Bentler, P.M. Multivariate analysis with latent variables: Causal modeling. Annual Review of Psychology, 1980, 31, 419-459.

Bentler, P.M., & Woodward, J.A. Nonexperimental evaluation research: Contributions of causal modeling. In L.E. Datta & R. Perloff (Eds.), Improving evaluations. Beverly Hills, Ca: Sage Publications, 1979.

Boruch, R.F. Bibliography: Randomized field experiments for planning and evaluating social programs. Evaluation, 1974, 2, 83-87.

Boruch, R.F., & Cordray, D.S. An appraisal of educational program evaluations: Federal, state, and local agencies. Washington, D.C.: U.S. Department of Education.

Bryk, A.S. An intvestigation of the effectiveness of alternative adjustment strategies in the analysis of quasi-experimental growth data. Unpublished doctoral dissertation, Harvard Graduate School of Education, 1977.

Bryk, A.S., Strenio, & Weisberg, H.I. A method for estimating treatment effects when individuals are growing. Journal of Educational Statistics, 1980, 5(1), 5-34.

Bryk, A.S., & Weisberg, HI.I Use of nonequivalent control group design when subjects are growing. Psychological Bulletin, 1977, 83, 950-962.

Bureau of Education for the Handicapped. Progress toward a free appropriate public education: An interim report to Congress. Washington, D.C.: Department of Health, Education, and Welfare, 1978.

Burstein, L. Analysis of multilevel data in educational research and evaluation. In D. Berliner (Ed.), Review of Research in Education, Vol. 8. Itasca, IL: F.E. Peacock, 1980. (a)

Burstein, L. The role of levels of analysis in the specification of education effects. In R. Dreeben & J.A. Thomas (Eds.), Educational production: A microanalysis of schooling. New York: Ballinger Press, 1980. (b)

Burstein, L. Investigating social programs when individuals belong to a variety of groups over time: Implications for Follow Through research and evaluation. Paper presented at the National Institute of Education Conference on Follow Through Research and Development, Pittsburgh, PA, March 1981.

Cain, G.C. Regression and selection models to improve nonexperimental comparisons. In C.A. Bennett and A.A. Lumsdain (Eds.), Evaluation and experiment, some critical issues in assessing social programs. New York: Academic Press, 297-317.

Campbell, D.T., & Erlebacher, A.E. How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. Hellmuth (Ed.), The disadvantaged child. Vol. 3, New York: Brunner/ Mazel, 1970.

Cicirelli, V.G., et al. The impact of Head Start: An evaluation of the effects of Head Start on children's cognitive and affective development. Athens, OH: Ohio University and Westinghouse Learning Corporation, 1969.

Cline, M.D., et al. Education as experimentation: Evaluation of the Follow Through planned variation model (Vols. 1A, 1B). Cambridge, MA: Abt Associates, 1974.

Cohen, D.K., & Garet, M.S. Reforming educational policy with applied social research. Harvard Educational Review, 1975, 45(1), 17-43.

Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, S., Weinfeld, F.D., & York, R.L. Equality of educational opportunity. Office of Education, U.S. Department of Health, Education, and Welfare. Washington, D.C.: U.S. Government Printin Office, 1966.

Cook, T.D. Dilemmas in evaluation of social programs. In M.B. Brewer and B.E. Collins (Eds.), Scientific inquiry and the social sciences, San francisco, CA: Jossey Bass, Inc., 1981, 257-287.

Cook, T.D., & Campbell, D.T. Quasi-experimentation. Chicago, IL: Rand McNally College Publishing Company, 1979.

Coulson, J.E. National evaluation of the Emergency School Aid Act (ESAA): Review of methodological issues. Journal of Educational Statistics, 1978, 3(1), 1-60.

Coulson, J.E., Ozenne, D.G., Hanes, S.D., Bradford, C., Doherty, W.J., Suck, G.A., & Hemenway, J.A. The third year of Emergency School Aid Act (ESAA) implementation. System Development Corporation, TM-5236/014/00, 1977.

Cronbach, L.J. Design educational evaluations. Stanford Evaluation Consortium, Stanford University, 1978.

Cronbach, L.J., Ambron, S.R., Dornbusch, S.M., Hess, R.D., Hornik, R.C., Phillips, D.C., Walker, D.F., & Weiner, S.S. Toward reform of program evaluation. San Francisco, CA: Jossey-Bass, Inc., 1980.

Cronbach, L.J., Rogosa, D.R., Floden, R.E., & Price, G.G. Analysis of covariance in nonrandomized experiments: Parameters affecting bias. Occasional Paper, Stanford Evaluation Consortium, Stanford University, 1977.

Cross, C.T. Title I evaluation -- A case study in congressional frustration. Educational Evaluation and Policy Analysis, 1979, 1(2), 15-22.

Daillak, R.H., & Alkin, M.C. Qualitative studies in context: Reflections on the CSE studies of evaluation use. Los Angeles, CA: University of California, Los Angeles, Center for the Study of Evaluation, 1981.

Firebaugh, G.L. Groups as contexts and frog ponds. In K. Roberts and L. Burstein (Eds.), New directions in the methodology of social and behavioral research. San Francisco, CA: Jossey-Bass Publishers, 1980.

Goldberger, A.S. Abnormal selection bias. 9006, Social Systems Research Institute, University of Wisconsin, Madison, 1980.

Goldberger, A.S. Methods for eliminating selection bias. Memorandum, Department of Economics, University of Wisconsin, Madison, 1979.

Goldberger, A.S. Selection bias in evaluating treatment effects: Some formal issufrations. Discussion paper 123-72. Madison: Institute for Research on Poverty, 1972.

Haney, W. A technical history of the national Follow Through evaluation, Vol. V, The Follow Through planned variation experiment. Cambridge, MA: The Huron Institute, 1977.

Heckman, J. The common structure of statistical models of truncation, sample selection, and limited dependent variables and a simple estimator for such models. Annals of Economic and Social Measurement, 1976, 5, 475-492.

Heckman, J. Addendum to 'sample selection bias as a specification error'. In E.W. Stromsdorfer and G. Farkas (Eds.), Evaluation studies review annual, vol. 5. Beverly Hills, CA: Sage Publications, 1980.

House, E. Evaluation with validity. Beverly Hills, CA: Sage Publications, 1980.

House, E. The logic of evaluative argument. CSE Monograph. Los Angeles, CA: CEnter for the Study of Evaluation, 1977.

House, E., Glass, G.V., McLean, L.D., & Walker, D.F. No simple answer: Critique of the 'Follow Through' evaluation. Harvard Educational Review, 1978.

Joreskog, K.G. A general method for analysis of covariance structures. Biometrika, 1970, 57, 239-251.

Joreskog, K.G. A general method for estimating a linear structural equation system. In A.S. Goldberger & O.D. Duncan (Eds.), Structural equation models in the social sciences. New York: Seminar Press, 1973.

Joreskog, K.G. Analyzing psychological data by structural analysis of covariance matrices. In D.H. Krantz, R. Atkins, D.Luce, and P. Supper (Eds.), Contemporary developments in mathematical psychology, Vol. II, San Francisco, CA: D.W. Freeman & Co., 1974.

Joreskog, K.G. Structural equations models in the social sciences: Specification, estimation, and testing. In P.R. Krishnaiah (Ed.), Applications of statistics, Amsterdam: North Holland Publishers, 1977.

Joreskog, K.G., & Sorbom, D. Statistical models and methods for analysis of longitudinal data. In D.J. Aigner & A.S. Goldberger (Eds.), Latent variables in socioeconomic models, Amsterdam: North Holland Publishers, 1977.

Joreskog, K.G., & Sorbom, D. Advances in factor analysis and structural equation models. Cambridge, MA: Abt Books, 1979.

Klibanoff, L.S., & Haggart, S.A. Report #8: Summer growth and the effectiveness of summer school. Technical Report #8 from the Study of the Sustaining Effects of Compensatory Education on Basic Skills, System Development Corporation, Santa Monica, CA, 1981

Maddahian, E. Statistical models for the study of cognitive growth. Unpublished doctoral dissertation, University of California, Los Angeles, 1981.

Maddala, G.S., & Lee, L.F. Recursive models with qualitative endogenous variables. Annals of Economic and Social Measurement, 1976, 5, 525-545.

Miller, M.D. Measuring between-group differences in instruction. Unpublished doctoral dissertation, University of California, Los Angeles, 1981

Munck, I.M.E. Model building in comparative education. Stockholm, Sweden: Almqvist & Wiksell International, 1979.

Muthen, B. Some categorical response models with continuous latent variables. In K.G. Joreskog & H. Wold (Eds.), Systems under indirect observation: Causality, structure and prediction. Amsterdam: North Holland Publishing Company, 1981.

Muthen, B., & Joreskog, K.G. Selectivity problems in quasi-experimental studies. Presented at the Conference on Experimental Research in Social Sciences, University of Florida, Gainesville, 1981.

National Institute of Education. Evaluating compensatory education. A report on the NIE Compensatory Education Study, 1977.

Olsen, R.J. Tests for the presence of selectivity bias and their relation to specifications of functional form and error distribution. Working paper No. 812, Yale University, 1979.

Raizen, S.A., & Rossi, P.H. (Eds.) Program evaluation in education: When? How? To what ends? Washington, D.C.: National Academy Press, 1981.Reichardt, C.S. The statistical analysis of data from non-equivalent group designs. In T.D. Cook and D.T. Campbell (Eds.), Quasi-experimentation, Chicago: Rand McNally, 1979.

Rogosa, D. Politics, process, and pyramids. Journal of Educational Statistics, 1978, 3(1), 79-86.

Rossi, P.H., & Lyall, K.C. Reforming public welfare: A critique of the negative income tax experiments. Russell Sage Foundation, 1976.

Sirotnik, K.A., & Oakes, J. A contextual appraisal system for schools: Medicine or madness? Educational Leadership, 1981 (in press).

Stallings, J.A., & Kaskowitz, D.H. Follow Through classroom observation evaluation 1972-1973. Stanford Research Institute, August 1974.

Smith, M.S., & Bissell, J.S. Report analysis: The impact of Head Start. Harvard Educational Review, 1970, 40, 41-104.

Strenio, J.F., Bryk, A.S., & Weisberg, H.I. An individual growth model perspective for evaluation of educational programs. Proceedings of the social science section, Annual Meeting of the American Statistical Association, 1977.

Stromsdorfer, E.W., & Farkas, G. Evaluation studies review annual, Vol. 5. Beverly Hills, CA: Sage Publications, 1980.

Sorbom, D. An alternative to the methodology for analysis of covariance. Psychometrika, 1978, 43, 381-396.

Sorbom, D. Structural equation models with structured means. To appear in K.G. Joreskog and H. Wold (Eds.), Systems under indirect observation: Causality, structure, prediction. Amsterdam: North Holland Publishing Co., 1981.

Sorbom, D., & Joreskog, K.G. The use of structural equation models in evaluation research. Presented at the Conference on Experimental Research in Social Sciences, University of Florida, Gainesville, 1981.

Tuma, N.B., Hannan, M.T., & Goroenveld, L. Dyanmic analysis of event histories. In E.W. Stromsdorfer, & G. Farkas, Evaluation studies review annual, Vol. 5, Beverly Hills, CA: Sage Publications, 1980.

Wargo, M.J. An evaluator's perspective. In M.J. Wargo & D.R. Green (Eds.), Achievement testing of disadvantaged and minority students for educational program evaluation. McGraw-Hill, 1977.

Tuma, N.B., & Hannan, M.T. Approaches to the censoring problem in analysis of event histories. In, K. Schuessler (Ed.), Sociological Methodology, San Francisco: Jossey-Bass, 1979.

Webb, N.M. Group process: The key ot learning in groups. In K.H. Roberts, & L. Burstein (Eds.), Issues in aggregation, vol. 6, New directions for methodology of social and behavioral science, San Francisco, CA: Jossey-Bass, 1980.

Wisler, C.E., & Anderson, J.K. Desinging a Title I evaluation system to meet legislative requirements. Educational Evaluation and Policy Analysis, 1979, 1(2), 47-56.