This is to certify that the

thesis entitled

ON THE EVALUATION OF SOCIAL ACTION PROGRAMS BY THEORY TESTING: AN EXAMPLE FROM COMPENSATORY EDUCATION

presented by

Jonathan Shapiro

has been accepted towards fulfillment of the requirements for

Ph.D. degree in Political Science

Cleo Cherryholmes

Major professor

Date 9/12/80

O-7639

ON THE EVALUATION OF SOCIAL ACTION PROGRAMS BY THEORY TESTING: AN EXAMPLE FROM COMPENSATORY EDUCATION

By

Jonathan Shapiro

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Political Science

1980

ABSTRACT

ON THE EVALUATION OF SOCIAL ACTION PROGRAMS BY
THEORY TESTING: AN EXAMPLE FROM
COMPENSATORY EDUCATION

By

Jonathan Shapiro

The goal of this dissertation is to demonstrate that the use of experimental design for evaluation research is not unproblematic. It is argued that the methodological properties of the experiment more likely satisfy the needs and interests of researchers rather than decision makers. However, the data generated in evaluation is to be utilized by decision makers rather than by researchers. The problems arising from the gap between method and informational needs in evaluation usually are manifested as nonusage by the decision make when policy is made.

In response to this, a research design is proposed which requires an evaluator to specify a theoretical model of the process by which program activities lead to outcomes and compare groups to all points in this process. The design is based on an argument raised by Edward Suchman concerning the conduct of evaluation. Suchman criticizes the conventional evaluation design which tends to focus on outcomes while neglecting process. He suggests that this

narrow focus tends to leave undetected important information about programs particularly when a program is shown to be ineffective.

The design created in the dissertation is used to reanalyze the data from the Ohio-Westinghouse evaluation of Head Start. The results indicate that there are conditions under which the proposed design is feasible and will generate greater amounts of useful evaluation data than conventional designs.

ACKNOWLEDGMENTS

I would like to thank my committee--Professors John Aldrich, Charles Ostrom, Frank Pinner, and especially Cleo Harlan Cherryholmes for their support and assistance through this long process. Others who were a necessary part of my success include my parents, Simon and Sara Shapiro; my wife, Heidi; and two special teachers, Myron Aranoff and Barry Rundquist. I would also like to thank Mrs. Nancy Heath for singlehandedly typing all the drafts of this dissertation.

TABLE OF CONTENTS

	Page
LIST OF TABLES	iv
INTRODUCTION	1
Chapter	
I. EXPERIMENTATION AND EVALUATION	3
The Status of Experimentation in Evaluation Research	3
A Critique of the Role of Experimentation in Evaluation	18
II. A META-THEORY AND METHODOLOGY FOR EVALUATION RESEARCH	32
III. AN APPLICATION OF THE PROPOSED RESEARCH DESIGN TO AN EMPIRICAL EXAMPLE: THE SPECIFICATION	58
IV. AN APPLICATION OF THE PROPOSED DESIGN TO AN EMPIRICAL EXAMPLE: THE DATA RESULTS	79
V. SUMMARY AND CONCLUSIONS	119
APPENDICES	122
A. CONSTRUCTS	123
B. SOLVING FOR THE REDUCED FORM	145
REFERENCES	152

LIST OF TABLES

Table		Page
1.	Variables Used in the Data Analysis	84
2.	A Comparison of the R-Square for the OLS and 2OLS Estimates of the Full Causal Model	87
3.	A Comparison of the R-Square for the OLS and 2SLS Estimates of the Full Causal Model for the Treatment and Control Samples	93
4.	Results of the OLS Estimations of the Full and Predictive Models for the Treatment Group	95
5.	Results of the OLS Estimations of the Full and Predictive Models for the Control Group	98
6.	A Comparison of Outcome Measures Between the Treatment and Control Groups by ANOVA and ANCOVA	101
7.	Results of the Stage One Chow Tests for Differences Between the Treatment and Control Groups on Significant Structural Variables	106
8.	Results of the Stage Two Chow Test for Differences Between the Treatment and Control Groups on Significant Variable x Treatment Interactions	107
9.	A Comparison of the Proportion of Explained Variance in the Treatment and Control Group Predictive Models .	109
10.	Results of the OLS Estimation of the Full and Predictive Causal Models for the Combined Sample	113

INTRODUCTION

This dissertation advances an alternative to the experimental research design conventionally employed in program evaluation. The alternative, based on the arguments of Edward Suchman (1966), advocates the analysis of a complex set of relationships to assess the effectiveness of social action programs. Suchman has challenged the inferences generated by experimental design as weaker than those created by the testing of theory. At issue is the way in which an evaluator goes about gathering data in order to make maximally valid inferences about program impact. It is the function of the research design to indicate to the evaluator the manner in which data are to be obtained. Thus, the dissertation will examine the logic underlying experimental design, i.e., the arguments why data should be collected in that particular way and contrast that with a design calling for theory based evaluation data.

The design constructed in this dissertation is based on Suchman's meta-theory of evaluation. A basic contention of the dissertation is that while Suchman's arguments about experimentation and theorizing are essentially correct, his position has not been seriously entertained by evaluators because the argument is incomplete. Demonstrating the limitations of experimentation and the benefits of theory based evaluation is not compelling unless a feasible research design can also be specified. Therefore, the dissertation will attempt to complete the argument for theory based evaluation by constructing, implementing, and critically examining a design based on Suchman's meta-theory of evaluation.

The structure of the dissertation is as follows. The first chapter discusses the nature and functions of research designs in general and the preeminent role of experimental designs in evaluation research in particular. The final section of Chapter I will present Suchman's criticism of the nature of the inferences generated by experimental designs. Chapter II contains a discussion of Suchman's meta-theory of evaluation and a research design loosely predicated on that meta-theory. Chapters III and IV represent the attempt to implement the design by reanalyzing the data from the Westinghouse-Ohio evaluation of Project Head Start (1969). The last chapter will assess how well the research design performed and consider the potential role of theory based evaluation data in future evaluation efforts.

One note of clarification: the presentation and critique of experimental design refers explicitly to the experiment as conceived and described in Campbell and Stanley (1963) and Cook and Campbell (1979), rather than to experimentation at a generic level. These two books seem to have the largest impact on current evaluation.

CHAPTER I

EXPERIMENTATION AND EVALUATION

The Status of Experimentation in Evaluation Research

According to Carol Weiss, "Experimental design has long been considered the ideal for evaluation" (Evaluating Action Programs: 6). By experiment is meant that one or more treatments (programs) are administered to some set of persons (or other units) drawn at random from a specified population; and that observations (or measurements) are made to learn how (or how much) some relevant aspect of their behavior following treatment differs from like behavior on the part of an untreated or control group also drawn at random from a specified population; and that observations (or measurements) are made to learn how (or how much) some relevant aspect of their behavior following treatment differs from like behavior on the part of an untreated or control group also drawn at random from the same population (Riecken and Boruch: 3, emphasis theirs). The initial task in this chapter is to determine why the experiment, as defined above, is accorded the status given by Weiss and other prominent evaluation researchers. To assess the utility of experimentation, the exact function of the experiment, as an integral part of evaluation methodology, must be explicated.

An experimental design is a specific type of research design. In general, a research design is a set of instructions to an investigator indicating the activities required to secure "adequate and proper data to which to apply statistical procedure" (Campbell and Stanley: 1).

As Kerlinger describes it (p. 327):

> Design is data discipline. The implicit purpose of all research design is to impose controlled restrictions on observations of natural phenomena. The research design tells the investigator, in effect: Do this and this; don't do that or that; be careful with this; ignore that; and so on. It is the blueprint of the research architect and engineer. If the design is poorly conceived structurally, the ultimate product will be faulty. If it is at least well conceived structurally, the ultimate product has a greater chance of being worthy of serious scientific attention.

For research involving an intervention, the design would stipulate how the treatment and control group are drawn (i.e., random selection and random assignment), when the intervention is to be administered, and when observation on the groups are to be taken.

By adequate and proper data is meant date which lead to maximally valid inferences about the effect and generalizability of an intervention. When attempting to attribute the reasons for a particular sample outcome, a researcher must be concerned about internal validity. An inference is internally valid when it correctly identifies the cause of the observed sample outcome. Threats to the internal validity of an inference are factors responsible for the outcome which the researcher fails to identify.

In evaluation, the intervention is usually a (social action) program. Thus, the evaluative inference is the assertion that a particular sample outcome is due to the program under analysis. The evaluative inference is internally valid when possible rival factors (other than the program) are eliminated as plausible causes of the outcome.

Externally validity concerns the ability of the inference to hold true in other samples or populations. External validity is basically a function of the generalizability of the sample as well as the nonreactivity of the experimental setting. As opposed to internal validity, external validity appears to be basically independent of the research design. (One exception is whether or not the research design calls for pretesting.) Therefore, to discuss the role of experimentation in evaluation is to discuss the most commonly accepted set of rules for gathering evaluation data assumed to be maximally internally valid. [Cook and Campbell note that for both theoretical and practical research, internal validity should always be of paramount concern (p. 83)].

The orientation toward experimentation in evaluation is, at least in part, a function of the notion that social action programs are structured in such a manner that the research setting resembles a laboratory situation. Rossi has observed that, "In principal, the evaluation of social action programs appears to be most appropriately undertaken through the use of experimental designs" (Caro: 239). He argues that important aspects of experimentation are present in social action settings. Two examples are the control sponsoring agencies exert over their programs and the general condition that ameleorative programs are not intended for general consumption, suggesting the availability of natural control groups.

Of even greater significance than the notion that evaluation can be done by experimentation is the normative assumption that evaluation should be done by experimentation. At least three identifiable arguments contribute to the widespread acceptance of this assumption among evaluators. The first of these maintains that the practices of the natural sciences should serve as a model for knowledge gathering in the social sciences. Therefore, adopting the primary methodology of natural science, the experiment is a prerequisite for the successful accumulation of knowledge in social science. In 1935 A. Stephen Stephan stated (Caro: 40):

> Students of human behavior have long envied the chemist and physicists who are releasing the secrets of nature through experimentation and laboratory procedure. The exacting methods of the laboratory have been responsible for the phenomenal advance of the physical sciences. The gap between the accumulated knowledge of the physical sciences and the social sciences is largely explained by the differences in the exact methods of the former and the floundering methods of the latter.

The essence of Stephan's paper was that the awakening enthusiasm in government agencies for rational and comprehensive planning meant that social scientists would be able to construct large scale experiments, which he considered to be the key element in the success of the natural sciences.

The second argument for the necessity of experimentation in evaluation is a logical extension of the position that policy making should be an experimental enterprise, i.e., policies may be enacted even if their outcomes are uncertain or unknown. Campbell (1971) has argued that effective evaluation of social policy can only occur when policies are treated as experiments. When administrators justify their policies by declaring in advance what the outcomes will be, they lose the flexibility to make use of evaluations which may indicate the need to modify or abandon a particular policy. If the justification for a policy was the need to attempt to resolve a serious social problem, rather than asserting some certain outcome, then the failure of a policy to create change could be tolerated. Thus, Campbell suggests that by justifying reform on the basis of the urgency of social problems as opposed to the certainty of outcomes, a policy could be regarded as only a potential solution and may be discarded in favor of an alternative when it is shown to be ineffective. The most obvious way to evaluate experimental programs would be by using experimental research designs.

Alice Rivlin (1971) discusses two policies that have been implemented in the manner suggested by Campbell. One was the New Jersey negative income tax experiment (pp. 94-102) and the other was the Follow Through program (102-106). In each case, evaluation was accomplished by treating program participants as the experimental group, creating control groups, and examining group differences on selected outcome measures; a typical (quasi) experimental design.

Gilbert and Mosteller (1972) argue that such an experimental approach is necessary to enact effective school policy while Rivlin observes on a more general level, ". . . unless we begin searching for improvements and experimenting with them in a systematic way, it is hard to see how we will make much progress in increasing the effectiveness of our social services" (Rivlin: 119). More recently, Bennett and Lumsdaine (1975) have noted that a good many decades of failure to solve basic social problems suggest that experimentation with new kinds of solutions is going to be necessary. A better future, they predict, ". . . may accrue to societies which actively seek it through innovation and experiment" (p. 534).

The third, and most pervasive, argument for experimental evaluation is based on the desirable methodological and theoretical properties ascribed to experimentation and in particular the effect of random assignment to groups. Random assignment, according to Riecken and Boruch,

> . . . is the essential feature of true experiments because it provides the best available assurance that experimental subjects (as a group) are so much like control subjects in regard to ability, motivation, experience, and other relevant variables (including unmeasured ones) that differences observed in their performance following treatment can safely be attributed to the treatment and not other causes with a specific degree of precision (p. 4).

Their main point is that, as opposed to passive research designs such as correlational studies, experiments ". . . generally allow inferences of superior dependability about cause and effect" (Riecken and Boruch: 9). It is the notion of cause, and attributing cause, that truly lies at the base of the argument for experimentation. Consequently, if the value placed on causal attribution in evaluation can be deduced, an explanation for the value placed on experimental design will have been generated.

In one sense, the value placed on causal attribution and the role of experimentation in evaluation are easily explained. Most evaluators, and authors of evaluation literature, come from a psychological or educational psychological tradition where the dominant method is experimentation and the dominant research aim is causal attribution. Cook and Campbell (p. 9) acknowledge the relationship between experimentation and causality by stating,

> . . . the deliberately intrusive and manipulative nature of experimentation is closely related to some philosophy of science conceptions of a particular type of cause, to most persons' everyday understanding of the notion of cause, and to the way that most changes would have to be made to improve our environment by introducing successful new practices and weeding out harmful ones.

Thus, the issue now appears to be why are causal attribution and experimentation valued in psychology and educational psychology.

The research tradition in psychology and educational psychology differs from that, for example, of economics or political science in a very definite manner. The use of experimentation to assess causality would seem to preclude the sorts of empirical descriptions, in the form of behavioral models, that are common in economics and political science. Given experimentation, empirical research is the assessment of the degree of disruption of some state of nature due to the researcher's interference (intervention) in that state. The causal assertions afforded by experimental design are not cause and effect hypotheses about why things are the way they are, but rather, assertions that changes in the normal state of affairs were caused by the intervention. In other words, the researcher can make causal inferences about why the treatment group differed from the control group, but not about the control group itself. This is why outcomes in experimentation are generally measured in terms of group differences or gain scores as opposed to the levels of the outcomes themselves. To the degree that differences between groups are useful pieces of information, the experiment, with its power to maximize the internal validity of causal inferences, is a critical tool of the empirical researcher. However, if explanations for natural states are the research goal, e.g., how does political preference occur, what leads to lower or higher intelligence, the experiment is not truly structured to provide such information. It is important to keep in mind that the causal inferences afforded by the experiment are of a particular (comparative) nature only.

The argument that experimental designs generate causal inferences of maximum probability can be made both from a philosophy of science and statistical analysis perspective. It has been noted by several authors (Cook and Campbell, 1975; Riecken and Boruch, 1974; Gilbert and Mosteller, 1972) that the assessment of causality is most easily accomplished through intervention and manipulation rather than by passive observation. In the context of disparaging correlational studies, Cook and Campbell (1975, p. 287) state

> Essential to the idea of an experiment is a deliberate, arbitrary human intervention--a planned intrusion or disruption of things as usual. Probably the psychological roots of the concept of cause are similar. Causes are preeminently things we can manipulate deliberately to change other things. Evidence of cause best comes as a result of such manipulation.

Thus the surest way to establish causality is to introduce it mechanistically, and utilize the change in some system or state of affairs as evidence of the causal impact of an intervention. George Box has succiently verbalized this notion in experimentation, "to find out what happens to a system when you interfere with it, you have to interfere with it (not just passively observe it)" (Gilbert and Mosteller: 372).

Again, while it is undoubtedly true that causal attribution is easiest when one controls the cause, the issue remains whether or not those types of causes are of interest, particularly with respect to evaluation research. The notion of control over the causal factor is not the only reason for favoring experiments. Cook and Campbell (1979) note that diverse arguments concerning conditions for causal attribution, such as David Hume's analysis of cause, Mill's Canons of Logic and Popper's falsificationism can all be fit to the experimental design.

According to Cook and Campbell (p. 10), Hume stressed three conditions for inferring cause and effect: (a) contiguity between the presumed cause and effect; (b) temporal precedence, in that the cause had to precede the effect in time; and (c) constant conjunction, in that the cause had to be present whenever the effect was obtained. By applying an intervention to a treatment group, shortly thereafter observing outcomes, and noting that the effect in the treatment group was not present in the control group, Hume's conditions could be fulfilled when the intervention did have an impact.

John Stuart Mill (Cook and Campbell: 18) held that three conditions were necessary for inferring cause: first, it had to precede the effect in time; second, the cause and effect had to be related; and third, other explanations of the cause-effect relationship had to be eliminated. Mill's methods of agreement, disagreement, and concomitant variation apply to the condition of eliminating alternative causal explanations. The Method of Agreement states that an effect will be present when the cause is present; the Method of Difference states that the effect will be absent when the cause is absent; and the Method of Concomitant Variation implies that when both of the above relationships are observed, causal inference will be all the stronger since other interpretations of the covariation between the cause and effect can be ruled out (Cook and Campbell: 18). Again, note that the presence of an effect in the treatment group, its absence in the control group, and the two together fit Mill's methods of causal inference. Mill's Method of Concomitant Variation reduces the plausability of spurious relationships, since any third variable truly causing the effect would be present in the control group which is identical to the treatment group except for the intervention.

"Among more contemporary philosophers of science, Popper (1959) has been the most explicit and systematic in recognizing the necessity of basing knowledge on ruling out alternative explanations of phenomena so as to remain . . . with only a single conceivable explanation" (Cook and Campbell: 20). Popper's basic notion is that one can never prove an hypothesis, if for all trials the data fit the hypothesis, one can only say the hypothesis has not yet been disconformed. He suggests that while hypotheses cannot be proven, they can be corroborated if they are not falsified (Salmon: 24). He suggests that highly corroborated hypotheses are required for explanation and prediction. Cook and Campbell see in his arguments the implication that the only process available for establishing a scientific theory is one of eliminating plausible rival hypotheses (p. 21). Again, such a concept is present in experimentation for the effect of randomization is to rule out rival plausible hypotheses by maximizing the probability of between group equivalence. Furthermore, Popper suggests that the hypothesis under test needs to be the most ampliative, that is, highly falsifiable, of the group of hypotheses under consideration (Salmon: 25). When one considers that the rival hypotheses, that is, threats to internal validity are simple notions such as maturation or history, the intervention is likely to imply a much more complex hypothesis, thereby meeting the condition of greatest falsifiability.

These three arguments concerning the conditions under which cause and effect can be inferred are based on the logical arguments about how cause may be revealed. In each case, the structure of experimental design would allow the conditions to be met when a causal relationship did exist. However, any relationship between variables requires statistical confirmation. To the extent that cause is implied by strong statistical relationships, it can be argued that the structural of experimental design is well suited to finding these statistical "causal" relationships when they exist due to the ability of the design to perform an essential statistical task, namely, controlling variance.

Kerlinger (1973) summarizes the manner in which experimental design controls variance as the "maxmincon principle" (p. 307), which is an acronym for the design functions of maximize the systematic variance or experimental variance; control extraneous systematic variance; and minimize error variance. Each variance type plays a particular role in the statistical analysis of data.

Experimental variance is the variance on outcome measures between groups due to treatment. By maximizing experimental variance, the researcher is able to "pull apart" (Kerlinger: 308) or distinguish among alternative treatments or between the treatment and no treatment condition. It is necessary to give the variance of a relationship the chance to show itself, therefore research should be conducted such that experimental conditions are as distinct as possible. If a tutoring program were to be evaluated, a true relationship would have a better chance of being discovered if the treatment group were to have one hundred hours of tutoring rather than ten.

The control of systematic extraneous variance is the statistical version of the issue of internal validity. Since effect is measured in terms of variance, an inference about an intervention would be internally valid, from a statistical perspective, when the variance would be due to treatment and not variance due to other factors. It is in the control of extraneous variance where the experimental design truly shines since, "theoretically, randomization is the only method of controlling all possible extraneous variables" (Kerlinger: 309-310). It is with respect to the control of extraneous variance that the experiment is a most powerful research tool.

Error variance is variance due to random factors which are basically uncontrollable and unpredictable. The random nature of error variance is assumed to be a function of the myriad of factors affecting relationships in all different ways. The minimization of error variance is based on two principles: (1) the reduction of measurement error through controlled conditions and (2) an increase in the reliability of measures (Kerlinger: 312). The minimization of error variance again allows the experimental variances to demonstrate significance since:

$$V_t = V_b + V_e$$

where V_t is the total variance in a set of measures

V_b is the between groups variance, presumably due to the influence of treatment

V_e is the error variance

Obviously, the larger the V_e, the smaller the V_b must be for a given amount of V_t (Kerlinger: 313). As Kerlinger points out, the equations for the t and f statistics, where

$$t = \frac{\text{statistic}}{\text{standard error of the statistic}}$$

and

$$F = \frac{V_b}{V_e}$$

indicate the same thing: in order for the numerators of the fractions on the right to be accurately evaluated for significant departures from chance expectations, the denominators should be accurate measures of random error (p. 313). Note that minimization of error variance, ceteris paribus, leads to maximum likelihood that significant differences will be assessed as such.

Houston (1972) has summarized the effects of variance control due to randomized experimental design (p. 62).

1. The model provides a specific inference regarding the existence of effects which can be causally attributed to program components and their interactions; estimates of the magnitude of these effects are also provided.

2. The stability of these inferences is known, being specified by the size and power of the statistical tests afforded by the model.

3. The generalizability of these inferences is known, being specified by the experimental design.

4. The internal validity of these inferences rests upon assumptions generally accepted by behavioral scientists (i.e., the effect of randomization).

Along with the desirable methodological properties of experimental design, the structure of an experiment lends itself to sound research practice. Riecken and Boruch (1974) argue that the very process of implementing an experimental evaluation helps to clarify the nature of the social problem under attack. They suggest that designing an experiment ". . . focusses attention on the variables of specific interest, forcing administrators to specify objectives and operations, thus linking the data with the policy decision to be made" (p. 6).

Taken together these arguments have served to entrench experimental design as the ideal method for assessing the effectiveness of social action programs. The regard that researchers out of the psychology and educational psychology tradition have for experimental design can be exemplified by the following three testimonials. In Experimental and Quasi-Experimental Designs for Research, Campbell and Stanley (1963) unabashedly describe experimentation

> . . . as the only means for settling disputes regarding educational practice, as the only way of verifying educational improvements and as the only way of establishing a cumulative tradition in which improvements can be introduced without the danger of a faddish discard of old wisdom in favor of inferior novelties (p. 2).

Hatry, Winnie and Fish (1973) label the controlled experiment the "Cadillac of program evaluation" (p. 56), while Tharp and Gallimore (1979) assert that ". . . the true experiment with random assignment of subjects to treated and untreated conditions remains the ideal effectiveness test . . ." (p. 41).

It is the position of this dissertation that these assertions are basically incorrect.

A Critique of the Role of Experimentation in Evaluation

Unlike critiques of the use of experimental design in evaluation in terms of feasibility (Weiss, 1972; Weiss and Rein, 1969), limitations in the conduct of experiments (Boruch, 1975) and ethical and moral considerations (Rossi and Williams, 1972), the argument advanced in this dissertation is that evaluation and experimentation are basically incompatible owing more to the structure of evaluation than to the structure of experimentation. That is, experimental design is a structure which permits causal inferences, both in the philosophy of science and in the statistical sense, and is feasible, as evidenced by randomized field experiments (Clark and Walberg, 1968; Crane and York, 1970) and Cook and Campbell's chapter on conditions under which randomization can take place (1979: 341-386). The problem is that the types of evaluation data of greatest utility cannot be generated by an experimental design. A brief description of evaluation may help to make this argument clearer.

Evaluation research is part of a decision-making process. Regardless of the unit of analysis: a school reading program, a state wide or national welfare program on federal level policy positions, the one factor that makes all types of evaluation similar is that the data are generated to facilitate some decisions about that which is being analyzed. In effect, although the research is carried out by an evaluator, the data are generated for use by a policy maker, for whom the evaluator is merely an agent. For evaluation to be useful, it must be carried out with the aims of the decision maker in mind. This is where evaluation differs from academic experimental research, the researcher is not the person for whom the data are being collected, yet this distinction appears generally to be overlooked by evaluators, with some particularly negative outcomes.

If the research goals of academic researchers and decision makers can be compared and contrasted, it can be shown why experimental design may be appropriate for academic but not for evaluation research. In a very real sense, the psychologist or educational psychologist, as academic researcher, has a great deal of freedom in choosing research topics. Hypotheses generated by the researcher can be general or specific, relevant or not relevant to practical concerns. Except for tenure decisions, the researcher has no real stake in the outcomes. If an hypothesis is falsified, another can be generated. Causal inferences can be made if the researcher is willing to ignore process and concentrate on outcomes. The causal inference concerns differences between groups due to intervention, not outcomes per se due to some causal process.

The academic researcher can ignore external validity concerns while maximizing internal validity. The preoccupation with internal validity leads to causal inferences within the sample, inferring to populations is generally ignored. The experiment seems particularly well suited to researchers interested in causal inference, albeit constrained causal inference about outcomes in the samples.

The decision maker, in contrast, has a different set of priorities. The decision maker has a large stake in the outcomes of research. The research hypothesis is predetermined by the goals of the program being evaluated. Acceptance of the null hypothesis cannot be taken lightly; it may signal the end of the program. The significance of accepting the null hypothesis in evaluation has led Suchman to argue that evaluation must focus on the process leading from program activities to outcomes so that the reason for program ineffectiveness can be determined (see below).

It is also the case that the decision maker utilizes evaluation data to make a decision about a program's likely performance in the future. That is, the decision to continue a program, expand it, or close it down is based on a forecast into the blind future, thus the decision maker is primarily interested in utilizing sample results to generalize to other situations. Consequently, external validity must be a preeminent concern in evaluation. The need of the decision maker to focus on process rather than outcomes and external as well as internal validity suggests that the experiment may not be the most appropriate research design. If this is true, several interesting issues suggest themselves.

The first is, if in evaluation research an experimental design is not utilized, can the evaluator make causal inferences? The answer is, the evaluator may or may not but the issue is basically irrelevant. Most social science disciplines, with the exception of psychology and educational psychology, don't require causal assertions.

Kerlinger points out that scientific research can be done without invoking cause and causal explanation.

> Evidence can be brought to bear on the empirical validity of conditional statements of the "If p, then q" kind, alternative hypotheses can be tested, and probabalistic statements can be made about p and q--and other p's and q's and conditions r, s, and t . . . the elements of deductive logic in relation to conditional statements, a probabilistic framework and method of work and inference, and the testing of alternative hypotheses are sufficient aids to scientific ex post facto work without the excess baggage of causal notions and methods presumably geared to strengthening causal inferences (p. 393).

Indeed, behavioral models in economics and political science rely solely on the notion of conditional relationships, that is, differing frequency distributions of the dependent variable given different levels of the independent, as the sole basis for establishing a relationship between variables.

A second issue related to appropriate designs for evaluation concerns the role of experimentation in evaluation. If experimental design will not be useful for decision makers, why does the evaluation methodology literative argue for experimentation. The answer seems to be that evaluation methodologists fail to recognize that evaluation data need to be geared to decision makers and not the evaluator. Consequently, the literature discusses evaluation as if the results were to be used by the evaluator. In fact, the two research design pieces most widely cited in evaluation literature, Campbell and Stanley's Experimental and Quasi-Experimental Designs for Research (1963) and Cook and Campbell's Quasi-Experimentation (1979) are not books about evaluation research. They are merely books on research design. The fact they are so heavily cited by evaluators indicates an insensitivity to the special characteristic of evaluation as part of administrative decision making. It is possibly the failure of evaluation to focus on the needs of the decision maker rather than the evaluator that accounts for the low usage rate of evaluation data in decision making (see, for example, Carol Weiss, Evaluating Action Programs, pp. 318-326).

Finally, Suchman's critique of experimentation is based on the lack of concern with process, that is, the causal process leading to outcomes. Suchman argues that the ultimate goals of social action programs are generally only indirectly affected by program activities. Programs attempt to change the intermediate process which is "causally" related to the ultimate objectives. Thus, there are two possible sources of failure (of a program to achieve its goals) (1) the inability of the program to influence the "causal variable" or (2) the invalidity of the theory linking the causal variable to the desired objective (Caro: 46).

Experimental design, with its exclusive focus on outcomes, would not be able to distinguish the first from the second type of failure. Suchman's response to this is the concept of distinguishing "program failure" from "theory failure." The concept is derived from a particular interpretation of the structure of social action programs.

The structure of a social action program can be represented by a series of sequential goals. Goals can be classified as immediate, intervening or ultimate based on their temporal proximity to the initiation of a program, the degree to which attainment of the goal is a direct function of program activities, and the extent of change implied by attainment of the goal.

Suchman discusses the example of a tuberculosis program where the ultimate goal is stopping the spread of infectious tuberculosis. The immediate goals are "Provision of appropriate x-ray facilities for general hospitals and the encouragement of the use of existing facilities for the x-raying of all adult admissions" (p. 70).

The intervening goal would be, "Isolation by prompt hospitalization of all infectious cases until rendered noninfectious" (p. 69). The long-range goal would be, "The earliest possible detection and isolation of all cases of reinfection tuberculosis" (p. 68) such that the ultimate goal of arresting the disease could be attained.

The goals are nested by virtue of the relationship between adjacent goals. Each higher order goal is dependent upon attainment of the preceding goal, therefore each lower order goal is a causal determinent of the subsequent goal. Thus, the series of goals constitutes not only the sequence of program intentions but is also a representation of the causal relationships which link program activities to outcomes. The causal relationships constitute the theory upon which the program is based. For the tuberculosis program, x-raying is assumed to be a valid and reliable method for identifying tubercular victims, thus the use of x-ray machines in hospitals would lead to prompt hospitalization and therefore isolation of infectious cases. It is further assumed that hospital treatment of tuberculosis is more effective and efficient than home treatments, prompt hospitalization would lead to the earliest possible detection and isolation of all cases of reinfection tuberculosis. Note that the disease could be arrested only if all hypothesized relationships hold. The concept is one of a ". . . cumulative chain of objectives progressing from the most immediate practical objective (installing x-ray machines) toward the ultimate ideal goal (arresting the disease) . . ." (Suchman: 54).

The notion of a linkage of program activities to ultimate goals via intervening variables has important implications for the assessment of social action programs. In particular, the process becomes the most important focus. If the evaluator is to demonstrate that treatment outcomes are due to the planned intervention of a program, and not chance or some unanticipated reason, the relationships between program effects and ultimate outcomes must be established. Thus, to infer that a program is achieving its ultimate goals, it must be established that (1) the program attained the intervening goals, and (2) the intervening goals are causally related to the ultimate goals.

Conversely, the failure of a program to realize its ultimate goals may be a result of two factors: (1) Program failure, where the program failed to attain the intervening goals, or (2) Theory failure, where the intervening goals were not causally related to the ultimate goals. The distinction is important in terms of administrative decisions concerning the future of a program and/or the feasibility of ultimate program goals. Assume that a decision maker can modify both program activities and program goals (policy). Unlike the success or failure outcomes afforded by experimentation and quasi experimentation, the use of a methodology assessing both program failure and theory failure would result in a family of evaluation outcomes. These outcomes, and appropriate administrative responses, can be summarized as follows:

		program failure	program success
theory	failure	A	B
	success	C	D

A. Modification of program activities would not result in the attainment of program goals. Both activities and policy would have to be modified.

B. The program attained the intervening goals but the ultimate goals are infeasible. This requires a modification of program policy.

C. The program failed to achieve the intervening goals but the ultimate goals are feasible. This requires modifying the program activities to attain the ultimate goals.

D. The program achieved the ultimate goals via the intervening goals. Such a program could be continued and expanded.

Finally, it may be the case that some of the intervening goals are attained and are related to some of the ultimate goals. Such a situation would constitute partial program and/or theory failure, indicating the need for partial program and/or policy modification. This would enable decision makers to enact incremental decisions concerning components of a program while maintaining the overall program structure.

The Ohio-Westinghouse evaluation of Head Start provides an example of the program failure/theory failure distinction. Is the finding that the program did not induce lasting achievement gains in the treatment group a function of theory failure or program failure? It has been argued that the nonsystematic manner in which the program was implemented hindered its effectiveness in stimulating children (Smith and Bissel, 1969). Jenson, however, argues that the greater part of variation in I.Q. is accounted for by heritability, therefore, any attempt to increase achievement via a stimulating environment can result in only limited success (HER, 1969: 1-123). The distinction is of administrative importance. If the Ohio-Westinghouse results are a function of program failure, i.e., the environmental theory is essentially correct, modification and reorganization of Head Start may result in more positive outcomes. However, if Jensen's position is valid, Head Start, in any form, will never be effective. It is the program goals which would have to be modified. Unless the process were examined, however, the issue would not be considered or even identified.

Conversely, results that favor the treatment group on outcome do not necessarily indicate that a program is effective. In discussing the importance of examining the linkage between program activities and outcomes, Suchman has argued that

> One of the most significant implications of this approach to the statement of evaluative hypothesis involves the challenge

> not only to demonstrate that effect B follows program A, but also to "prove" that effect B was really due to program A. Some administrators may argue that so long as B occurs it does not really matter whether A was the actual cause of B. However, if A is spurious, one may institute an expensive, broad program based on A only to find (or, even worse, not to find because the valuation is not continuous) that the desired effect no longer occurs because of a change in the "true" cause which may have been only momentarily related to A (Suchman: 87-87).

The failure of the experimental design to account for the relationship between program activities and program goals leads to the situation that neither positive nor negative results on outcome measures indicate what the effect of the program has been. A significant effect may be due to treatment, however, unless the treatment is examined, not just assumed, the actual treatment may have been quite different from what was intended.

Another aspect of experimental design criticized by Suchman is that inferences concern group differences rather than variable levels. However, the ratio of difference over the total score may be an important statistic. That ratio would reflect the impact of the treatment on outcomes relative to other influences on the outcomes. But, the experiment does not offer any explanation for the levels of outcomes not due to treatment.

In basic research, where the variables are placed into relationships by theoretical argument, recognition of a multiplicity of causes (of which treatment is but one) and effects is manifested by the use of multicausal models in which no effect has a single cause and each cause had multiple effects. Suchman argues that the logical conditions of a "multiplicity of causes" and an "interdependence of events" applies in evaluation research. Since social action programs are only disposing, contributory or precipitating rather than determining causes of outcomes, he observes that

> . . . any "explanation" of the success or failure of program A to achieve effect B must take into account the preconditions under which the program is initiated, the events which intervene between the time the program begins and the effects are produced, and the consequences that follow upon the effects. <u>Thus, no program is an entity unto itself, but must be viewed as part of an ongoing social system</u> (emphasis added) (Caro: 40).

From this perspective, a social action program intervenes into a social process with the objective of manipulating certain variables in that process. Changes in the dependent variables (the ultimate program goals) are then a function of the relationship between the intervening (the direct program goals) and dependent variables. Suchman calls this an "input-process-output" model of social change where the inputs, treatment effects, are translated into outputs, the program goals, via the social process. The treatment/outcome relationship is nested in a larger set of causal relationships.

Underlying a social action program is what Suchman calls the evaluative hypothesis. The hypothesis that "Activities A, B, C will achieve objectives X, Y, Z implies some logical reason for believing that the program activities have some causal connection to the desired effect" (Caro: 45). The evaluative hypothesis contains the set of causal relationships which lead from program activities to outcomes. It is the theoretical argument which justifies accounting for outcomes in terms of the program and therefore must be logical, plausible, and testable. Evaluation is a test of the valuative hypothesis. It tests the validity of the argument that the development of the desired effect can be explained in terms of program variables.

The evaluative hypothesis is a theoretical statement which is testable. Consequently, evaluation research is conducted as basic research, since it involves the testing of a theoretical model. That model comes from the causal reasoning underlying program activity. Thus, evaluation research is basic research, and is carried out by the rules established for any theory testing research paradigm. In espousing a view of the unity of research, Suchman states:

> The scientific method is not bound by either subject matter or objective. Hence, evaluation research has no special methodology of its own. As "research" it adheres to the basic logic and rules of scientific method as closely as possible. . . . In other words, evaluative research is still research, and it differs from nonevaluative research more in objective or purpose than in design or execution (Suchman: 81-82).

Thus, the ground rules are established. Evaluation research must be conducted by the rules and criteria that govern all research. Along with a concern for assessing the impact of a program, an evaluation researcher must contend with methodological issues such as problem formulation, concept formulation, hypothesis and theory testing, and inference generation. In Evaluative Research, Suchman states, "Research begins with a hypothesis. . . . In the Evaluation, that hypothesis is a statement of a causal relationship between some program activity and some desired effect." Even when the hypothesis is not explicitly specified by the evaluator, it is implicitly present.

Suchman notes that, ". . . (the hypothesis) is often overlooked by the evaluative researcher who may tend to forget that a test of 'Does it work?' presupposes some theory as to why one might expect it to work" (Suchman: 86).

A research design compatible with this view of evaluation must allow for a theoretical model of the relationship between program and outcome. The testing of this model is the essential evaluation task and at the center of the proposed research design. However, this theory based design is not the only evaluation consideration.

One point that requires emphasis is that the theory testing design should not be viewed as a replacement for the experimental paradigm but rather as an expansion of experimentation to include the assessment of intervening variables and goals and to examine the relationship between goals of different levels. The basic operation of experimentation, the comparison of groups on particular attributes is to be maintained. However, the position adopted here is that comparing groups on outcome measures will not provide sufficient information about a program such that inferences can be made. The comparisons must include assessments of intervening as well as outcome variables and the relationships beween variables. One implication of this is that, for the most part, the assumptions, conditions, and requirements for valid and reliable inferences in experimentation apply as well to the theory testing approach. For example, the desirability of randomized assignment to groups applies here. However, it will be argued that the negative effects of nonrandomized assignment may be mediated by the proposed design.

The second chapter in this dissertation examines Suchman's arguments about the conduct of evaluation research in more detail. The theory testing design proposed at the end of Chapter II is then based on Suchman's position and some of the arguments in this chapter. Chaters III and IV will then provide a test of the theory based research design.

CHAPTER II

A META-THEORY AND METHODOLOGY FOR EVALUATION RESEARCH

The task in the second chapter of this dissertation is to present the major arguments in Suchman's meta-theory of evaluation and develop a methodology consistent with these arguments. While the focus in the preceding chapter was on the distinction between program failure and theory failure, Suchman, in fact, develops a global argument concerning the conduct of evaluation, touching on a series of related theoretical and methodological issues. It should be kept in mind that Suchman's concern is for the development of an approach to the assessment of social action programs which incorporates two evaluation foci:

1. Assessment of the evaluative hypothesis. That is, examination of the linkages between program activities and outcomes since, by assumption, the impact of treatment on the dependent variables travels along a causal path filtering through a set of intervening variables.

2. Assessment of the larger social environment within which the program operates. Particular attention must be paid to nontreatment factors related to outcomes since, by assumption, the program constitutes only one of a series of causal influences on the dependent variables.

Suchman introduces his meta-theory by noting the long-standing distinction between basic and applied research. His intention is to demonstrate that a necessary condition for sound evaluation is the inclusion of basic research principles (Caro: 51). Upon reflection, it would seem that highlighting evaluation by distinguishing basic and applied research is a questionable strategy given the lack of intersubjectively agreed upon definitions of the two research types. Consider the following example.

The essential argument concerning this distinction is that basic and applied research differ by the purpose for which each is intended. Further, the different purposes are better served by different methodologies such that there is one way to do basic research (theory testing an empirical description) and another way to do applied research (impact studies, intervention analysis). Cook and Campbell (1979) use the argument of purpose specific research methods as the foundation for their contention that different priority orderings of validity concerns exist for basic and applied research.

Cook and Campbell elaborate upon Campbell and Stanley's (1963) discussion of internal and external validity by introducing two additional categories of validity; construct validity and statistical conclusion validity. Construct validity is the issue of whether a construct created by a researcher truly and only measures the concept it represents (pp. 38-39). Statistical conclusion validity concerns inferences about covariation between variables on the basis of statistical evidence (p. 37).

One implication of enumerating four validity types is that Cook and Campbell perceive the researcher striving for causal inference being confronted by four questions:

1. If a statistical relationship exists between variables, will it be detected (statistical conclusion validity)
2. Can the outcomes that emerge in a sample be attributed to the independent variable (internal validity)
3. Can the outcomes be generalized to other samples, settings or times (external validity)
4. Inferences are made in terms of concepts; can it be demonstrated that the constructs employed truly reflect the concepts of interest (construct validity)

According to Cook and Campbell (pp. 82-83), the different purposes for which basic and applied research are intended lead to different rankings of concern for these validity types. For basic research, that is, for investigators with theoretical interests, the types of validity, in order of importance, are internal, construct, statistical conclusion and external validity (p. 83), while the priority ordering for many applied researchers is something like internal validity, external validity, construct validity of the effect, statistical conclusion validity and construct validity of the cause (p. 83).

These rankings indicate, as clearly as a position paper, exactly what Cook and Campbell see as the important methodological differences between basic and applied research. Ignoring for a moment the top priority accorded internal validity in both cases (discussed below), the major difference between basic and applied research concerns the ranking of external validity, that is, the importance attached to the generalizability of outcomes. They argue that since few theories specify crucial target settings, populations, or times to or across which generalization is desired, external validity is of relatively little importance to basic research (p. 83).

Applied research, on the other hand, is concerned with testing whether a particular problem has been alleviated by a treatment. It is crucial that any demonstration of change be made in a context which permits either wide generalization or generalization to the specific target settings or persons in whom the researcher or his clients are particularly interested (p. 83); thus the high ranking for external validity.

Finally, the primacy of internal validity for both basic and applied research is because Cook and Campbell are writing about experimentation (p. 84). The unique purpose of experiments is to provide stronger tests of causal hypotheses than is permitted by other forms of research (p. 84). Given that the unique original purpose of experiments is cause related, internal validity has to assume a special importance in experimentation since it is concerned with how confident one can be that an observed relationship is causal (p. 84). It would appear that, for Cook and Campbell, overriding any distinctions between applied and basic research is the unifying assumption that all research is best conducted by experimental designs.

In light of Suchman's argument that evaluation research should be conducted as basic research, and given Cook and Campbell's priority ordering for basic research, the conclusion could be drawn that Suchman does not believe the evaluative hypothesis is generalizable. However, the conclusion is incorrect. Suchman perceives basic research as aiming at the formulation of theoretical generalizations while applied research stresses action in a highly specific situation (Suchman: 75). He further argues that the generalizability of basic research results is because theory testing focusses on the discovery of knowledge which is not context specific. The specificity of applied research is because such research is a specific application of knowledge in a given context (p. 75).

Both Suchman and Cook and Campbell adopt the position that generalizability is a high priority issue in evaluation. For Suchman, this derives from an argument that evaluation needs to be conducted as basic research while Cook and Campbell assert that external validity is crucial in evaluation precisely because it is applied research. Clearly, Cook and Campbell are not on the same wave length with Suchman. An explanation for these diametrically opposed definitions, as well as a means for characterizing methodological approaches to evaluation, may be possible provided the distinction between basic and applied research is replaced by a distinction touched upon in Chapter I: research aimed at causal attribution and research aimed at specifying conditional relationships.

This confusion concerning definitions can be accounted for by the infatuation with causal attribution exhibited in psychology and educational psychology. One effect of the concern for causal inference is that internal validity is always more important than external validity. Cook and Campbell's statement that internal validity is the sine qua non of causal inference (p. 84) readily illustrates this point. They also argue that increases in external validity must be at the expense of internal validity. Campbell and Stanley (1963) note that both internal and external validity are important but ". . . they are frequently at odds in that features increasing one may jeopardize the other (p. 5). (If internal validity is the sine qua non, where does that leave external validity?) Cook and Campbell suggest that often ". . . jeopardizing internal validity for the sake of increased external validity usually entails a minimal gain for a considerable loss" (p. 84). Observe that the greater the degree to which a researcher disrupts some social situation, the easier it will be to establish internal validity. External validity, however, is greatest when the researcher (1) uses unobtrusive measures, (2) creates a nonreactive setting, and (3) refrains from pretesting groups (Campbell and Stanley, 1963). Thus, the levels of internal and external validity are inversely related, and are a function of the degree to which a researcher actively intervenes in the process under study. For example, while the absence of pretesting increases the external validity of outcomes, it also reduces the researcher's assurance that pretreatment equivalence exists between groups, lowering the level of internal validity.

If increasing external validity decreases internal validity, it may reasonably be asked why external validity is so important in applied research settings. The answer, according to Cook and Campbell, is that in the applied setting a researcher can afford to be less concerned with precise causal inference. The researcher, ". . . is relatively less concerned with determining the causally efficacious components of a complex treatment package, for the major issue is whether the treatment as implemented caused the desired change" (p. 83). Since the main emphasis is on outcomes and not process, that is, "less concern with determining the causally efficacious components," the reduced causal emphasis allows for increased concern with external validity. This is not to suggest that internal validity is not paramount, after all Cook and Campbell are still advocating experimentation, but it is to suggest that the degree of concern allocated to internal validity is less in the applied than in the basic research setting. Thus, generalizability is most likely to occur when causal inference is not so rigorously pursued.

For Suchman, generalizability is not as much a function of the research setting as it is a function of the types of variables used for analysis and the level of abstraction of the variables (p. 75). This notion is attributed to Hovland (p. 77) who distinguishes between program and variable testing in evaluation research. Program evaluation refers to a test of a total product with the purely practical objective of determining whether exposure to the program was accompanied by certain desired effects (the argument advocated by Cook and Campbell). Variable testing, on the other hand, is concerned with singling out specific components of the program, as indices of some more generalizable stimuli, and testing the effectiveness of these variables. Contrary to the Cook and Campbell position, Suchman states, "Program testing has almost no generalizability, being applicable solely to the specific program being evaluated. Generalizations (to other products, populations, times) have the status of untested hypotheses. For Suchman, as opposed to Cook and Campbell, generalizability is much more a function of data rather than research design.

These differing notions of what leads to particular levels of generalizability can again be accounted for by differentiating the experimental and nonexperimental approaches to research. In particular, while Suchman discusses the generalizability of relationships between variables, Cook and Campbell's concern is with the generalizability of an intervention or treatment. This can be deduced from their discussion of what promotes or prevents external validity. Consider the test of some program where the treatment is administered and some measurement takes place. Outcomes are a function of the intervention, in this case treatment plus anything else in the research setting that affects outcomes, for example, if pretreatment testing took place, then the intervention plus pretesting are the treatment. If in actual operation pretesting does not occur in the program, the external validity issue is whether treatment plus pretesting is basically the same intervention as treatment by itself. Generalizability in this case is generalizabiltiy of treatment since unobtrusive measures, no pre-treatment testing and a nonreactive setting make the treatment in the experiment as much like treatment in the population as possible. Thus, external validity concerns the generalizability of the treatment. Crucially, it would seem that no provisions are made to assess the generalizability of the relationships that occur. Apparently, any significant relationships are assumed to be generalizable.

In contrast, Suchman sees generalizability as a function of timelessness and spacelessness (p. 78). That is, the generalizability of relationships is a function of the degree to which the results are independent of the situation in which they are studied. Basic research aims at discovering knowledge (about relationships) which holds true in any (or at least many) situations. To the extent that evaluative research can focus upon the general variables underlying a program and test the effects of these variables rather than the effectiveness of the program as a whole, it may hope to produce findings of greater general significance. For example (Suchman: 77):

> An evaluation of the effectiveness of a prenatal clinic may be set up on a program basis according to some administrative design and then determining the number of mothers who attend. Such an evaluation may enable one to decide whether or not to continue this specific clinic but it will have only limited value for planning similar clinics in different areas or for different populations. However, if the clinic is established to test some specific action principle or variable, for example, the relative effectiveness of personal versus formal appeals for attendance, the results would have greater transferability to other situations. In this sense one might argue for the greater ultimate "practicality" of variable as opposed to program testing because of its stronger potential for generalization and accumulated knowledge.

Suchman's emphasis on generalizable relationships rather than on causal inference would seem to mark him as a nonexperimentalist. He notes that the essential evaluation task is not describing the relationship between treatment and outcomes, but the elaboration of how and why the treatment was able to achieve the objectives. This task, he suggests, is at the heart of evaluative research (Caro: 50). The test of a program comes not from establishing covariation (or even causality) between program and outcome, but by the basic research procedure of specification through statistical elaboration of this zero order relationship. The evaluative hypothesis contains this statistical elaboration, that is, it contains variables which impinge upon the original relationship. The emphasis on explaining the program effects by the evaluative hypothesis leads Suchman to declare

> One of the most significant implications of this approach to the statement of evaluative hypotheses involves the challenge not only to demonstrate that effect B follows program A, but also to "prove" that effect B was really due to program A. Some administrators may argue that so long as B occurs, it does not matter whether A was the actual cause. This will be legitimate insofar as A is not a spurious cause of B. However, if A is spurious, one may institue an expensive, broad program based on A only to find (or, even worse, not to find because the evaluation is not continuous) that the desired effect no longer occurs because of a change in the "true" cause which may have been only momentarily related to A. <u>To achieve this test of "spuriousness," the evaluative project must include an analysis of the intervening process between programs and results</u> (Suchman: 87, emphasis added).

By advocating the control of nontreatment variables through statistical elaboration rather than randomization, Suchman implies that inferences generated by experimental design are deficient. In particular, control through randomization leads to research where analysis of process becomes difficult, if not impossible. Suchman distinguishes the "descriptive" part of an evaluation, where the zero order relationship is assessed, and can be assessed by experimental designs, and the "explanatory" part of evaluation, where the analysis of process establishes the causal connections between what was done and the results that were obtained. Thus, "making sense" of the descriptive analysis is the basic reason for adding a concern with process to the evaluation study (Suchman: 66). Consequently, the process becomes the major focus of evaluation.

It should be noted that Suchman's use of intervening variables to test for spuriousness is not, strictly speaking, a technically correct argument. For a relationship to be spurious, the control variable has to be antecedent, leading to both zero order variables. If the control variable is intervening, the relationship may be contingent, but the dependent variable could not occur in the absence of the independent variable. A complete elaboration of the bivariate relationship between treatment and outcome would require both antecedent and intervening variables. The effect of including control variables is that (Caro: 50):

> . . . any "explanation" of the success or failure of program A to achieve effect B must take into account the preconditions under which the program is initiated, the events which intervene between the time the program begins and the time the effects are produced, and the consequences that follow upon the effects. Thus no program is an entity unto itself but must be viewed as part of an ongoing social system.

One can imagine constructing a model by which a program is to be assessed that contains two basic types of relationships. The evaluative hypothesis would contain those intervening variables that, by design, contribute to the relationship between treatment and outcomes. The second set would be a set of relationships that constitute the context within which the evaluative hypothesis is located. Two types of variables will be found in the context, those which impinge upon the evaluative hypothesis, disrupting or strengthening the causal flow from treatment to outcome, and variables, independent of the program, which also influence outcomes. Suchman sees the construction and situation of the program within some social context as being a basic research activity. He states (Caro: 50-51)

> In social research we generally deal with multicausal models in which no event has a single cause and each event has multiple effects. No single factor is a necessary and sufficient cause of any other factor. These logical conditions of a "multiplicity of causes" and an "interdependence of events" applies equally to evaluative research. It means that activity A becomes only one of many possible actions or events which bring about (or deter) the desired effect. . . . The significance of this model of "causality" is that evaluations of success must be made in terms of conditional probabilities involving attacks upon causal factors which are only disposing, contributory, or precipitating rather than determining. The effect of any single activity will depend upon other circumstances also being present and will itself reflect a host of antecedent events. Any single activity will, in turn, have a great many effects, many of them unanticipated, and some of them even undesirable.

Suchman's position is rather clear as to how an evaluator should go about assessing a social action program. The relationship between treatment and outcomes is the same as a zero order hypothesis in basic research. Measures of association may indicate the strength of the zero order conditional relationship between treatment and outcome, but causal understanding comes about only when the relationship undergoes statistical elaboration, that is, when variables impinging upon the zero order relationship are controlled by explicitly incorporating them in the analysis. The implication is that evaluation takes place by examining a program within a contextual model representing that social process the program is intended to affect (one evaluates in context, not in a vacuum).

One problem with Suchman's argument that, "evaluation research should be conducted as basic research, in basic research one accounts for a zero order relationship by statistical elaboration, therefore in evaluation the treatment outcome relationship needs to be assessed in context," concerns his notion of what is basic research. This issue has already been raised. One could reasonably argue that Suchman's assumption that explanation is a function of elaboration, rather than randomization, may just be because he is a sociologist, and sociological research is generally nonexperimental. There seems to be no reason why control of variables is more efficaciously accomplished by elaboration rather than by randomization. In fact, Suchman's arguments about this issue are the foundation for this dissertation. The contention is not that control by elaboration is methodologically superior to control by randomization. With respect to internal validity, elaboration is weaker. Rather, the crucial argument is that experimental design is inappropriate in the evaluation setting because control by randomization precludes a study of the causal process, and it is inferences about the process which have the greatest use for decision makers.

Making the case for the utility of the study process in evaluation requires several arguments. Suchman's major justification for studying process is, of course, based on his distinction between program failure and theory failure. He notes that an analysis of process can have both administrative and scientific significance, particularly where the evaluation indicates that a program is not working as expected from an administrative viewpoint, program failure and theory failure should lead to different administrative responses. In general, the response to program failure would emphasize more systematic implementation of the program, while theory failure would indicate the need for a different program strategy. Obviously, program failure could not be distinguished from theory failure unless the evaluative hypothesis is explicated and tested. Additionally, examination of the process would minimize the probability of a Type II error when a positive treatment effect is detected.

A second justification for examining the process concerns the implication of focussing on the general concepts underlying a social action program. By constructing and testing a general model of the process influenced by a program, the results will contribute to the body of academic knowledge concerning a particular policy area. For example, an evaluation of Head Start using Suchman's approach would be accomplished by testing a model representing the achievement and motivation process for preschool children, an area of research important to psychologists, educational psychologists, and sociologists (and at least one political scientist). David Cohen (1975) has argued that the knowledge generated by basic research is important to policy makers because general policy orientations (as opposed to specific policy decisions) are predicated on the state of basic knowledge in a policy area. Suchman comments that his approach to evaluation, ". . . combines evaluation with research and attempts to make a contribution to basic knowledge as well as to administrative decision making" (Suchman: 68).

A third argument for the utility of studying process is more complex, requires more assumptions and, if valid, would be of greater significance than the other arguments. The argument concerns the type of data that has the greatest likelihood of utilization by decision makers. Since utilization is the major goal of evaluation (Weiss: 1972), the significance of this argument should be obvious.

The argument is predicated on the assumption that, most often, decision makers are constrained to act incrementally, where by "incremental" is meant decisions that result in only small changes in the condition of some situation. Many models of decision making (see, for example, Etzioni, Lindblom, Seidman, Wildausky or Allison) are based on the notion that decisions tend to be small in scope and effect. A further assumption of this argument is that information, to have an impact on incremental decisions, must be incremental in nature. That is, the level of information must be as specific and focused as the decision it is supposed to influence.

One implication of specifying the process by which a program affects outcomes, is that information about the program is reduced to the extent that data reflects components of a program rather than the program as a unitary entity. If it is more feasible for decision makers to make lower level program decisions, they would require data that says something about the subprogram level. Such data is available when the process is broken down into components parts. As an interesting aside, consider how this view of what usable evaluation data needs to be contrasted with the Cook and Campbell argument that in applied research the researcher is relatively less concerned with determining the causally efficacious components of a complex treatment package (p. 83).

Weiss, using an incremental model of decision making, argues that decision makers have greater use for data about ". . . which elements of the program worked or didn't work and how and why" rather than for global findings generated by outcome focussed research (p. 323). Weiss states that information about elements of a program can be obtained only with

1. The explication of the theoretical premises underlying the program, and the direction of the evaluation to analysis of these premises.

2. Specification of the "process model" of the program--the presumed sequence of linkages that lead from program input to outcome, and the tracking of the process through which results are supposed to be obtained.

3. Analysis of the effectiveness of components of the program, or alternative approaches, rather than all-or-nothing, go or no-go assessment of the total program (Weiss: 323).

Thus, the reduction of the program data into component parts may be more relevant to the decision maker constrained to behave incrementally than a macro-statement of program impact.

The final argument used to make the care for control by elaboration brings the discussion full circle since it concerns the decision makers' need for data that are externally valid and generalizable. It has been suggested earlier that for decision makers generating forecasts about program effectiveness into a blind future, generalizability of the evaluation results is of paramount concern. It has also been suggested that the generalizability of an experiment concerns the generalizability of treatment while research which is nonexperimental attempts to generalize relationships. It would seem that what decision makers would like to generalize is the effect of the program, that is, they need some idea of what treatment groups would look like after treatment. This type of generalizability more closely corresponds to generalizability of relationships rather than generalizability of treatment.

There is another, more philosophical, level at which the generalizability of sample data can be considered. In a sense, the generalizability of an inference depends upon the degree to which the inference can be proven true. If the conclusion in an inference is necessarily true, it should hold in all situations. The question of generalizability of inductive inferences raised by Hume (Salmon: 11) can be understood in terms of two distinctions about inferences. One distinction is fundamental; the distinction between demonstrative and nondemonstrative inferences. A demonstrative inference is one whose premises necessitate its conclusion; the conclusion cannot be false if the premises are true (Salmon: 8). A nondemonstrative inference simply fails to be demonstrative.

The second distinction between inferences is related to the first. Inferences can be ampliative or nonampliative depending on whether the conclusion is contained in the premises (nonampliative) or whether the conclusion exceeds the scope of the premises (Salmon: 8). Demonstrative inferences are nonampliative, their truth preserving nature comes by sacrificing any extension of content; the conclusion is totally contained in the premises. Yet the scientific method is based on predicting the future from the present by generating lawlike generalizations; thus science requires ampliative inferences.

Salmon summarizes Hume's position succinctly (p. 11):

> We cannot justify any kind of ampliative inference. If it could be justified deductively it would not be ampliative. It cannot be justified nondemonstratively because that would be viciously noncircular. It seems, then, that there is no way in which we can extend our knowledge to the unobserved. We have, to be sure, many beliefs about the unobserved, and in some of them we place great confidence. Nevertheless, they are without rational justification of any kind.

The responses to Hume's position take two directions. One could respond that, since generalizability cannot be guaranteed, inductive inferences should not be attempted. While this position seems rather extreme and unrealistic, it is, in fact, the reason for the emphasis in experimental design on internal validity. Cook and Campbell state that Campbell and Stanley (1963), in light of Hume's paradox, explicitly reject inductive inference (p. 86). The primacy of internal validity is because its problems are deductively soluble (Cook and Campbell: 86). In effect, experimentalists prefer to account for what happened within a sample, rather than generalize from a sample, because causal inferences concerning only the sample under analysis are demonstrative. This position is not immune to Hume's paradox since the preference for demonstrative inference is also a preference for nonampliative inference.

The problem remains that decision makers require ampliative inferences. The fact that in experimental research ampliative inferences are either foregone, or made without reference to Hume's problem of induction, means that generalizable inferences based on experimental design are likely to not be very useful to decision makers. The response to Hume should not be either foregoing ampliative inference or ignoring the paradox. If generalizability of the evaluation outcomes is necessary, one needs to acknowledge the paradox to understand what the problems of inductive inference are, and then make the best effort possible to render the results generalizable, even though it is known that the research design will not yield perfectly justified inferences. The two activities required to test an hypothesis or model is (1) specification of the variable in the model, and (2) a statistical test of the model with some sample. For the most part, the emphasis in empirical research has been on the second task.

However, the force of Hume's argument is to deny the justification of inferences, based on sample results, about some unobserved population.

This suggests that if inferences of greater generalizability are to be generated, emphasis would have to be moved from statistical analysis to the specification of the model. The generalizability of inferences would be more a function of the power of the argument in terms of logic and plausibility, which placed variables in theoretical relationship rather than the statistical results from the test of the model in a sample. To the degree that, in evaluation, the generalizability of the results would be a function of the logic, plausibility and detail with which the relationship between program and outcomes is specified, the inferences generated under statistical elaboration will be superior to those of experimental design. This is because the effect of statistical elaboration is to examine the bivariate relationship under as many conditions as possible, by incorporating control variables that may impinge on the zero order relationship. On the other hand, in experimental design the zero order relationship is examined in a conditionless state, where all other variables are held constant. Thus, the experimental setting cannot be expected to approximate the outside world at all, making generalizations difficult from the perspective of logic and plausibility.

These four arguments attempt to stress the reasons why evaluation is more appropriately undertaken by modelling programs rather than assessing them experimentally.

It has been argued that Suchman's position concerning statistical elaboration, the study of process, and the distinction between program failure and theory failure are reasonable. However, Suchman fails to specify a research design which would incorporate these concerns. Suchman does discuss some research designs, but they only serve to weaken his arguments. Problems arise when Suchman makes the case for a particular type of analysis and then suggests designs that are inappropriate.

On the one hand, Suchman has argued that the heart of evaluative research is to elaborate upon how and why program activities were able to achieve objectives (Caro: 50). Because evaluative research needs to be conducted as basic research, crucial significance is attached to an analysis of the process whereby "a" related to "b," that is, intervening variable analysis (Suchman: 79). Thus, the fundamental methodological task in evaluation is statistical elaboration.

On the other hand, the designs for evaluation research that Suchman discusses are experimental designs straight out of Campbell and Stanley (Suchman: 91-111). The major part of his discussion on methods of evaluation simply reiterates standard experimental concerns. As for intervening variable analysis (the heart of evaluative research) Suchman deals with this in one sentence, "We cannot here go into the rather technical details of intervening variable analysis" (Suchman: 109). The implication is that intervening variable analysis is a back up method to experimental design, and may or may not be applied.

This dissertation takes a stronger position than Suchman and argues that the outcomes of evaluation are meaningful only when the intervening process is accounted for. The analysis of outcomes and process is a unitary research aim. Therefore, the research design suggested in this dissertation calls for the simultaneous analysis of process and outcomes.

In order to develop a methodology consistent with Suchman's concerns, the proposed research design combines the experimental concept of comparing a treated and nontreated group and the nonexperimental concept of control by elaboration, that is, assessing a program in context. The design rests on some fairly simple assumptions. It is assumed that the need for some social action program is the result of identifying some social problem. The social problem is some negative outcome of some social process. This implies that the aim of a social action program, to change the social condition of some treatment group, will result in some changes in the social process for the treatment group. Thus, the degree to which a social action program affects a treatment group will be indicated by the degree to which the social process differs between a treatment and control group.

The research design calls for comparisons between a treatment and nontreatment group. However, the comparisons to be made will not be solely in terms of outcomes but rather in terms of the social process the social action program is supposed to affect. This suggests that the initial evaluation task is the construction of a theoretical model representing the social process of interest. The model consists of the evaluative hypothesis (treatment components, intervening variables, and outcomes) located within a relevant context (variables not affected by the program which nonetheless impact on outcomes).

The theoretical model is fit both to a treatment and non-treatment sample. The basic utility of this research design lies in the different types of comparisons a researcher can make to assess different aspects of the program being studied. The first consideration must be with the power of the model to represent the social process of interest. The most appropriate way to go about this would seem to be by assessing the explanatory power of the model in the control group. It is for the untreated group that the ability of the model to account for some "naturally" occurring social condition can be most clearly documented. The task of model validation is most important because an unavoidable characteristic of this design is that assessment of program effects can be no better than the model used to represent the process. If the control group model shows poor fit, deviations in the treatment group from this baseline are meaningless. Thus, the first requirement is a well-specified, valid theoretical approximation of the relevant social process. Ideally, the researcher would start out with two or more candidate models and utilize that model with the greatest explanatory power.

If the model is acceptable, then the effects of the program can be determined by discovering those points at which the control group and treatment group models differ. An appropriate assessment to be made at this point concerns the program failure and theory failure distinctions. If the treatment group does not exceed the control group in levels of the outcome variables, the reasons for program ineffectiveness can be ascertained. Program failure would have occurred if the treatment failed to activate the intervening variables. This would be the case if, upon comparison of the levels of the intervening variables between groups, no differences in favor of the treatment group are detected.

Theory failure would occur if the relationships specified in the evaluative hypothesis failed to hold. The two junctures where theory failure can be assessed are the point connecting components of treatment with the intervening variables, and the point connecting the intervening variables to outcomes. Theory failure would primarily be assessed in terms of the relationships between variables in the treatment group.

The effects of statistical elaboration also would be assessed in the treatment group. Two types of relationships need to be examined. The first concerns those variables in the treatment group model which impinge, either in a positive or negative way, on the relationship between treatment components and outcomes. The second type of assessment is a comparison between treatment and nontreatment variable effects on outcomes. Such an assessment would indicate the degree of change a program could induce relative to nontreatment variables affecting outcomes. This research design, calling for the construction and testing of alternative theoretical representations of a social process of interest, and the types of between group and within group assessments that are possible, should lead to the sorts of inferences, useful to decision makers, that Suchman argues are based in an appropriate model for the conduct of evaluation.

One point that requires emphasis is that the proposed research design does not totally reject the experimental method. Both the assessment of outcome differences and program failure are based on comparisons between groups. Thus, the conditions under which experimental research is optimal, such as random assignment, apply with equal force to this design. However, it will be suggested that the negative implications of nonrandom assignment are not as severe for the proposed research design compared to quasi-experiments.

To this point, the dissertation has compared Suchman's arguments with those concerning experimentation. This was to set off characteristics of Suchman's approach with the best possible case for experimental design. However, most evaluation occurs in situations where random assignment to groups is not feasible. The significance of the proposed research design may be greater when comparing it to quasi-experiments under nonrandom assignment than when it is compared to true experiments when random assignments is possible. The trouble with quasi-experimentation is that not only is analysis of the process precluded, but the researcher does not even have the grounds for making causal inferences; the primary reason for doing experiments.

The basic problem in the quasi-experimental setting is that pretreatment equivalence between groups cannot be fully determined, therefore it is never clear whether post-treatment differences are due to treatment or some other differentially distributed variable. Aiken (1980) suggests that one problem of inference with quasi-experiments occurs when nonrandom assignment to group is a function of a variable that is also related to outcomes. When this selection variable is not included in the outcome assessment, the resulting estimate of program effect is biased and inconsistent. While the design proposed in this dissertation will not correct this problem, the fact that the theoretical model includes variables related to outcome implies that those variables can be examined for a relationship with selection. To the extent that significant differences between groups on these variables can be found, sources of the bias and inconsistency in the estimate of treatment effect can be identified, since these variables are related to selection and outcome, yet excluded from the outcome equation. Thus, the proposed design permits a diagnostic though not corrective strategy for dealing with this problem. At the least, a researcher could identify when the situation is occurring, and accordingly reduce the degree of belief in the results.

The next two chapters in this dissertation present the assessment of the strengths and weaknesses of the proposed design. The data from the Ohio-Westinghouse evaluation of Head Start are reanalyzed according to the proposed design. Chapter III contains the specification of the process model which is to be tested and Chapter IV reports the results of the analysis.

CHAPTER III

AN APPLICATION OF THE PROPOSED RESEARCH DESIGN TO AN EMPIRICAL EXAMPLE: THE SPECIFICATION

In this and the following chapter, the proposed research design, calling for the analysis of a social process model, is utilized to reanalyze the data from the Ohio Westinghouse evaluation of Head Start. There are two major goals of this reanalysis. One is illustrating the data results generated by the proposed design. Equally important is the goal of generating substantive inferences about Head Start's effectiveness as indicated by the data.

The initial task is the specification of a theoretical representation of the social process targettted by the Head Start program. The model consists of two parts: a statement of the evaluative hypothesis, that is, the set of relationships by which program activities lead to outcomes, and the context within which the evaluative hypothesis is located. The evaluative hypothesis can be deduced from the sequence of program goals specified for the program, under the assumption that each lower level goal is a necessary causal condition for attainment of the higher level one. The sequence of immediate, intervening, and ultimate goals can be stated in the form of a set of hypotheses.

Because Head Start was a War on Poverty program, its most long-range goal, in conjunction with other War on Poverty programs, was to contribute to a reduction in the economic and social attainment disparities among societal groups by increasing the level of economic and social attainment for the disadvantaged.

Thus,

$$X \longrightarrow Y_n \tag{1}$$

where X represents Head Start and

Y_n is increased economic and social attainment.

Because a large part of Head Start was devoted to educational programs, it must have been assumed that economic and social attainment is influenced by education.

In particular

$$X \longrightarrow Y_{n-1} \longrightarrow Y_n \tag{2}$$

where Y_{n-1} represents increased educational achievement and attainment on the part of disadvantaged groups.

However, Head Start was a pre-school program, so it must have been hypothesized that a student's academic success is influenced by the student's entering capabilities. Therefore,

$$X \longrightarrow Y_{n-2} \longrightarrow Y_{n-1} \longrightarrow Y_n \tag{3}$$

where y_{n-2} represents increased achievement potential and motivation upon entering school.

Since Head Start was a pre-school program, students entering school were no longer within the program environment or directly receiving treatment. By the definitions in Chapter I, Y_{n-2} is, therefore, an ultimate program goal. Consequently, any lower level goals are intervening or immediate by definition. The intervening goals can be attained as a direct result of program activities. At this point, such intervening goals are causal determinants of achievement potential and motivation. There exists great intersubjective agreement on the part of program designers, evaluators and other scholars that the causal mechanism Head Start adopted to induce higher level of achievement potential and motivation was an enriched environment.

Datta, for example, observed that one of the inspirations for the creation of Head Start was "an accumulation of evidence . . . showing that environmental factors in the early childhood years are particularly powerful in shaping children's future growth and development" (p. 5). Jenson similarly asserts that the major underpinning of compensatory education is the "'deprivation hypothesis,' according to which academic lag is merely the result of social, economic and education deprivation . . ." (p. 2). And Miller notes that Head Start curriculum planners developed a program structure particularly shaped to provide an environment that stressed experiences assumed to be lacking in the lower class home and neighborhood (p. 216). It is clear that the goal sequence becomes

$$X \longrightarrow I \longrightarrow y_{n-2} \longrightarrow y_{n-1} \longrightarrow y_n \quad (4)$$

where I is the intervening variable "enriched environment."

If P is defined as the set of immediate goals such as designing the program, identification, and collection of a treatment group and program implementation, a version of the full set of Head Start goals is

$$X \longrightarrow P \longrightarrow I \longrightarrow y_{n-2} \longrightarrow y_{n-1} \longrightarrow y_n \quad (5)$$

Verbally, the hypothesis states that given a Head Start program with a treatment group, program activities will lead to an enriched environment for the group. The enrichment will result in increased achievement potential and motivation for the group upon entering school. The enhanced capabilities will result in greater achievement and attainment. Increased achievment and attainment will lead to increased economic and social attainment such that inequalities among social groups will be reduced. (Clearly, all the hypotheses include a <u>ceteris paribus</u> assumption.)

The Ohio-Westinghouse evaluation was conducted as if equalized achievment potential and achievement motivation among children entering school was the ultimate goal. Thus, the reanalysis will focus on the sequence

$$X \longrightarrow P \longrightarrow I \longrightarrow Y_{n-2} \quad (6)$$

This is the underlying foundation for the structure of the evaluative hypothesis. The evaluative hypothesis will need to contain a persuasive argument as to why the Head Start program (X), when a treatment group has been identified and treatment specified (P), will lead to an enriched environment for the treatment group (I), resulting in improved achievement potential and motivation on the part of program participants (Y_{n-2}). The appropriate evaluative hypothesis and specification of content are dependent upon the ultimate goals ascribed to the program.

The issue of accurately determining the appropriate program goals has been extensively discussed (Weiss, 1972). The determination is fundamental since the assessment of program effectiveness is based on the degree to which the program attains the ascribed goals. To the extent that accurate goal determination is difficult with conventional research designs, it is also difficult with the proposed design. This is because nothing in the proposed design inherently allows for more accurate goal determination. The proposed design is intended to deal with problems of inference not problems of goal identification. In this case, the conventional goal identification procedure of utilizing program documents will be followed.

The Cooke Committee (1965), charged with framing the form and objectives of the Head Start program detailed a threefold approach to the development of Head Start services.

1. Provision of comprehensive services with particular attention to health and nutrition

2. Emphasis on the importance of strengthening family life and the ability of the parents to be primary advocates, change agents, and educators for their children

3. Focus on the child's motivational and social development and on the achievment of competence in everyday life, including academic preparation for school (Datta: 5)

With respect to the third concern, the Cooke memorandum specified two major objectives:

1. Improving the child's mental processes and skills with particular attention to conceptual and verbal skills

2. Establishing patterns of success for the child that will create a climate of confidence for his future learning efforts (Datta: 5)

The implication of cognitive (conceptual and verbal skills) and affective (confidence) goals of Head Start is that the evaluative hypothesis, and ultimately the social process model, will have to contain cognitive and affective input processes.

According to Datta (p. 6), the process by which the Head Start program was to effect change in cognitive ability and motivation was inspired by ". . . an accumulation of theory and evidence that environmental factors in the early childhood years are particularly powerful forces in shaping children's future growth and development." An additional focus of Head Start was on the effect of the parent/child relationship on the child's preparation for school. For the evaluative hypothesis to reflect these aspects of the program the cognitive and affective processes must originate in the child's pre-school (home) environment and, for the most part, center on parent/child interactions in the home.

Finally, the Head Start program was explicitly intended for economically disadvantaged families and pre-school children. It must have been assumed that a family's economic status led to a particular home environment and preparation of children for school. To incorporate this assumption, the evaluative hypothesis assumes socio-economic status of the family to be a determinant of the home environment.

Given this general framework the evaluative hypothesis will assume the following form

$$\begin{tikzcd} X_1 \arrow[r, dash] & X_2 \arrow[r, dash] & y_1 \arrow[d, leftrightarrow] \arrow[dl, dotted, leftrightarrow] \\ X_1 \arrow[r] & X_3 \arrow[r] & y_2 \end{tikzcd} \qquad (7)$$

where X_1 = economic status

X_2 = cognitive aspects of home environment

X_3 = affective aspects of the home environment

y_1 = achievement potential

y_2 = achievment motivation

and the dotted double headed arrows indicate potential relationships between the cognitive and affective processes.

While Figure 7 represents an explication of the causal dynamics of the evaluative hypothesis, the social process model will not be fully specified until the social context surrounding the evaluative hypothesis is explicated. According to Suchman, at this point an evaluator needs to draw on, and can ultimately contribute to, the state of knowledge in a particular social science or public policy area. Completing the social process model such that an evaluation of Head Start can procede requires the utilization of prior research about the achievement and motivation process in young children. Two sets of literature, focussing on the cognitive and affective components of achievement and motivation, will be reviewed to provide additional structure for the social process model.

Although cognitive models of learning have been studied across several social science areas, the findings tend to converge to a single general assertion: the degree to which a child is cognitively prepared for school is a function of pre-school interactions between the child and parents.

Iverson and Walberg (1979: 2) state that from a theoretical perspective, four approaches to the measurement and study of home environment and learning may be distinguished:

1. Sociological surveys that include socio-economic measures such as parent education, income, and occupation
2. Family constellation studies that analyze the number, birth order, and spacing of children in the family

3. The work of the "British School" that emphasized parental experiences and aspirations for the child and objects and material conditions in the home
4. The work of the "Chicago school" that emphasizes specific social-psychological or behavioral processes thought conducive to learning

Examination of samples of each research type, however, suggest that the differences are not theoretical but methodological.

The two factors which Duncan (1963) found constituted valid indicators of socio-economic status were occupational and educational attainment of the parents. Of the two, educational attainment was deemed to be of greater significance. Most subsequent sociological research, for example, Sewall, et al. (1970), Hauser (1971), and Duncan, Featherman and Duncan (1972) all utilized parental economic and educational attainment as determinants of learning. It is clear that in and of themselves, income and education of the parents do not lead to characteristics of the child and, therefore, the variables only serve as indicators of the level of the child's pre-school environment. The income level indicates the availability of material resources, for example, books and games, travel, etc., which a child can avail himself of in preparation for school. It is also assumed that the amount of time parents have to interact with children is a function of income. Educational attainment, it would seem, indicates something about the parents' valuation of schooling and it is assumed that part of the parent/child interaction consists of the parents relaying to and instilling in the child their (the parents') attitude toward education.

An example of an early family constellation study is Beverly Duncan's (1966) where she hypothesized relationships between achievement and, along with socioeconomic status, the number of siblings and whether the family is intact or broken. Anastasi (1956) reviewed 110 studies of number of siblings and achievement and generally found negative correlations between family size and I.Q. (Cicirelli: 1979, p. 366).

According to Victor Cicirelli (p. 366), the question of the effect of birth order on ability and achievement has been motivated both by the psychoanalytic conception of the unusual role of the first born and by observation of the over-representation of the first born among the eminent (Schachter, 1963). Although it is generally concluded that achievement is negatively related to number of siblings and order of birth, it is not clear how much of the relationship is due to the amount of interaction between parent and each child, the intended underlying concept, and the spurious relationship possible due to the negative relationship between SES and family size and the positive relationship between SES and achievement. It is clear, however, like educational attainment and income, the constellation studies are based on the notion that parent/child interactions, which necessarily decrease per child as the family gets larger, is the primary determinant of early school achievement.

The difference between the research of the British and Chicago schools of research on the home environment concerns the issue of what are appropriate indicators of parent/child interactions. Dave (1963) and Wolff (1964), at the University of Chicago, developed lists of parents' behaviors and parent/child interactive behavior that seem likely to foster intellectual growth. These process variables are measured by trained home interviewers asking questions such as "Do you read to your child?" and "Do you discuss his grades with him?" (Iverson and Walberg: 3). Sets of process variables are aggregated to indicate "presses" in the home environments. Examples of such presses include academic guidance, achievement, intellectuality of the home, and work habits of the family all of which are assumed to influence academic achievement. Other processes investigated by the Chicago school have focussed on activeness of the family (Dolan, 1978) and language models (Majoribanks, 1972, and Kifer, 1975).

In contrast, studies in the British school tradition (Fraser, 1959; Peaker, 1967; Wiseman, 1976; Majoribanks, 1976; Schaffer, 1976) focus on parents' experiences, attitudes and material conditions in the home rather than on the parent/child interaction patterns. Typical questions from the "Survey of Parents of Primary School Children" (Fouden, et al., 1967) include "What do you feel about the ways teachers control the children of (name of school)?" and "Has the teacher talked to you about the methods used at (name of school)?" (Iverson and Walberg: 3). Fraser (1959) used reading habits of the parents as a home environment measure while Claeys and DeBoerke (1976) and Schafer (1976) used the Parent Attitude Research Instrument developed by Schafer (1958) (Iverson and Walberg: 6).

At issue, still, is what constitutes a reliable and valid indicator of parent/child preschool interactions. Iversen and Walberg (1979) suggest an inverse relationship between the cost of obtaining measures by particular indicators and the degree to which the indicator validly and reliably measures the underlying concept. By the standards of face, construct and predictive validity, family SES and constellation are less accurate but less expensive proxies for aspirations, conditions and processes in the home that facilitate or hinder cognitive ability. Walberg also suggests that the relative efficacy of the British and Chicago school models has yet to be determined (p. 7).

Despite Walberg's contention that the four indicator types attempt to measure the same underlying concept, it appears possible to distinguish the sociological, constellation, and British school variables as indicators of inputs into the process resulting in a particular level of home environment and the Chicago school instrument as an enumeration of the resultant home environment patterns. That is to suggest that SES, constellation, and the British school variables lead to the interaction patterns measured by the Chicago school. To test this hypothesis, the following sequence is proposed as the cognitive component of the theoretical representation of the achievement and motivation process:

$$\begin{tikzcd} X_1 \arrow[r] & X_4 \arrow[r] & y_1 \\ X_2 \arrow[ur] & & \\ X_3 \arrow[uur] & & \end{tikzcd} \tag{8}$$

where X_1 = SES

X_2 = number of siblings

X_3 = parental attitudes and values

X_4 = parent/child interactions

Y_1 = achievement potential

This, however, is only half the model since it was asserted that Head Start also embraced affective goals. Thus, the research on affective outcomes must be investigated to complete the specification of the process model. Two types of relationships need specification. The first concerns the variables describing inputs and outputs of the affective process. The second concerns variables linking the affective and cognitive processes.

According to Lazar, et al. (1978) many intervention programs (including Head Start) specifically set noncognitive goals such as increasing self-esteem (hypothesized above as an intervening goal), enhancing social and emotional development and influencing attitudes related to school success (p. 82). It was assumed that part of the deficiency suffered by disadvantaged children was a lack of educational motivation and goals for the future.

The focus of the affective process, ultimately, is on achievement motivation. The concept was originally developed by H. M. Murray.

Murray, a psychologist, argued that it was possible to identify a variety of innate needs that give the human personality its enduring effects. One of these needs, it was asserted, was a need for achievement (Bigge and Hunt: 99).

The concept was refined by Atkinson and then McClelland. Atkinson asserted that people tend to approach and engage in achievement related tasks given some satisfactory probability of success and avoid task with low probabilities of success. Further, it was assumed that the motive for success would be strongest when people feel responsible for the outcomes of their behavior, when there is quick feedback of results and when there is some risk of failure, although Atkinson assumes that everyone has some motive for success (Bigge and Hunt: 101).

McClelland (1955), hypothesized that achievement motivation was primarily a function of affective determinants and primarily family based. McClelland hypothesized that family behavior and child rearing practices establish learning experiences for the child which, ". . . create enduring personality patterns that persist through adulthood and determine achievement motivation" (Maehr: 204). By encouraging independence, challenge seeking, and delay of gratification through exhortation, modelling or selective reinforcement, the parent not only establishes appropriate behavior patterns but, most importantly, creates affective responses that cause the child to approach or avoid achievement situations (Maehr: 205).

Kahl (1965) took the notion, as it related to compensatory education programs, one step further and discussed achievement orientation where achievement orientation included achievement motivation and those values, attitudes, norms and goals which seem important for success in school and later jobs (Lazar: 85). Lazar cites a paper by Spenner and Featherman (1977) which indicated that achievement motivation in its different forms can play an appreciable independent role in determining academic success.

Bigge and Hunt (1980) state that two elements relevant to achievement motivation theory have only recently been added: (1) a more complete and balanced cognitive theory, and (2) the analysis of how both the causes that people attribute to their wanting to do things and the actual doing of them affects motivation and performance (p. 103). The first point implies that the affective and cognitive processes are interdependent. This will be discussed shortly. The second point concerns research that has been done (Weiner, Rotter, Heider, Deci) on the elaboration of the relationship between achievement and achievement motivation.

Weiner (1979) suggests that the relationship between achievement and motivation for a given individual is mediated by that individual's attribution for achievement, that is, the individual's perception of why the achievement occurred. The most important impact of attribution concerns the locus of control. Internal locus of control implies that an individual will feel he/she was responsible for successful achievement i.e., achievement was due to ability and effort. Those with external locus of control would attribute achievement to factors outside personal control, for example, luck or low task difficulty. Implicitly, the effect of external locus of control is that the individual does not take credit for his/her achievement, thus, no positive effects, such as increased motivation, can occur since this is not viewed as personal accomplishment. On the other hand, those with internal locus of control would perceive achievement as a personal accomplishment, and the payoffs from such achievement may lead the person to higher motivation, i.e., to want to continue to achieve. To the degree that locus of control is related to SES, the relationship between achievement will be stronger for advantaged rather than disadvantaged children. This leads to the hypothesis that if Head Start was ineffective, the relationship between achievement and motivation should be stronger in the control as opposed to the treatment group. This hypothesis will be examined in the data analysis.

From a strictly affective perspective, the prime determinant of motivation is assumed to be the child's self-concept (Vyuroglu and Walberg, 1979). The more capable a child perceives him or herself, the greater the motivation to achieve can realistically be. The concept has several interpretations but Walberg and Uguroglu note that "While there is little agreement regarding one definition, . . . the general factor of self-perception whereas in many motivational measures such as self-concept, selfhood, self-actualization and self-competence." The argument that such self-perception is the most important determinant of motivation has been advanced by Lazar (1978) who considers self-esteem, Cicirelli (1969) who considers self-esteem, Circirelli (1969) who is interested in self-concept and Uguroglu and Walberg who suggest that self-image is reflected in the notion of locus of control, such that, high self-image implies internal locus of control and low self-image suggests external locus.

It is further assumed that parental attitudes affect a child's motivation. However, this relationship is indirect. Parental attitudes relate to the child's development of self-image which, in turn, is related to motivation. Thus, the impact of family on achievement motivation is assumed to be filtered through the child's perception of himself or herself. The affective process is, therefore,

$$\begin{tikzcd} X_5 \arrow[r] & X_8 \arrow[r] & y_2 \\ X_6 \arrow[ur] & & \\ X_7 \arrow[uur] & & \end{tikzcd} \qquad (9)$$

where X_5 = parents' aspiration for the child

X_6 = parents' expectation for the child

X_7 = parents' attitude toward the child

X_8 = self-image

The second point raised in the discussion of attribution theory was that the cognitive and affective process are interrelated. Based on this assumption, a relationship between achievement and motivation was hypothesized. To extend this notion, it is assumed that motivation and self-image flow causally to achievement. A great deal of research has examined the relationship between self-image and achievement. Maehr (1978), Bandura (1977), Bloom (1976), Cattel (1975) and Johnson (1974) specifically point out in their work the importance of the self-view as a primary correlate of learning (Uguroglu and Walberg: 5).

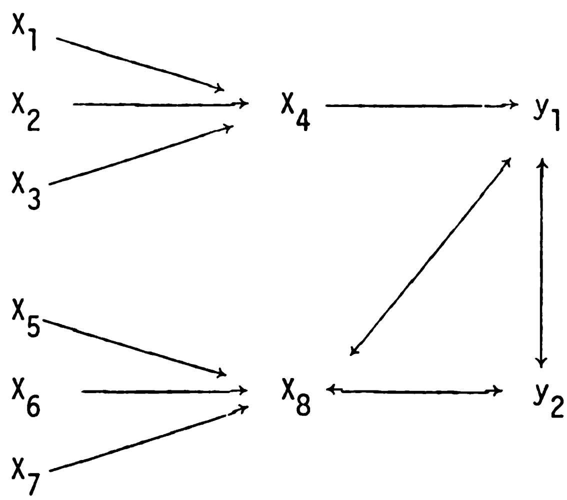
Scheirer and Kraut (1979), however, note that most studies of the relationship between self-concept and achievement fail to reject the null hypothesis. The reason, they suggest, is the faulty causal assumption that self-concept leads to achievement. Rather, they assert that the proper specification is that achievement leads to self-concept (p. 144).

Thus, Scheirer and Kraut assume that attitude is a function of behavior and not vice versa. A logical extension of this argument, which will be pursued here, is that self-concept and achievement exert simultaneous influence. Anderson (1978) tested such a model and found the relationship to be significant in both directions.

A similar argument can be made for the simultaneous relationship between motivation and achievement. In addition, it is assumed that the higher a child's motivation, the higher his/her self-image. These assumptions suggest the argument that achievement, achievement motivation, and self-image are all simultaneously related.

$$\begin{tikzcd} & y_1 \arrow[d, leftrightarrow] \\ X_8 \arrow[ur] \arrow[r, leftrightarrow] & y_2 \end{tikzcd} \tag{10}$$

where the variables have been defined above. The complete model therefore, consists of the following relationships



(11)

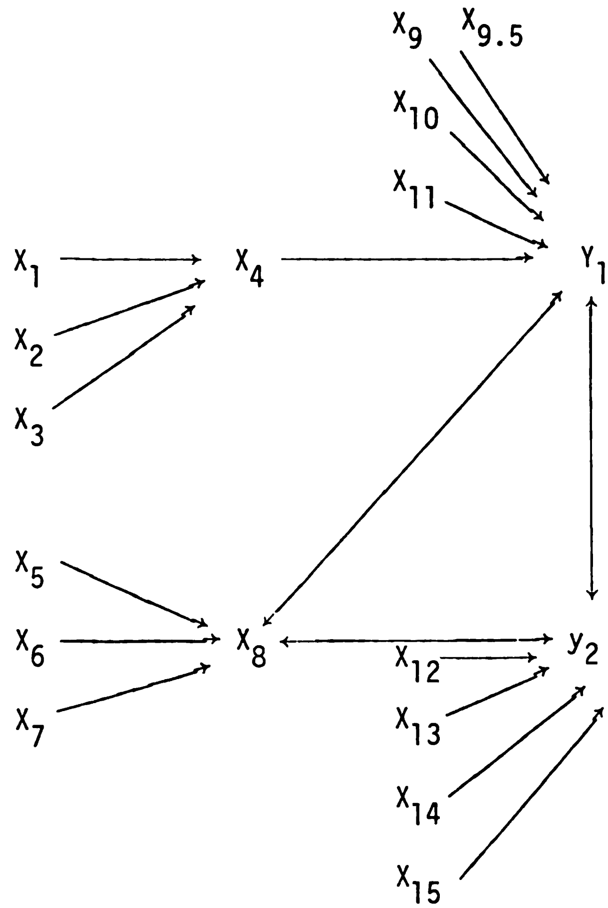
where all variables have been previously defined (see pages 70 and 74). One important point is that the simultaneous relationships may be methodologically tidier than empirically compelling. In particular, the causal relationship between y_2 and X_7 may not be reasonable. If the other simultaneous relationships hold, any y_2/X_8 association may be spurious. Therefore, the data analysis will need to carefully examine these simultaneous relationships.

This model constitutes a representation of the theoretical assumptions, explicit or otherwise, underlying the Head Start program. Given this process, the intent of Head Start was to intervene and ameliorate the inequitous affects of background with respect to cognitive and affective variables. In particular, the program was to intervene between the background variables and intervening goals, home learning environment (X_4) and self-image (X_8). To attain the ultimate goals, achievement (Y_1) and achievement orientation (y_2), the relationships between the intervening and ultimate goals must hold.

To complete the social process model, it is necessary to move beyond program related variables to an inclusion of the variables unaffected by the program but still related to the ultimate program goals. Walberg and Iverson (1979) suggest that some of the variables related to cognitive achievement are sex, race, and age of the child. With respect to achievement motivation, Maehr suggests the importance of post-program variables such as the child's attitude toward norming groups, for example, society at large, teachers, schools, and friends is important. The model as it will be tested has the form shown on the following page.

The test of the research design in Chapter IV will take the following steps: An assessment of the explanatory power of the model examining the control group. Assessment of the effectiveness of the program with respect to the achievement and achievement motivation by between group comparisons, assessment of program failure by between group comparisons with respect to the intervening variables, home learning environment, and self-image, and the assessment of theory failure by testing the explanatory power of the model for the treatment group.

In this case the test is more an assessment of the program designers' interpretation of the academic theory than a formal theory test. Theprogram strategy reflects their understanding of the implications of the theory. Thus, the theory test can occur at more than one level.



(12)

where: X_1 = parental attitudes and aspirations for child's education

X_2 = family constellation

X_3 = socio-economic status

X_4 = home learning environment

X_5 = parental vocation aspiration for child

X_6 = parental attitude toward child

X_7 = parental vocational expectation for child

X_8 = child's self-image

X_9 = race

$X_{9.5}$ = Sex

X_{10} = kindergarten attendance

X_{11} = age

X_{12} = child's attitude toward peers

X_{13} = child's attitude toward school

X_{14} = child's attitude toward home

X_{15} = child's attitude toward society

y_1 = achievement potential

y_2 = achievment motivation

CHAPTER IV

AN APPLICATION OF THE PROPOSED DESIGN TO AN EMPIRICAL EXAMPLE: THE DATA RESULTS

The social process model specified in Chapter III was fit to the data from the Ohio-Westinghouse evaluation of Head Start. One intention of this chapter is to suggest the proposed design has greater utility, on a practical level, than conventional evaluation designs both for the evaluator and design maker. To this end, three sets of inferences generated by the design are reported. The significance of these inferences is that they are unique to designs which explicity call for analysis of process. Thus, they would be unattainable by outcome focussed, experimental evaluation. However, even with the proposed design, the inferences are weak. This is because evaluative inferences are about change, requiring dynamic data, but the design used by the Ohio-Westinghouse evaluators collected data from only one point in time. Inferences can be no stronger than the data used to generate them; inferences concerning change based on the Ohio-Westinghouse data must have somewhat lowered degrees of belief.

The inferences are based on the sorts of analyses permitted by fitting the social process model to a treatment and control group. The analyses involve (1) the treatment group compared to the control group, (2) the treatment group by itself, and (3) the treatment group combined with the control group. The inferences concern (1) program effectiveness, (2) policy concerning compensatory education programs and (3) knowledge of the process of achievement and motivation in young children.

The simultaneous relationships hypothesized in the model of achievement and motivation render the ordinary least squares (OLS) estimates problematic. When a system of equations requires simultaneous solution, OLS estimates are likely to be biased and inconsistent Kmenta (302-303). This is a consequence of the right hand side endogenous variables' correlation with the error term. Consider the two equation system:

$$y_t = \beta_o + \beta_1 z_t + \beta_2 x_t + \varepsilon_1 \tag{1}$$

$$z_t = \beta_o + \beta_1 y_t + \beta_2 R_t + \varepsilon_2 \tag{2}$$

it is likely that for (1) z_t and ε_1 are correlated if

$$z_t = f(y_t)$$

and

$$Y_t = f(\varepsilon_1).$$

Kmenta (1971: 302-303) demonstrates that a consequence of the non-independence of regressors and the error term is inconsistency in the OLS estimates. If the right hand side endogenous variables could be "purged" of the error-related component, the resulting estimates would be asymptotically efficient and consistent.

In the reduced form of a system, each endogenous variable is expressed in terms of the exogeneous variables and disturbances. By computing an instrumental variable Y* as a function of the reduced form coefficients, the Y* would be uncorrelated with error. Substituting the Y*'s into the structural equations for the Y's would produce consistent estimates when the transformed structural equations are estimated by OLS. This procedure is known as two-stage least squares (2SLS) where the first stage is calculation of the instrumental variables by the reduced form coefficients and the second stage is OLS estimation of the transformed structural equations. However, derivation of the reduced form, such that unique solutions for each endogenous variable exist, requires each simultaneous equation to be identified, that is, there must exist unique instruments for each replaced right hand side endogenous variable. The structural equation model to be estimated here has the following form (based on Chapter III).

Equation (1) $X_4 = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon_1$

Equation (2) $X_8 = \beta_{10} + \alpha_{11} Y_1 + \alpha_{12} Y_2 + \beta_{11} X_5 + \beta_{12} X_6 + \beta_{13} Y_2 + \varepsilon_2$

Equation (3) $Y_1 = \beta_{20} + \alpha_{21} X_8 + \alpha_{22} Y_2 + \beta_{21} X_9 + \beta_{22} X_{10} + \beta_{23} X_{11} + \varepsilon_3$

Equation (4) $Y_2 = \beta_{30} + \alpha_{31} X_8 + \alpha_{32} Y_1 + \beta_{31} X_{12} + \beta_{32} X_{13} + \beta_{33} X_{14} + \varepsilon_4$

The structural model is block recursive where block 1 = equation 1 and block 2 = equations 2, 3 and 4. Each equation is overidentified since in each there are more excluded exogenous than included endogenous variables minus one (Shaprio, 1979).

The reduced form model is as follows (solution in Appendix B).

Equation (1R) $X_4 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon_1$

Equation (2R) $$X_8 = \frac{\left[\frac{(\alpha_{32})(\alpha_{22}\alpha_{11} + \alpha_{12})}{(-\alpha_{22}\alpha_{32} + 1)} + (\alpha_{11})\right] B + A}{\left[\frac{(-\alpha_{11}\alpha_{32} - \alpha_{31})(\alpha_{22}\alpha_{11} + \alpha_{12})}{(-\alpha_{22}\alpha_{32} + 1)}\right]} + \frac{\left[\frac{(\alpha_{22}\alpha_{11} + \alpha_{12})}{(-\alpha_{22}\alpha_{32} + 1)}\right] C}{(-\alpha_{21}\alpha_{11} + 1)}$$

Equation (3R) $$Y_1 = \frac{\left[\frac{(\alpha_{22})(\alpha_{31}\alpha_{12} - 1)}{(-\alpha_{31}\alpha_{22} - \alpha_{21})} + (\alpha_{12})\right] C + A + \left[\frac{(\alpha_{31}\alpha_{12} - 1)}{(\alpha_{31}\alpha_{22} - \alpha_{21})}\right] B}{\left[\frac{(-\alpha_{32}\alpha_{22} - 1)(\alpha_{31}\alpha_{12} - 1)}{(-\alpha_{31}\alpha_{22} - \alpha_{21})}\right] + (-\alpha_{32}\alpha_{12} - \alpha_{11})}$$

Equation (4R) $$Y_2 = \frac{\left[\frac{(\alpha_{31})(\alpha_{11}\alpha_{21} - 1)}{(-\alpha_{11}\alpha_{31} - \alpha_{32})} + (\alpha_{21})\right] A + B + \left[\frac{(\alpha_{11}\alpha_{21} - 1)}{(-\alpha_{11}\alpha_{31} - \alpha_{32})}\right] C}{\left[\frac{(-\alpha_{12}\alpha_{31} + 1)(\alpha_{11}\alpha_{21} - 1)}{(-\alpha_{11}\alpha_{31} - \alpha_{32})}\right] + (\alpha_{12}\alpha_{21} - \alpha_{22})}$$

where $A = (\beta_{10} + \beta_{11} X_5 + \beta_{12} X_6 + \beta_{13} X_7 + \varepsilon_2)$

$B = (\beta_{20} + \beta_{21} X_9 + \beta_{22} X_{10} + \beta_{23} X_{11} + \varepsilon_3)$

$C = (\beta_{30} + \beta_{31} X_{12} + \beta_{32} X_{13} + \beta_{33} X_{14} + \varepsilon_4)$.

However, the use of 2SLS rather than OLS in the presence of simultaneity does not necessarily lead to optimal estimates (Shapiro, 1979). Because consistency is an asymptotic property, the variance of the two-stage estimates may be large in finite samples. In particular, while the bias of the 2SLS estimates is smaller than for OLS estimates, the variance tends to be greater. Thus, the choice between the 2SLS and OLS estimates is a function of the trade-off between the deviation of the estimates from the true parameter values and the precision of the estimates (Shapiro: 349). If the true parameter values were known, a method for choosing the estimate with the best "mix" of bias and variance would be to calculate and compare the mean square error (MSE) of the estimates for a particular equation where MSE is defined as

$$\text{MSE}(\hat{\theta}) = \text{Variance } (\hat{\theta}) + [\text{Bias } \hat{\theta}]^2 \qquad \text{(Rao and Miller: 64)}$$

and utilize the minimum mean square error estimates.

In the absence of information on true parameter values, the choice between the OLS and 2SLS estimates is unclear. Johnston (1972) discusses a variety of Monte Carlo studies which compared the properties of OLS and some simultaneous methods (including 2SLS) under particular conditions and reports that the differences amont methods tend to be slight but 2SLS generally outperforms OLS (417). The model specified in Chapter III was estimated by both procedures. The variables included in the analysis are listed in Table 1 and Appendix A. As indicated in Table 2, the OLS procedure yielded a better fit to the data particularly for the affective equations. Thus the

TABLE 1.--Variables Used in the Data Analysis

Variable Name	Concept	Operationalization
HLE	Home learning environment--parent/child interactions and child's behavior related to achievement	Scale of parent/child interactions and child's behavior in the home
ACH	Achievement potential	Mean of the nonzero scores on the subunits of the Metropolitan Readiness Test
SELF	Self-image	Scale of self-concept questions where the child selects which of two figures he/she more closely resembles
MOTIV	Achievement orientation or motivation	Teacher's assessment of the child's achievement motivation by the Children's Behavior Inventory
EDASP	Parental aspiration for child's educational attainment level	Single item coded from finish grade school to attend graduate school
EDEXP	Parental Expectation for child's educational attainment level	Single item coded as for EDASP
VOCASP	Parental aspiration for child's occupation upon completion of schooling	Single Item coded from unskilled worker to major professional
VOCEXP	Parental expectation for child's occupation upon completion of schooling	Single item coded as for VOCASP
SIBS	Number of siblings	Number of children living at home up to nine
SES	Socio-economic status	Scale of parental educational and occupational attainment, plus income

TABLE 1.--Continued

Variable Name	Concept	Operationalization
BEHAV	Parental behavioral response to child's educational and occupational decisions	Scale of items of hypothetical situations calling to child's educational and occupational decisions
CONSERV	Degree of parental conservatism concerning the desirability of school treating the whole child as opposed to teaching the basics	Scale of questions concerning the appropriate scope of school concerns where a higher score indicates lower conservatism
DEEMP	Deemphasis of education by parents, particularly deemphasis of the importance of achievement	Scale of attitude questions where a higher score indicates lower deemphasis
FUTILE	Parental futility about the possible positive effect of education on their children's lives	Scale of attitude questions where a higher score indicates lower futility
GRIPES	Parental disapproval of the condition of their child's school	Scale of attitudes question where a higher score indicates less gripes or higher satisfaction
IMP	Importance of education to children's lives	Scale of attitude questions where a higher score indicates lower importance
SEX	Gender of child	Response to question, "Are you male or female"
KIND	Kindergarten attendance	Response to question whether or not child attended a kindergarten
RACE	Race	Response to question, "Are you White, Black, Mexican American, Puerto Rican, American Indian, or other"
AGE	Age	Question coded from 5 years to 10 years by year

TABLE 1.--Continued

Variable Name	Concept	Operationalization
VASP	Parental aspiration for child's ultimate vocational attainment	Scale of items where for each item parents choose one of three listed occupations which they would most like their child to have
VEXP	Parental expectation for child's ultimate vocational attainment	Scale of items as in VASP except parents choose occupations they think is most likely to be attained by their child
ATT	Parental attitude toward child	Scale of items indicating type and intensity of parent/child relationship
SCHOOL	Child's attitude toward school	Scale of attitude questions about school situation and sad, happy and neutral faces. Child selects face indicating either negative, neutral or positive attitude. Higher scores indicate positive attitude, median scores indicate neutral attitude and lower scores indicate negative attitude
HOME	Child's attitude toward the home	Scale of attitude questions about the home scored as for school
PEERS	Child's attitude toward peers	Scale of attitude questions about peers scored as for school
SOCIETY	Child's attitude toward society	Scale of attitude questions about society scored as for school
GROUP	Group assignment	Response to question of being in treatment or control group

TABLE 2.--A Comparison of the R-Square for the OLS and 2OLS Estimates of the Full Causal Model (N = 432)

Dependent Variable	Procedure	R^2
HLE	OLS	.3859
HLE	2SLS	.3859
SELF	OLS	.1639
SELF	2SLS	.0761
ACH	OLS	.3304
ACH	2SLS	.2197
MOTIV	OLS	.2228
MOTIV	2SLS	.0936

rather arbitrary decision was made to utilize the OLS estimates despite the known biases primarily because the 2SLS results are generally uninterpretable.

In light of the decision to utilize the OLS estimates, it must be recognized that if the sample sizes are "sufficiently large," the standard errors are likely to be inflated, resulting in conservative significance tests of the individual partial slopes. The explanation for the low R^2 for the 2SLS in this particular data set will emerge in the context of the theoretical relationships found in the causal model.

The initial set of results concerns the conventional evaluation issue of program effectiveness. With an experimental or quasi-experimental research design, decisions about program effectiveness are based on comparisons of the treatment and control group on relevant outcome measures. The primary argument in this dissertation is that if evaluation includes explicit assessment of the program process as well as outcomes, useful information, for example accounting for the success or failure of a program, can emerge. In particular, with respect to program effect, estimation of the social process model permits the assessment of program failure and theory failure, a distinction which necessarily goes unattended in experimental and quasi-experimental research. As a baseline with which the results of the analysis can be compared, the following is a brief description of the results (for the first grade) of the original Ohio-Westinghouse evaluation and a reanalysis of the data by Smith and Bissell (1970).

The Report of the Westinghouse-Ohio National Evaluation of Head Start was issued in April of 1969. The report focussed on both the summer and year-long programs and their effects through three years of school. The analysis was conducted as an ex-post facto quasi-experiment, of the form (Campbell and Stanley, 1963)

$$\begin{array}{cc} X & O_1 \\ \hdashline & O_1 \end{array}$$

for each of the three years analyzed. The determination of program impact in each case was on analysis of covariance using variables such as socio-economic status as covariates. The basic question, according to the executive summary (1969), that the evaluators confronted was

> To what extent are the children now in the first, second, and third grades who attended Head Start programs different in their intellectual and social-personal development from comparable children who did not attend? (Caro: p. 343).

The overall finding, according to the evaluators, was

> In sum, the Head Start children cannot be said to be appreciably different from their peers in the elementary grades who did not attend Head Start in most aspects of cognitive and affective development measured in this study, with the exception of the slight but nonetheless significant superiority of full-year Head Start children on certain measures of cognitive development" (Caro: 346).

This general statement accurately reflects the specifics of the first grade results. Two cognitive measures, the Metropolitan Readiness Test and the Illinois Test of Psycho-linguistic Abilities, and two affective measures, the Self-Concept Index and the Cumulative Behavior Inventory, were applied to the full year and summer Head Start and control group samples. The summer program was found to not have an impact on either cognitive or affective outcomes. Although the full year Head Start groups also was not superior on either affective measure, small but statistically significant gains were found both for the Metropolitan Readiness subtest for listening and for the overall test score. Thus, the general conclusion for the first grade year-long program was limited cognitive impact and no affective impact. For a program intended to treat the "total" child, such results were viewed as negative and disappointing.

In an effort to mitigate the negative impact of the Ohio-Westinghouse result, Smith and Bissell (1970) reanalyzed a portion of the data and claimed to find a far more positive influence of the program. On inspection, however, it must be concluded that the specifics of the reanalysis, in light of the original findings, tended to constrain the results to a particular, strongly positive, outcome. For example, although the affective Head Start goals were as important as the cognitive ones, Smith and Bissell chose to reanalyze only the cognitive data. They do not indicate the reason for their decision (p. 79), but in the original evaluation, the only positive outcomes were cognitive ones.

Although in the original analysis, both summer and year-long samples were selected for three years, Smith and Bissell examined only the first grade, year-long sample. They selected the sample, they say, because there is little evidence to suggest a significant impact in summer programs (p. 79) and because the first grade sample is least likely to confound the impact of Head Start with schooling. It also happens that the first grade year-long sample was the only group for which the original evaluation found a statistically significant cognitive impact.

Smith and Bissell analyzed only the Metropolitan Readiness scores even though the Ohio-Westinghouse evaluators also administered the Illinois Test of Psycholinguistic Abilities. They suggest that the reason for focussing on the MRT was the high reliability of the test and the traditional use of readiness tests by elementary schools as a cue for relating to children as students (p. 80). (It was only for the Metropolitan Readiness Test that significant gains for the treatment group were found by the original evaluators.)

Finally, Smith and Bissell reduced the original sample of 432 first grade, full-year treatment and control group subjects to a subsample for which the greatest gains were documented in the original study (p. 90). Thus, the "reanalysis" was performed on the subsample (N=40) for which the greatest gains had been observed, taken from the only sample for which statistically significant results were obtained, utilizing only the one specific cognitive test for which statistically significant results were obtained. Consequently, the not surprising result they reported was, ". . . the Head Start Group scored significantly higher than the control group on the Metropolitan Readiness Test by a large enough margin for us to consider the differences 'educationally significant'" (p. 101). Their effort, clearly, is not a reanalysis but a reassertion of the original findings that Head Start had some significant cognitive impact.

Subsequent reanalyses by Barrow (1973) and Magidson (1977) have shown that a positive cognitive impact occurred for the summer group although Bentler and Woodward (1978) have challenged Magidson's findings. However, it is still the case that no reanalysis of the Ohio-Westinghouse data has shown the original evaluation to be in error. Therefore, for the full-year first grade sample, the base line result remains: the program exerted some small but statistically significant impact on cognitive outcomes but no significant impact on affective outcomes.

The assessment of program effectiveness was accomplished by comparing the treatment and control groups in terms of the achievement and motivation process. This was done by fitting the social process model to the treatment and control groups separately. The original estimation was done by both OLS and 2SLS, and Table 3 indicates that for each group, the OLS estimates provided better fit.

The OLS estimation was applied twice in each group. The initial estimation of the full causal model included variables that proved to be statistically nonsignificant. Since the inclusion of irrelevant explanatary variables reduces the efficiency of the OLS estimates (Kmenta: 396-399), a second set of equations was specified for each group where the regressors were only those variables found significant at .075 in the test of the full models. The reason for decreasing the critical value from the conventional .05 level was a recognition that the use of OLS to estimate simultaneous relationships potentially leads to conservative t-tests when sample sizes are sufficiently large. This was an attempt to avoid type II errors for borderline cases given the likely properties of the significance tests. This involved two variables, both of which were found to be significant at the .05 level in the predictive models. The results

TABLE 3.--A Comparison of the R-Square for the OLS and 2SLS Estimates of the Full Causal Model for the Treatment and Control Samples ($N_T = N_C = 216$)

Dependent Variable	Procedure	R^2
	Treatment	
HLE	OLS	.3637
HLE	2SLS	.3637
SELF	OLS	.1716
SELF	2SLS	.0667
ACH	OLS	.3089
ACH	2SLS	.1601
MOTIV	OLS	.2383
MOTIV	2SLS	.0886
	Control	
HLE	OLS	.4355
HLE	2SLS	.4355
SELF	OLS	.1964
SELF	2SLS	.1196
ACH	OLS	.3966
ACH	2SLS	.2501
MOTIV	OLS	.2299
MOTIV	2SLS	.1453

of the estimation of the full and predictive models for the treatment and control groups are contained in Tables 4 and 5.

The argument underlying the distinction between program failure and theory failure is that, from a decision maker's perspective, accounting for program effectiveness or ineffectiveness is a necessary precondition for valid program policy decisions. In particular, it has been argued (Chapter I) that certain failures should lead to the cancellation of a program while others may simply result in program modification. Similarly, successes which cannot be accounted for by program activities should not necessarily lead to positive program decisions since the true cause of the success may not be present when the program is continued or expanded (Suchman: 86-87).

The distinction between program failure and theory failure as the root of program ineffectiveness is the distinction between the failure of a program to attain its intervening goals and the failure of the intervening goals to be causally related to the ultimate program objectives. It has already been suggested that the assessment of program failure could be accomplished by comparing the treatment and control groups in the relevant indicators while theory failure could be assessed by theory testing (Chapter I).

A comparison of the treatment and control groups on the outcome measures achievement potential and achievement motivation indicates the failure of the Head Start program in both the affective and cognitive domains (Table 6).

TABLE 4.--Results of the OLS Estimations of the Full and Predictive Models for the Treatment Group (N = 216)

Independent Variable	Standardized β	PROB >|T|
	Full Model	
Dependent Variable: HLE	4-A PROB > F: .0001	R-SQUARE: .3637
BEHAV	.1781795	.0026
CONSERV	.07670027	.2408
DEEMP	.05359591	.3881
FUTILE	.23281390	.0026
GRIPES	-.05444690	.3929
IMP	.01836867	.7634
SES	.08247331	.2497
EDASP	.25259253	.0002
EDEXP	.09085303	.1297
VOCASP	-.04873454	.4685
VOCEXP	-.00993399	.8808
SIBS	-.16529512	.0050
Dependent Variable: SELF	4-B PROB > F: .0001	R-SQUARE: .1716
ATT	.12670180	.0739
MOTIV	.14419408	.0473
EDASP	.02043630	.7716
EDEXP	.08023104	.2459
VOCASP	-.05897304	.4247
VOCEXP	-.05250511	.4864
ACH	.29508340	.0001
VASP	.11399656	.1736
VEXP	-.15466791	.0833

TABLE 4.--Continued

Independent Variable	Standardized β	PROB >|T|
	4-C	
Dependent Variable: ACH	PROB > F: .0001	R-SQUARE: .3089
HLE	.12772097	.0366
MOTIV	.33951356	.0001
SELF	.20273565	.0011
SEX	-.01858952	.7540
KIND	-.04024300	.4934
RACE	-.07800079	.1905
AGE	.21043485	.0005
	4-D	
Dependent Variable: MOTIV	PROB > F: .0001	R-SQUARE: .2383
HOME	-.0775770	.4360
PEERS	.19476702	.0634
SCHOOL	-.04406022	.6838
SELF	.0100815	.1317
SOCIETY	.05675828	.5991
ACH	.44410722	.0001
	4-E	
Dependent Variable: HLE	PROB > F: .001	R-SQUARE: 3435
BEHAV	.18324020	.0015
FUTILE	.30992877	.0001
EDASP	.26843590	.0001
SIBS	-.18336451	.0012
	4-F	
Dependent Variable: SELF	PROB > F: .0001	R-SQUARE: .1459
MOTIV	.14168211	.0468
ACH	.27606603	.0001

TABLE 4.--Continued

<table>
<tr><th>Independent Variable</th><th>Standardized β</th><th>PROB >|T|</th></tr>
<tr><td colspan="3">4-G</td></tr>
<tr><td>Dependent Variable: ACH</td><td>PROB > F: .0001</td><td>R-SQUARE: 3006</td></tr>
<tr><td>HLE</td><td>.13225337</td><td>.0271</td></tr>
<tr><td>MOTIV</td><td>.32893084</td><td>.0001</td></tr>
<tr><td>SELF</td><td>.21765590</td><td>.0004</td></tr>
<tr><td>AGE</td><td>.1991127</td><td>.0008</td></tr>
<tr><td colspan="3">4-H</td></tr>
<tr><td>Dependent Variable: MOTIV</td><td>PROB > F: .0001</td><td>R-SQUARE: .2266</td></tr>
<tr><td>PEERS</td><td>.17289260</td><td>.0048</td></tr>
<tr><td>ACH</td><td>.46212547</td><td>.0001</td></tr>
</table>

NOTE: Variables with statistically significant coefficients are underlined (.05).

TABLE 5.--Results of the OLS Estimations of the Full and Predictive Models for the Control Group (N = 216)

Independent Variable	Standardized β	PROB >|T|
	Full Model	
	5-A	
Dependent Variable: HLE	PROB > F: .0001	R-SQUARE: .4355
BEHAV	.07878787	.1573
CONSERV	.08132765	.1958
DEEMP	.01193914	.8325
<u>GRIPES</u>	-.11309756	.0590
IMP	-.01662113	.7747
<u>SES</u>	.15151980	.0497
<u>EDASP</u>	.25573909	.0001
EDEXP	.01922029	.7411
VOCASP	-.04177586	.5237
VOCEXP	-.04177586	.4828
SIBS	-.07110054	.2066
	5-B	
Dependent Variable: SELF	PROB > F: .0001	R-SQUARE: .1964
ATT	-.06140048	.3873
<u>MOTIV</u>	.16887051	.0259
EDASP	.08998658	.2132
EDEXP	-.00794319	.9075
VOCASP	-.02963294	.6700
VOCEXP	-.07843064	.2663
<u>ACH</u>	.31665178	.0001
VASP	-.05563141	.4798
VEXP	.00582852	.9429

TABLE 5.--Continued

Independent Variable	Standardized β	PROB > |T|
Dependent Variable: ACH	PROB > F: .0001	R-SQUARE: .3966
HLE	.27993238	.0001
MOTIV	.30396902	.0001
SELF	.25244421	.0001
SEX	.07800378	.1620
KIND	-.17258329	.0025
RACE	-.04328100	.4477
AGE	.09383141	.0856
	5-D	
Dependent Variable: MOTIV	PROB > F: .0001	R-SQUARE: .2299
HOME	.09932634	.3551
PEERS	.08101707	.4720
SCHOOL	.02305629	.8351
SELF	.12263074	.0686
SOCIETY	-.19127094	.0717
ACH	.37096017	.0001
	5-E	
Dependent Variable: HLE	PROB > F: .0001	R-SQUARE: .4134
FUTILE	.33721312	.0001
GRIPES	-.11371088	.0351
SES	.20026919	.0053
EDASP	.27445095	.0001
	5-F	
Dependent Variable: SELF	PROB > F: .0001	R-SQUARE: .1758
MOTIV	.12608512	.0720
ACH	.34689401	.0001

TABLE 5.--Continued

Independent Variable	Standardized β	PROB > |T|
	5-G	
Dependent Variable: ACH	PROB > F: .001	R-SQUARE: .3895
HLE	.28982720	.0001
MOTIV	.31654957	.0001
SELF	.26290572	.0001
KIND	-.17827622	.0013
	5-H	
Dependent Variable: MOTIV	PROB > F: .0001	R-SQUARE: .2165
SELF	.11986401	.0720
ACH	.40372373	.0001

TABLE 6.--A Comparison of Outcome Measures Between the Treatment and Control Groups by ANOVA and ANCOVA

6-A

ANOVA ($N_T = N_C = 216$)

Variable	$\overline{X}_T$	S_T	$\overline{X}_C$	S_C	Significance
ACH	8.5120	2.8659	8.1991	2.6648	.2150
MOTIV	55.6772	21.5440	57.2255	20.3650	.4412
HLE	19.9722	6.441	19.3472	6.0267	.2984
SELF	33.1806	8.8984	33.6296	6.8541	.5552

6-B

ANCOVA

Variable	Significant of Treatment Coefficient
ACH	.0698
MOTIV	.2127
HLE	.0767
SELF	.3652

It should be noted that the failure of the program to attain the ultimate cognitive goal, in contrast with the original evaluation finding, may be a function of using a different set of covariates; the Ohio-Westinghouse evaluators only included SES in their analysis of covariance. What must be emphasized is that the primary issue of the proposed approach is not whether the treatment group statistically outperformed the control group (although it is important) but whether the outcomes can be accounted for by the evaluative hypothesis such that external validity of the program impact is maximized.

Having established the ineffectiveness of the program, the import of the causal modelling methodology is that for both the affective and cognitive failure, the type of failure, and therefore the possibility of rectifying the failure, can be determined. If, as has been the assumption, the program could not directly influence the ultimate goals, then the failures of Head Start must be explained in terms of the intervening goals and the relationship between the intervening and ultimate goals.

Inspection of Table 6 indicates that for both the intervening affective and cognitive goals, program failure occurred. That is, the failure of the program to significantly affect achievement potential and achievement motivation for the treatment group is a function of the failure of the program to significantly affect the home learning environment and self-image of the treatment group. The critical question is, what kinds of failure are these.

Inspection of Table 4 for theory failure indicates that the cognitive and affective failures are of two different types. If the theory failure concept is expanded to include not only the criterion of a causal relationship between the intervening and ultimate goals, but also the crucial question of the manipulability of the intervening goal, the differing natures of the affective failure and cognitive failure become apparent. With respect to the cognitive process, the results of the estimation of the causal model indicate the essential validity of the theoretical specification. The results indicate that the home environment can be manipulated by modifying the parents' attitude toward education (FUTILE), their aspiration for the child's educational attainment (EDASP) and their attitude toward their child's educational decisions (BEHAV) as well as by directly influencing the child's behavior (elements of HLE). Secondly, the significant standardized β (.13) for the regression of achievement on home learning environment indicates that ceteris paribus, the greater the impact of the program on home learning environment, the greater the impact of the program on achievement. Thus, a program strategy of increasing achievement by enriching the environment should be moderately successful (note the magnitude of β HLE compared to the others in Table 4-G) and, therefore, attainment of the ultimate cognitive goal by Head Start is feasible pending program operations which would attain the intervening goal.

The assessment of the theoretical viability of the affective process leads to quite different conclusions. Table 4-F indicates that, for the treatment group, there exists no set of exogenous variables by which changes in self-image could be affected, i.e., there is no indication of how to attain the goal. Furthermore, attainment of the intervening goal (though desirable) would not lead to attainment of the ultimate affective goal owing to the lack of a significant causal relationship (Table 4-F, β SELF) between self-image and achievement motivation.

Based on the results of the program failure/theory failure assessment, program designers would have to consider dropping the affective goals from statements of program intent. Thus, the first major inference to emerge from the proposed method: (Program impact inference)

> The affective and cognitive failures of the Head Start program are of two different kinds. The cognitive failure is simple program failure, which can be rectified by program modification. The affective failure is theory failure, which could not be rectified by changes in the program. The affective goals must be considered unattainable.

Additional relevant information about program impact can be derived from a comparison of the estimated models for the treatment and control groups. Two questions of interest are (1) how did the program affect the overall process of achievement and motivation for the treatment group (not just outcomes), and (2) is the theoretical specification a sufficient representation of the causal process in the two groups.

Methodologically, the first question can be handled by comparing the predictive model for each group in terms of the patterns of significant variables and the coefficients of commonly significant variables. The technique used to assess the impact of the program on the causal process is a two-stage Chow test where the first stage is an application of Gujarati's dummy variable procedure (1970) and the second stage replaces each structural variable by two dummy variables (suggested in a discussion, by Edward Haertel).

In the stage one Chow test dummy variables measuring the variable by treatment interaction were included in regressions which contained the predictive model regressors for the combined data set. A significant coefficient for any dummy variable indicates a significant treatment by variable interaction. Table 7 indicates the results of the stage one Chow tests.

For significant dummy variables the stage two Chow tests replaced the relevant structural variables and variable by treatment interaction with two dummy variables: a variable by treatment interaction and a variable by control interaction. The second stage Chow test allowed direct comparisons of the differences between the treatment and control groups for the significant stage one interactions.

In an equation by equation inspection of group differences, the greatest impact of treatment is found in the achievement potential (ACH) equation. In the achievement equation age, the impact of kindergarten and the impact of the home learning environment are all interactively significant. The interactive effect of kindergarten attendance, a nonsignificant influence in the treatment group, is perhaps the easiest to explain. Since the treatment group's initial school experience is the Head Start program, kindergarten may simply duplicate that experience and provide no unique contribution to the

TABLE 7.--Results of the Stage One Chow Tests for Differences Between the Treatment and Control Groups on Significant Structural Variables

Independent Variable	Standardized β	F
Dependent Variable: HLE	7-A F 13, 418 at .05 = 1.64	
Intercept x Treatment	-.24237	.714
EDASP x Treatment	.03373	.043
SIBS x Treatment	-.10933	1.075
BEHAV x Treatment	.28539	2.316
FUTILE x Treatment	-.00425	.000
GRIPES x Treatment	.04016	.825
SES x Treatment	-.07183	.275
Dependent Variable: SELF	7-B F 7, 424 at .05 = 2.03	
Intercept x Treatment	-.41611	2.709
MOTIV x Treatment	.01978	.016
ATT x Treatment	.43853	3.462
Dependent Variable: ACH	7-C F 11, 420 at .05 = 1.81	
Intercept x Treatment	.01955	.007
KIND x Treatment	.16462	2.449
AGE x Treatment	.26212	2.758
HLE x Treatment	.26506	3.281
MOTIV	.03271	.064
Dependent Variable: MOTIV	7-D F 7, 424 at .05 = 2.03	
Intercept x Treatment	-.22343	.961
SELF x Treatment	-.07546	.172
ACH x Treatment	.-2887	.033
PEERS x Treatment	.23051	2.207

TABLE 8.--Results of the Stage Two Chow Test for Differences Between the Treatment and Control Groups on Significant Variable x Treatment Interactions

Independent Variable	Standardized β	F
8-A Dependent Variable: HLE	F 6, 425 at .05 = 2.12	
BEHAV x Treatment	.34106	12.828
BEHAV x Control	.26615	7.818
8-B Dependent Variable: SELF	F 5, 426 at .05 = 2.23	
Intercept x Treatment	-.45586	4.167
Intercept x Control	.45586	4.167
ATT x Treatment	.30773	3.993
ATT x Control	-.12564	0.553
8-C Dependent Variable: ACH	F 8, 423 at .05 = 1.96	
KIND x Treatment	.06663	.863
KIND x Control	.021529	9.295
AGE x Treatment	.399640	16.774
AGE x Control	.77160	3.413
HLE x Treatment	.228560	6.621
HLE x Control	.510110	29.526
8-D Dependent Variable: MOTIV	F 5, 426 at .05 = 2.23	
PEERS x Treatment	.14274	3.128
PEERS x Control	.18008	4.956

child's achievement potential. The explanations for the interactive effects of age and home learning environment are less clear.

A possible explanation for the differential effect of home learning environment on achievement is based on a possibly unanticipated consequence of the program. It is clear that one goal of Head Start was to provide a better home learning environment, indicated by the intention of making the parents and family of the treatment group the primary agents of change (Datta: 6). The data indicate, however, (Table 8-C) that the effect of the progam was to reduce the strength of the relationship between home learning environment and achievement. It may be, despite the emphasis on family, that the primary impact of the program on the children occurred at the program center, rather than in the home. This effect would not be reflected in the patterns of parent/child interactions in the home. For the participants, as opposed to the control group, achievement may be much more a function of the internalized variables self-image and achievement motivation rather than a function of the external influence home learning environment. Clearly, further research on this relationship is required. Finally, as indicated in Tables 7 and 8, no other equation displays such substantive interactive effects of treatment.

With respect to the question of the sufficiency of the theoretical specification of the causal process for the treatment and control groups dynamics, the data indicate that, except for the achievement orientation equation, the causal model does a better job

TABLE 9.--A Comparison of the Proportion of Explained Variance in the Treatment and Control Group Predictive Models

Dependent Variable	Group	R^2
HLE	Treatment	.3435
HLE	Control	.4134
SELF	Treatment	.1459
SELF	Control	.1758
ACH	Treatment	.3006
ACH	Control	.3895
MOTIV	Treatment	.2266
MOTIV	Control	.2165

of explaining the relationship in the control group. The conclusion to be drawn from this, particularly under the assumption of essentially equivalent groups, is that unanticipated treatment effects are introducing disruptions of the "normal" causal relationships, rendering the original theoretical specifications insufficient for the treatment group. Consequently, future evaluations of Head Start would require more elaborate specifications of the process model such that more accurate assessment of treatment impact would be possible.

The second major set of inferences unique to the proposed methodology concerns the use of the results in making recommendations for compensatory education programs in general. The question of interest is, do the sample results for the treatment group reflect the type of programs which would maximally treat disadvantaged children? If it can be assumed that the treatment group constituted a representative sample of disadvantaged children, one policy relevant result clearly stands out. The result concerns the optimal structure of compensatory education programs given both affective and cognitive goals. In light of the empirical support for the cognitive hypotheses, and the nonsupport for the affective hypotheses, it must be concluded that programs focussing on strictly cognitive inputs will result in more systematic and predictable cognitive outcomes than those attempting to manipulate self-image and achievement motivation. However, one strongly supported causal influence on self-image and achievement motivation is indicated, namely achievement. Thus, while achievement is viewed as a cognitive output variable, with respect to self-image and achievement motivation, it was actually an input [see Tables 4-F (β ACH) and 4-H (β ACH)]. The results indicate that achievement, self-image and achievement motivation can best be maximized by compensatory education programs that focus <u>exclusively on cognitive inputs and strategies</u> and allow the child's increased achievement (assuming an effective program) to lead to gains in self-image and achievement motivation. This position has been argued by Bereiter and Engleman with respect to their successful compensatory education model (1966). Thus, the second major inference arising from the causal modelling methodology is: (Policy Inference)

> Compensatory education programs should focus exclusively on cognitive inputs and strategies. If successful, such a program would not only lead to increased achievement potential, but ultimately to improved self-image and achievement motivation.

A corollary inference concerns the implications of the simultaneous relationships found to exist between achievement and achievement motivation and achievement and self-image. Since the estimation indicates that, dynamically, higher achievement will lead to higher self-image and motivation, and ultimately to even higher achievement, the true effect of a compensatory education program (1) will not be adequately represented by a simple pretest/posttest design and (2) may not be capable of being accurately estimated by existing evaluation methodologies.

The third major inference due to the causal modelling methodology concerns the state of theoretical knowledge about the achievement and motivation process. The decision to estimate the process model in a combined treatment and control group sample, to take advantage of the positive effects of a large sample size, is conditioned upon the results of the Chow test discussed earlier. The danger of combining the samples without applying the Chow test is the possibility of masking significant interaction terms. In fact, the original purpose of the Chow test was to determine whether the regression coefficients from separate samples were similar enough to permit estimation from a combined sample (Datta: 173-174). Having already investigated the significant interactions, it was decided that combining the samples would not lead to any incorrect inferences and that the combined sample would have greater external validity for inferences to a general pre-school population.

The results of the combined sample theory test indicated that about 35 percent of the variance in the cognitive variables and about 18 percent of the variance in the affective variables is accounted for.

As originally hypothesized, family constellation, parental attitudes and SES are all found to impact on home learning environment. The equation for achievement potential also was empirically supported, causal influences being the home learning environment, age and previous school attendance, and self-image and achievement motivation. One mildly surprising finding is the lack of a relationship between race and achievement. One possible explanation is that the concepts for which race can be a proxy, for example, home environment or attitudes, are explicitly incorporated in the model, leaving no unique contribution to be made by race.

As in previous cases, the results for the affective equations are weak. In fact, all the explained variance in self-image and almost all the explained variance in achievement motivation (where attitude toward peers was statistically significant) are due solely to the relationship of each with achievement and the other affective variable. Consequently, the simultaneous nature of the affective variables and achievement has been verified.

Several interesting implications of the theory testing results can be identified. The first, of course, is the substantial impact

TABLE 10.--Results of the OLS Estimation of the Full and Predictive Causal Models for the Combined Sample (N = 432)

10-A

Dependent Variable: HLE PROB > F: .0001 R-SQUARE : .3859

Independent Variable	Standardized β	PROB > $\lvert T \rvert$
EDASP	.25049728	.0001
EDEXP	.05570617	.1753
VOCASP	-.04324742	.3251
VOCEXP	-.02871574	.5140
SIBS	-.12012762	.0028
SES	.11691079	.0386
BEHAV	.12638538	.0016
FUTILE	.28356438	.0001
CONSERV	.07062244	.1134
IMP	-.00868929	.8341
DEEMP	.02788315	.4962
GRIPES	-.07665772	.0756

10-B

Dependent Variable: SELF PROB > F: .0001 R-SQUARE: .1639

Independent Variable	Standardized β	PROB > $\lvert T \rvert$
EDASP	.0474128	.3450
EDEXP	.4635176	.3359
VOCASP	-.04454357	.3770
VOCEXP	-.06320957	.2158
ACH	.29294768	.0001
VASP	.02432341	.6682
VEXP	-.05674001	.3372
MOTIV	.13824166	.0068
ATT	.03516506	.4698

TABLE 10.--Continued

Independent Variable	Standardized β	PROB > |T|
10-C Dependent Variable: ACH	PROB > F: .0001	R-SQUARE: .3304
SEX	.03841423	.3446
KIND	-.09841059	.0165
RACE	-.06222178	.1323
AGE	.14665075	.0003
HLE	.20019297	.0001
MOTIV	.3269982	.0001
SELF	.23326110	.0001
10-D Dependent Variable: MOTIV	PROB > F: .0001	R-SQUARE: .2228
AHC	.40454549	.0001
SCHOOL	-.02536922	.7397
HOME	-.00187574	.9794
PEERS	.16185268	.0330
SOCIETY	-.06744206	.3698
SELF	.12034217	.0098
10-E Dependent Variable: HLE	PROB > F: .0001	R-SQUARE: .3795
EDASP	.25594224	.0001
SIBS	-.12732560	.0014
SES	.11243873	.0267
BEHAV	.12857804	.0012
FUTILE	.29459424	.0001
GRIPES	-.08658262	.0381
10-F Dependent Variable: SELF	PROB > F: .0001	R-SQUARE: 1491
ACH	.30357779	.0001
MOTIV	.13913450	.0054

of the home on early scholastic success, the basic assumption upon which the Head Start program rested. Secondly, the data results indicate the problems with a theoretical specification that posits separate cognitive and affective processes. The misspecification is twofold: one common set of inputs leads to cognitive and affective outcomes and there exists no (except for peer influence) set of exogenous variables uniquely related to the affective outcomes. Finally, the simultaneous relationship among achievement, achievement motivation and self-image suggests that a large degree of academic success is a function of the initial level of these attributes when a child begins school.

The data results also point to methodological issues that will confront research on the achievement and motivation process. First, the significant relationships between the affective variables and achievement indicates that analyses of achievement focussing strictly on cognitive input variables are necessarily misspecified. With respect to OLS estimates, the nature of the misspecification results in biases in the estimates. If the excluded affective variables are correlated with the included regressors, the estimates will be biased and inconsistent. If the excluded affective variables are uncorrelated with the included regressors, the estimates are unbiased but the standard errors are inflated (Kmenta: 392-394).

The second methodological issue concerns the estimation of models of achievement and motivation in light of the significant simultaneous relationships between the outcome variables. Since OLS estimates are known to be biased and inconsistent when the right hand side endogenous variables are correlated with the error term, consistent estimation of the equations requires a simultaneous technique such as 2SLS. However, one of the implicit assumptions underlying 2SLS is a well-specified theoretical model, such that unique instruments can be obtained for each endogenous variable. The lack of exogenous variables related to the affective variables means that the instruments for self-image and achievement motivation necessarily contained large amounts of error, reducing the R-square and increasing the MSE when the 2SLS estimation was applied. That is why the OLS estimates were minimum MSE. Consequently, the lack of exogenous variables related to the affective outcomes indicates that the analysis of the achievement and motivation process will be done with an estimation procedure yielding biased and inconsistent estimates.

Thus, the third major inference arising from the application of the causal modelling methodology is: (Basic knowledge inference)

> The process of achievement and achievement motivation cannot be represented by two distinct input processes. Rather, there exists one set of cognitive inputs related to achievement, and achievement, in turn, impacts on achievement motivation and self-image. The three outcome variables are simultaneously related, therefore, models of achievement, to be correctly specified, must contain the affective variables.

In summary, this dissertation adopted the position that the emphasis on experimental and quasi-experimental design in the evaluation methodology literature indicates an insensitivity toward the special nature of evaluation research as part of a larger decision-making process. The implication of this insensitivity is an underestimation of the importance of external validity concerns with respect to evaluation inferences.

In an effort to maintain the emphasis on internal validity but maximize the generalizability of evaluation results, estimation of a social process model has been proposed. Two conditions required validation to support an argument for utilizing the proposed design: (1) that the approach was feasible, i.e., it is no more difficult to implement than conventional research designs, and (2) the informational payoff from the proposed design exceeds that of conventional research designs. Realization of the first condition occurred when the design was successfully applied to an existing evaluation data set. Thus, the data needs of the causal modelling methodology were satisfied by the types of data conventionally gathered.

To fulfill the second condition, three sets of evaluation related inferences were generated which experimental and/or quasi-experimental research designs logically could not produce. The inferences concerned (1) assessment of program failure and theory failure to account for program ineffectiveness, (2) implications of the results for general compensatory education policy, and (3) the state of theoretical knowledge which, in the long run, decision makers require to formulate general policy positions.

With respect to the Head Start program, the specific inferences were: (1) Head Start exhibited program ineffectiveness both for the affective and cognitive goals. However, the failures are of two different types in that the cognitive goal is feasible while the affective goal is not; (2) compensatory education programs will maximize attainment of cognitive and affective goals by focussing exclusively on cognitive inputs. Attainment of the affective goals would be due to the relationship between those variables and achievement; (3) theoretically, the affective and cognitive outcomes are a function of one set of input variables. The simultaneous relationship among achievement, achievement motivation and self-image implies the necessity for inclusion of affective variables in analyses of achievement or specification error will occur.

Finally, in comparison with the specific nature of these inferences, the general recommendation of the Ohio-Westinghouse evaluators concluded:

> . . . we strongly recommend that large scale efforts and substantial resources continue to be devoted to the search for more effective programs, procedures and techniques for remediating the effects of poverty on disadvantaged children (Circirelli: 347).

The irony of this recommendation, as is now clear, is that at least a partial answer was contained in the data already collected by the evaluators. Estimation of a social process model was necessary to even suggest what an appropriate compensatory education strategy would look like. Finally, the basic conclusion of the Ohio-Westinghouse evaluators was that Head Start is an ineffective program. However, the results of this analysis suggest that even though the program as implemented was ineffective, the strategy of increasing cognitive ability by enriching the environment is feasible.

CHAPTER V

SUMMARY AND CONCLUSIONS

The intention of this dissertation was to demonstrate that the use of experimental research designs in the conduct of evaluation leads to inadequate inferences given the informational needs of decision makers. In particular, the focus on internal validity, at the expense of generalizability, does not yield information a decision maker can use to forecast program effectiveness into a blind factor. The problem appears to stem from the inability of evaluation research methodologists to differentiate evaluation, as part of a decision process, from academic research utilizing experimental designs.

It was then suggested that the data needs of decision makers would be better served by a research design based on arguments, developed by Edward Suchman, on the conduct of evaluation. The essence of the design is the specification of the evaluative hypothesis, the theoretical reasoning linking program inputs to intended outcomes and embedding the evaluative hypothesis within a larger model representing the social process the program is aimed at. The effect is a statistical elaboration of the zero order relationship between treatment and outcomes leading to an interpretation of program effectiveness in terms of relevant antecedent and intervening test variables.

The advantages of utilizing the proposed design was argued at both a methodological as well as "practical" level. Methodologically, the explication of a social process model maximized the generalizability of the data results. On a practical level, the design permits inferences about program failure and theory failure, general policy for a given issue area, and the social process of interest. None of these inferences could have been generated by a conventional experimental or quasi-experimental design.

An application of the proposed design was performed by reanalyzing the data from the Ohio-Westinghouse evaluation of Head Start. The evaluative hypothesis attempted to relate the program strategy of enriching the home environment and improving self-image to the outcomes of increased cognitive ability and achievement motivation. The evaluative hypothesis was embedded within a larger social process model of achievement and motivation. Tests of the social process model indicated that (1) the cognitive failure was program failure while the affective failure was theory failure, thus (2) compensatory education programs should be focussed strictly on cognitive outcomes and (3) the cyclical nature of the simultaneous relationships among achievement, motivation, and self-image. This information, quite obviously, was contained in the data generated by the Ohio-Westinghouse evaluators, however, their use of a quasi-experimental design did not permit these inferences to emerge.

The empirical application was not without problems. The dynamic nature of the evaluative inferences cannot be captured by the cross-sectional data obtained in the posttest only design. Thus, inferences of change are based on differences between groups rather than differences over time, a weaker type of evidence. This problem would be overcome by the proposed design requirement of data collection at more than one point in time.

A second problem with the proposed application is that the operationalization does not truly fit the proposed design. In particular, the evaluative hypothesis does not relate program activities to outcomes; the elements of treatment have not been specified. This problem stems from a lack of appropriate variables in the data set. In the application only outcomes are modelled, from intervening to dependent variables, but a complete specification would need to include those variables representing the treatment itself. Thus, a complete evaluation of Head Start would require a more elaborate social process model.

Despite these shortcomings, it would seem that the research design, and the attendant arguments concerning its advantages, has been validated. Evauation relevant information, above and beyond that of the original analysis, was generated by the design. Future applications where a complete model specification and appropriate data collection can occur with the proposed design in mind, should yield more definitive evidence for the positive effects of program evaluation through the testing of a relevant social process model.

APPENDICES

APPENDIX A

CONSTRUCTS

APPENDIX A

CONSTRUCTS

Home Learning Environment: Sum of Scores of the following:

Number of Toys That Child Has Which Could be Used in Playing School
1 0
2 1-2
3 3-5
4 6-9
5 10 or more

Number of Books Child Has To Read
1 0
2 1-2
3 3-5
4 6-9
5 10 or more

How Often Child Reads by Himself at Home
1 seldom or never
2 sometimes
3 often
4 regularly
5 extremely often

How Often Respondant Reads with Child
1 seldom or never
2 sometimes
3 often
4 regularly
5 extremely often

Length of Time Child Reads or Was Read to Day Before Interview
1 not at all
2 up to 15 minutes
3 15-30 minutes
4 30 minutes-1 hour
5 more than 1 hour

Number of Games Child Has
1 none
2 one or two
3 three to five
4 six to nine
5 ten or more

How Often Child Plays with Games
1 seldom or never
2 at least once a week
3 several times a week
4 at least once a day
5 at least several times a day

How Often Respondant Plays Games with Child
1 seldom or never
2 at least once a week
3 several times a week
4 at least once a day
5 at least several times a day

Ways in Which Respondant Is Preparing Child for School
1 nothing
2 help with social skills
3 help with attitudes
4 help with academic skills
5 help with a combination of social skills, attitudes, and academic skills

Achievement: Mean of the nonzero scores of the subunits of The Metropolitan Readiness Tests

Word Meaning
Listening
Matching
Alphabet
Numbers
Copying

Self-image: Sum of the scores of the following items:

Children's Self-Concept Index (CSCI)

CSCI 1 The balloon-child is learning a lot in school.
The flag child isn't learning very much.

The child responds by marking an (X) under the child who is more like himself.

1 Balloon-child
2 Flag-child

CSCI 2 The Parents think the balloon-child is OK.
The parents want the flag-child to do better.

Response Codes are the same as for CSCI.

CSCI 3 Some children hate the balloon-child.
Children like the flag-child.

Response Codes are the same as for CSCI 1.

CSCI 4 The balloon-child likes to please others.
The flag child does not care how others feel.

Response codes are the same as for CSCI 1.

CSCS 5 Children know the balloon-child can't do many things right.
Children know the flag-child can do things right.

Response Codes are the same as for CSCI 1.

CSCI 6 The balloon-child is sad a lot of the time.
The flag-child is happy most of the time.

Response Codes are the same as for CSCI 1.

CSCI 7 Children talk to the balloon-child.
Children do not talk to the flag-chid.

Response Codes are the same as for CSCI 1.

CSCI 8 It's real hard for the balloon-child to learn things.
It's real easy for the flag-child to learn things.

Response Codes are the same as for CSCI 1.

CSCI 9 The balloon-child gives up easily.
The flag-child likes to finish his work.

Response Codes are the same as for CSCI 1.

CSCI 10 Many people like the balloon-child.
Nobody likes the flag-child.

Response Codes are the same as for CSCI 1.

CSCI 11 Children know the balloon-child.
Children do not know the flag-child.

Response Codes are the same as for CSCI 1.

CSCI 12 Things are going to get worse for the balloon-child.
Things are going to get better for the flag-child.

Response Codes are the same as for CSCI 1.

CSCI 13 The balloon-child does not push or scare others.
The flag-child would like to push or scare others.

Response Codes are the same as for CSCI 1.

CSCI 14 The balloon-child feels good inside most of the time.
The flag-child feels bad inside most of the time.

Response Codes are the same as for CSCI 1.

CSCI 15 The balloon-child doesn't have much fun at school.
The flag-child has a lot of fun at school.

Response Codes are the same as for CSCI 1.

CSCI 16 Most people think the balloon-child is good.
Most people think the flag-child is bad.

Response Codes are the same as for CSCI 1.

CSCI 17 The Balloon-child would like to live in some other place.
The flag-child likes where he lives.

Response Codes are the same as for CSCI 1.

CSCI 15 The balloon-child does things better than other children.
The flag-child is not as good at things as other children.

Response Codes are the same as for CSCI 1.

CSCI 19 There are many things the balloon-child does not know.
The flag-child knows many things.

Response Codes are the same as for CSCI 1.

CSCI 20 Next year the balloon-child will do things better.
The flag-child will never be able to do things better.

Response Codes are the same as for CSCI 1.

Description

CSCI 21 The balloon-child hates himself most of the time.
The flag-child likes himself most of the time.

Response Codes are the same as for CSCI 1.

CSCI 22 Most grown-ups don't care about the balloon-child.
Grown-ups like to help the flag-child.

Response Codes are the same as for CSCI 1.

CSCI 23 The balloon-child would like to live with some other family.
The flag-child is happy with his own family.

Response Codes are the same as for CSCI 1.

CSCI 24 The balloon-child is strong enough to do the things he wants to.
The flag-child is too weak to do many things.

Response Codes are the same as for CSCI 1.

CSCI 25 The balloon-child is real clumsy or awkward.
The flag-child is not clumsy or awkward.

Response Codes are the same as for CSCI 1.

CSCI 26 The balloon-child likes to do things by himself.
The flag-child needs to have others help him.

Response Codes are the same as for CSCI 1.

Achievement Motivation: Sum of scores of the following items:

Classroom Behavior Inventory (CBI)

CBI 1. Does he ask questions for information about people, things, etc.?

0 Unable to observe
1 Never
2 Rarely
3 Half of the time
4 Often
5 Almost always

CBI 2. Does he continue working when not under direct supervision?

Response Codes are the same as for CBI 1.

CBI 3. Is he receptive to ideas and suggestions of adults?

Response Codes are the same as for CBI 1.

CBI 4. Does he stay with a task until it is completed?

Response Codes are the same as for CBI 1.

CBI 5. Is he easily distracted by things going on about him? Response Codes are the same as for CBI.

CBI 6. Does he show pride in his work?

Response Codes are the same as for CBI 1.

CBI 7. Does he need to be praised frequently?

Response Codes are the same as for CBI 1.

CBI 8. Does he try to perform his tasks <u>better</u> than others in his class?

Response Codes are the same as for CBI 1.

<u>Description</u>

CBI 9. When faced with a difficult assignment, does he work at it until he gets it?

Response Codes are the same as for CBI 1.

CBI 10. Does he try to do his best on tasks he undertakes?

Response Codes are the same as for CBI 1.

CBI 11. Is he unduly upset or discouraged if he makes a mistake or does not perform well?

Response Codes are the same as for CBI 1.

CBI 12. Is he receptive to the ideas and suggestions of his peers?

Response Codes are the same as for CBI 1.

CBI 13. Does he need attention or approval from adults to sustain him in his work?

Response Codes are the same as for CBI 1.

CBI 14. Does he try to figure things out for himself before asking for help?

Response Codes are the same as for CBI 1.

CBI 15. Does he have a tendency to discontinue activities after exerting a minimum of effort?

Response Codes are the same as for CBI 1.

CBI 16. Does he prefer the new, unfamiliar and novel tasks to the habitual, familiar ones?

Response Codes are the same as for CBI 1.

CBI 17. Does he do better in self-initiated tasks rather than in tasks that are teacher-initiated?

Response Codes are the same as for CBI 1.

CBI 18. Is he careful and methodical in the jobs he undertakes?

Response Codes are the same as for CBI 1.

CBI 19. Does he find it difficult to work or play by himself, thus requiring the company of other children?

Response Codes are the same as for CBI 1.

CBI 20. Does he seem confident that he can do what is expected of him?

Response Codes are the same as for CBI 1.

CBI 21. Does he settle difficulties calmly, on his own, without appealing to others?

Response Codes are the same as for CBI 1.

CBI 22. Does he seem disinterested in the general quality of his work?

Response Codes are the same as for CBI 1.

Parent's Aspiration for Child's Educational Attainment:

Respondant's Aspirations for Child's Level of Education

1 finish grade school
2 attend junior high school
3 finish high school
4 take vocational work in high school
5 take vocational work after high school
6 go to college
7 finish college
8 go to graduate school
9 don't know

Parent's Expectations for Child's Education Attainment:

How Much Education Respondant Thinks Child Will Actually Get.

Response Codes are the same as for educational aspiration.

Parent's Aspiration for Child's Vocational Attainment:

Kind of Job Respondant Would Like to See Child Get After Child Finishes Schooling

1 unskilled worker
2 semi-skilled worker
3 skilled worker
4 owner of little business, clerical sales, or technical
5 administrative personnel, owner of small business, semiprofessional
6 manager or proprietor of medium-sized business, lesser professional
7 executive, proprietor of large concern, major professional
8 don't know

Parent's Expectation for Child's Vocational Attainment:

Kind of Job Respondant Thinks That Child Will Actually Get After Child Finishes Schooling

Response Codes are the same as for vocational aspiration

Number of Siblings:

Number of Children Living at Home.

Code number is response to the question except that 9 or more children is coded as 9.

Socio-economic Status: Sum of scores of the following items:

Mother's Education

1 graduate school
2 completed college
3 some college
4 high school graduate
5 some high school
6 seventh to ninth grade
7 less than seventh grade

Father's education.

Response codes are the same as for mother's education.

Mother's Occupation

1 executive, proprietor of large concern, major professional, etc.
2 manager or proprietor of medium-sized business or lesser professional
3 administrative personnel at large concern, owner of small independent business or semi-professional
4 owner of little business, clerical worker, sales worker, or technician
5 skilled worker
6 semi-skilled worker
7 unskilled worker

Father's Occupation.

Response Codes are the same as for mother's occupation.

Total Family Income

1 less than $2,000
2 $2,000 to $3,999
3 $4,000 to $5,999
4 $6,000 to $7,999
5 $8,000 to $9,990
6 $10,000 to $14,999
7 over $15,000

Parental behavioral response to child's educational and occupational decisions: Sum of scores of the following items:

VABI Behavior Items

What would you do if your child is going to college and needs money to finish his/her education?

1 Weak action
2 Moderate action
3 Strong action

What would you do if your child wants to drop out of school at age 16?

Response same as for question 1.

What would you do if your child graduates from high school and is still uncertain what he/she wants to do?

Response same as for question 1.

What would you do if you wanted your child to go to college, but he/she did not want to go?

Response same as for question 1.

What would you do if your child gets a job that you don't think is good enough for him/her?

Response same as for question 1.

Parental Conservatism:

Sum of scores of the following items (with appropriate recodes):

What They Teach the Kids Is Out of Date

1 strongly agree
2 agree
3 don't know
4 disagree
5 strongly disagree

Most Teachers Do Not Want to be Bothered by Parents Coming to See Them.

Response same as for question 1.

Sports and Games Take Up Too Much Time.

Response Codes are the same as for question 1.

Not Enough Time Is Spent Learning Reading, Writing and Arithmetic.

Response Codes are the same as for question 1.

Teachers Who Are Very Friendly Are Not Able to Control the Children.

Response Codes are the same as for Question 1.

Parental Deemphasis of Education: Coded as for Conservatism, sum of scores of the following items (with appropriate recodes)

People Who Don't Have Much Education Enjoy Life Just as Much as Well Educated People.

In School There Are More Important Things Than Getting Good Grades.

The Teachers Make the Children Doubt and Question Things That They Are Told at Home.

Parental Futility for Education: Coded as for conservatism, some of the scores of the following items (with appropriate recodes)

Most Teachers Probably Like Quiet Children Better Than Active Ones.

I Can Do Very Little to Improve the Schools.

Kids Cut Up So Much That Teachers Can't Teach.

If I Disagree with the Principal There Is Nothing or Very Little I Can Do.

Most Children Have to be Made to Learn.

Parental Gripes about Education: Sum of scores of the following Items (with appropriate recodes)

The Teachers Expect the Children Always to Obey Them.

1 strongly agree
2 agree
3 don't know
4 disagree
5 strongly disagree

The Classrooms Are Overcrowded.

Response Codes are the same as for question 1.

There Are Some Children in the School I Would Not Want My Child To Play With.

Response Codes are the same as for question 1.

Once in a While It Should Be OK for Parents to Keep Their Children Out of School to Help Out at Home.

1 strongly disagree
2 disagree
3 don't know
4 agree
5 strongly agree

Parental Importance of Education for Children: Coded as for Gripes, sum of scores of the following items (with appropriate recodes)

The Best Way to Improve the Schools is to Integrate Them.

Most Teachers Would be Good Examples for My Children.

A Man Can Often Learn More on a Job Than He Can in School.
1 strongly disagree
2 disagree
3 don't know
4 agree
5 strongly agree

Most of the Teachers Are Not Trained As Well As They Should Be
1 strongly disagree
2 disagree
3 don't know
4 agree
5 strongly agree

Sex:
1 male
2 female

Kindergarten Attendance:
1 kindergarten
2 no kindergarten

Race:
1 white
2 black
3 Mexican American
4 Puerto Rican
5 American Indian
6 other

Age:
1 5 years old
2 6 years old
3 7 years old
4 8 years old
5 9 years old
6 10 years old

Parental Vocational Aspiration Scale (Boys): Sum of scores of the following items

VAEI-M1. If you had your wish and your son could have the opportunity, which one job would you like most for your son to be in?
1 farm hand
2 telephone repairman
3 doctor

VAEI-M2. Same question.
1 shoe repairman
2 small business owner
3 politician

VAEI-M3. Same question.
1 factory worker
2 fireman
3 college professor

VAEI-M4. Same question.
1 garbage collector
2 bill collector
3 government official

VAEI-M4. Same question.
1 night watchman
2 social worker
3 clergyman

VAEI-M6. Same question.
1 parking attendant
2 druggist
3 accountant

VAEI-M7. Same question.
1 milkman
2 machinist
3 engineer

VAEI-M8. Same question.
1 bartender
2 bricklayer
3 newspaper editor

VAEI-M9. Same question.
1 restaurant cook
2 bookkeeper
3 author

VAEI-M10. Same question.
1 hospital attendant
2 electrician
3 banker

VAEI-M11. Same question.
1 delivery man
2 carpenter
3 lawyer

VAEI-m12. Same question
1 truck driver
2 policeman
3 airplane pilot

VAEI-M13. Same question.
1 bus driver
2 plumber
3 psychologist

VAEI-M14. Same question.
1 construction worker
2 cashier
3 dentist

VAEI-M15. Same question.
1 taxi driver
2 car salesman
3 scientist

VAEI-M16. Same question.
1 waiter
2 photographer
3 mayor

VAEI-M17. Same question.
1 usher
2 store manager
3 astronaut

VAEI-M18. Same question as for position 330.
1 custodian
2 TV repairman
3 corporation president

VAEI-M19. Same question as for position 330.
1 chauffeur
2 barber
3 college administrator

VAEI-M20. Same question as for position 330.
1 gas station attendant
2 insurance agent
3 judge

Parental Vocational Expectation Scale (Boys):

Same choices as for vocational aspiration except the question is:

What kind of job do you think your son will actually get?

Parental Vocational Aspiration Scale (Girls):

VAEI-F1. If you had your wish and your daughter could have the opportunity, which one job would you like most for your daughter to be in?
1 store clerk
2 beautician
3 nurse

VAEI-F2. Same question.
1 field worker
2 office machine worker
3 singer

VAEI-F3. Same question.
1 elevator operator
2 jeweler
3 scientist

VAEI-F4. Same question.
1 baby sitter
2 dental assistant
3 psychologist

VAEI-F5. Same question.
1 dishwasher
2 court reporter
3 doctor

VAEI-F6. Same question.
1 fountain worker
2 telephone operator
3 musician

VAEI-F7. Same question.
1 ticket taker
2 saleslady
3 magazine editor

VAEI-F8. Same question.
1 cleaning lady
2 cashier
2 actress

VAEI-F9. Same question.
1 grocer checker
2 bookkeeper
3 dancer

VAEI-F10. Same question.
1 metermaid
2 stenographier
2 college professor

VAEI-F11. Same question.
1 maid
2 secretary
3 clothes designer

VAEI-F12. Same question.
1 factory worker
2 advertising agent
2 teacher

VAEI-F13. Same question.
1 clothes presser
2 policewoman
3 artist

VAEI-F14. Same question.
1 restaurant cook
2 photographer
3 school principal

VAEI-F15. Same question.
1 school bus driver
2 census taker
3 airline stewardess

Parental Vocational Expectation Scale (Girls):

Same choices as for vocational aspriation expect the question is:

What kind of job do you think your daughter will actually get?

Parental Attitude Toward Child: Sum of scores of the following items.

How Well Respondant Gets Along with Child
1 poorly
2 not very well
3 fairly well
4 well
5 very well

How Often Child "Gets on Respondant's Nerves"
1 many times a day
2 at least once a day
3 several times a week
4 at least once a week
5 seldom or never

How Often Respondant Becomes Angry with Child
1 many times a day
2 at least once a day
3 several times a week
4 at least once a week
5 seldom or never

How Often Child Does Something for Which He Needs to be Punished
1 many times a day
2 at least once a day
3 several times a week
4 at least once a week
5 seldom or never

Strongest Punishment Respondant Would Give Child
1 severe physical
2 mild physical
3 taking away privileges
4 scholding
5 ignoring child, dirty looks, etc.

Satisfaction That Child Has Given Respondant
1 none
2 very little
3 some
4 considerable
5 very much

<u>Child's Attitude Toward Schools: Sum of the scores of the following items</u>:

CARI 2. Bobby is on his way to school. He gets to school. He opens the door and goes inside. Which one is Bobby's face?

Response Code
1 positive attitude
2 neutral attitue
3 negative attitude

CARI 8. The principal says, "From now on, the school will be open on Saturday morning for children who want to come to read, to play games, or to make things." Karen says, "Oh, Jane, that's a good idea. Let's come over here on Saturday." Jane says, "Well . . . " Which one is Jane's face?

Response Codes are the same as for question 1.

CARI 12. Ann is at school. Her teacher says, "Come to the office with me, Ann." The principal wants to see you. They get to the office. Ann sees the principal. Which one is his face?

Response Codes are the same as for question 1.

CARI 16. The teacher says, "Class, let's put our chairs together in a circle." She says, "Kathy, come put your chair here next to mine today." The class sits down. Kathy is next to her teacher. Which one is kathy's face?

Response Codes are the same as for question 1.

CARI 21. Mark is working at school. Mark's teacher comes over. She looks at Mark's work. Which one is the teacher's face?

Response Codes are the same as for question 1.

CARI 25. Julie is in school. Each child is telling about his favorite food. The teacher calls on Julie. Which one is Julie's face?

Response Codes are the same as for question 1.

CARI 31. Ray is painting at school. He spills some paint on the floor. He doesn't know what to do about it. He sees the teacher coming over. Which one is the teacher's face?

Response Codes are the same as for question 1.

<u>Child's Attitude toward the Home: Some of scores of the following items</u>

CARI 4. Joe is playing at home. He sees his brother and sister coming. They say, "Joe, can we play, too?" Which one is Joe's face?

Response Codes are the same as for school.

CARI 7. Hank takes some of his school work home. He shows it to his mother and father. They look at Hank's work. Which way do they look?

Response Codes are the same as for question 1.

CARI 10. May is on her way here from school. She gets to her house. She stops for a minute in front of her door. Which face is May's?

Response Codes are the same as for question 1.

CARI 13. Jill is at home. Her father comes in. Her father says, "Come here, Jill. I want to talk to you about something." Which face is Jill's?

Response Codes are the same as for question 1.

CARI 20. Molly is at home with her mother and father, her brother and her sister. She starts to leave the room. Mother says, "Stay here, Molly, our whole family is together. Which one is Molly's face?

Response Codes are the same as for question 1.

CARI 24. Betty drops some of her food at the table. She starts to pick it up. She sees her mother looking at her. Which one is her mother's face?

Response Codes are the same as for question 1.

CARI 26. Phil comes home early from school. His mother sees him come in. She says, "Why are you home so soon?" Which one is his mother's face?

Response Codes are the same as for question 1.

CARI 29. Tom and Bill want to go inside to play. Tom says, "Let's go to your house, Bill." Bill says, "No. My folks are always mean." Bill says, "What about your house, Tom?" Which one is Tom's face?

Response Codes are the same as for question 1.

Child's Attitude toward Peers: Sum of scores of the following items

Children's Attitudinal Range Indicator (CARI)

CARI L. Sally is at school. A new girl comes to the class. At recess the new girl comes over to talk to Sally. Which one is Sally's face?

Response Codes are the same as for home.

CARI 6. Jerry is at home. He tells his mother, " I don't know what to do." Jerry's mother says, "Go play with your friends." Which face is Jerry's?

Response Codes are the same as for question 1.

CARI 9. The boys are playing a game. Don says, "I want to play the game with you." The boys say, "O.K., but you must obey all our rules." Which face is Don's?

Response Codes are the same as for question 1.

CARI 15. The boys are on the playground. Each one is showing how strong he is. It is Carl's turn. The boys are watching him. How do the boys look?

Response Codes are the same as for question 1.

CARI 18. Janet is coming up the walk toward school. She sees some children in her class. Some of the kids say, "Hi, Janet." Which one is Janet's face?

Response Codes are the same as for question 1.

CARI 22. Alice has made a picture at school. The teacher tells Alice it is a good picture. Alice shows it to the other children after school. How do their faces look?

Response Codes are the same as for question 1.

CARI 27. John is out on the playground. He sees a group of children playing a game. One of the boys says, "Come and play with us, John." Which one is John's face?

Response Codes are the same as for question 1.

CARI 30. Peggy is with some other girls. They want to have a club. One of the other girls says, "We need more kids in our club." She says, "What do you think, Peggy?" Which one is Peggy's face?

Response Codes are the same as for question 1.

Child's Attitude Toward Society: Sum of scores of the following items

CARI 3. Polly is playing outside. A delivery man drives up in his truck. He comes over to Polly. Which one is the man's face?

Response Codes are the same as for Peers.

CARI 5. Lynn and her friend are walking to the story. They pass a house in their neighborhood. Some people are sitting on the porch. Lynn says, "Oh, they're looking at us!" How do the people look?

Response Codes are the same as for question 1.

CARI 11. A fireman comes to Tom's house. He says, "I want to look around in your house to see that it is safe." Tom's mother talks to the fireman. Which one is her face?

Response Codes are the same as for question 1.

CARI 14. All the neighbors are going to have a meeting at Mike's house. Mike's mother is getting ready. Mike answers the door. Some neighbors come in. Which one is Mike's face?

Response Codes are the same as for question 1.

CARI 17. Steve is outside his house playing ball. Steve sees the neighbor man coming up to his house. The neighbor man stops to talk to Steve. Which face is the neighbor man's?

Response Codes are the same as for question 1.

CARI 28. Sue's mother asks her to go to the store. Sue gets to the store. The store-man sees Sue. Which face is the store-man's?

Response Codes are the same as for question 1.

CARI 32. Rita is playing with Nancy at school. Rita says, "I don't like the neighborhood where I live. Everything is so ugly." She says, "Is your neighborhood nice, Nancy?" Which one is Nancy's face?

Response Codes are the same as for question 1.

Group Assignment:

Type of Treatment
1 Head Start
2 Control

APPENDIX B

SOLVING FOR THE REDUCED FORM

APPENDIX B

SOLVING FOR THE REDUCED FORM

Given the block recursive model:

Block 1 =

Equation 1: $X_4 = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$

Block 2 =

Equation 2: $X_8 = \beta_{10} + \alpha_{11} Y_1 + \alpha_{12} Y_2 + \beta_{11} X_5 + \beta_{12} X_6$

$+ \beta_{13} X_7 + \varepsilon_2$

Equation 3: $Y_1 = \beta_{20} + \alpha_{21} X_8 + \alpha_{22} Y_2 + \beta_{21} X_0 + \beta_{22} X_{10}$

$+ \beta_{23} X_{11} + \varepsilon_3$

Equation 4: $Y_2 = \beta_{30} + \alpha_{31} X_8 + \alpha_{32} Y_1 + \beta_{31} X_{12} + \beta_{32} X_{13}$

$+ \beta_{33} X_{14} + \varepsilon_4$

Within Block 2, the relationships are nonrecursive, thus a reduced form for the block can be obtained by solving for X_8, Y_1 and Y_2 as follows:

Let $A = (\beta_{10} + \beta_{11} X_5 + \beta_{12} X_6 + \beta_{13} Y_7 + \varepsilon_2)$

Let $B = (\beta_{20} + \beta_{21}X_9 + \beta_{22}X_{10} + \beta_{23}X_{11} + \varepsilon_3$

Let $C = (\beta_{30} + \beta_{31}X_{12} + \beta_{32}X_{13} + \beta_{33}X_{14} + \varepsilon_4)$

Then, equation 2, 3, and 4 can be rewritten

Equation 2': $X_8 = \alpha_{11}Y_1 + \alpha_{12}Y_2 + A$

Equation 3': $Y_1 = \alpha_{21}X_8 + \alpha_{22}Y_2 + B$

Equation 4': $Y_2 + \alpha_{31}X_8 + \alpha_{32}Y_2 + C$

which can be re-expressed as

Equation 2'': $X_8 - \alpha_{11}Y_1 - \alpha_{12}Y_2 = A$

Equation 3'': $- \alpha_{21}X_8 + Y_1 - \alpha_{22}Y_2 = B$

Equation 4'': $-\alpha_{31}X_8 - \alpha_{32}Y_1 + Y_2 = C$

The equations can be solved for X_8, Y_1 and Y_2 by a form of Gaussian elimation.

To solve for X_8

1. Multiply (3'') by α_{11} and add to (2'')

$$(-\alpha_{21}\alpha_{11} + 1)X_8 + 0 + (-\alpha_{22}\alpha_{11} - \alpha_{12})Y_2 = A + \alpha_{11}B \qquad (5'')$$

2. Multiply (3'') by α_{32} and add to (4'')

$$(-\alpha_{11}\alpha_{32} - \alpha_{31})X_8 + 0 + (-\alpha_{22}\alpha_{32} + 1)Y_2 = \alpha_{32} B + C \qquad (6'')$$

3. Multiply (6'') by

$$\frac{-(-\alpha_{22}\alpha_{11}-\alpha_{12})}{(-\alpha_{22}\alpha_{32}+1)} \quad \text{and add to (5'')}$$

$$\frac{(-\alpha_{11}\alpha_{32}-\alpha_{31})(\alpha_{22}\alpha_{11}+\alpha_{12})}{(-\alpha_{22}\alpha_{32}+1)} + (-\alpha_{21}\alpha_{11}+1)X_8 =$$

$$\frac{(\alpha_{32}B + C)(\alpha_{22}\alpha_{11}+\alpha_{12})}{(\alpha_{22}\alpha_{32}+1)} + (\alpha_{11}\,B + A)$$

4. Solving for X_8

$$X_8 = \frac{\dfrac{(\alpha_{32}\,B + C)(\alpha_{22}\alpha_{11} + \alpha_{12})}{(-\alpha_{22}\alpha_{32}+1)} + (\alpha_{11}\,B + A)}{\dfrac{(-\alpha_{11}\alpha_{32}-\alpha_{31})(\alpha_{22}\alpha_{11}+\alpha_{12})}{(-\alpha_{22}\alpha_{32}+1)} + (-\alpha_{21}\alpha_{11}+1)}$$

5. Simplifying somewhat yields

$$X_8 = \frac{\dfrac{(\alpha_{32})(\alpha_{22}\alpha_{11} + \alpha_{12}) + (\alpha_{11})}{(-\alpha_{22}\alpha_{32}+1)} \; B + A + \dfrac{(\alpha_{22}\alpha_{11} + \alpha_{12})\,C}{(-\alpha_{22}\alpha_{32}+1)}}{\dfrac{(-\alpha_{11}\alpha_{32} - \alpha_{31})(\alpha_{22}\alpha_{11} + \alpha_{12})}{(-\alpha_{22}\alpha_{32}+1)} + (-\alpha_{21}\alpha_{11}+1)}$$

To solve for Y_1

1. Multiply (4'') by α_{12} and all to (2'')

$$(-\alpha_{31}\alpha_{12} + 1)X_8 + (-\alpha_{32}\alpha_{12} - \alpha_{11})Y_1 + 0 = A + \alpha_{12}C \qquad (5'')$$

2. Multiply (4'') by α_{22} and add to (3'')

$$(-{}_{31}{}_{22} - {}_{21})X_8 + (-{}_{32}{}_{22} + 1)Y_1 + 0 = B + {}_{22}C \qquad (6'')$$

3. Multiply (6'') by

$\dfrac{-(-\alpha_{31}\alpha_{12} + 1)}{(-\alpha_{31}\alpha_{22} - \alpha_{21})}$ and add to (5'')

$$\left[\frac{(-\alpha_{32}\alpha_{22} + 1)(\alpha_{31}\alpha_{12} - 1)}{(-\alpha_{31}\alpha_{22} - \alpha_{21})} + (\alpha_{32}\alpha_{12} - \alpha_{11})\right] Y_1$$

$$= \frac{(\alpha_{22}C + B)(\alpha_{31}\alpha_{12} - 1)}{(-\alpha_{31}\alpha_{22} - \alpha_{21})} + (\alpha_{12}C + A)$$

4. Solving for Y_1

$$Y_1 = \frac{\dfrac{(\alpha_{22}C + B)(\alpha_{31}\alpha_{12} - 1)}{(-\alpha_{31}\alpha_{22} - \alpha_{21}} + (\alpha_{12}\,C + A)}{\dfrac{(-\alpha_{32}\alpha_{22} + 1)(\alpha_{31}\alpha_{12} - 1)}{(-\alpha_{31}\alpha_{22} - \alpha_{21})} + (-\alpha_{32}\alpha_{12} - \alpha_{11})}$$

5. Simplifying somewhat yields

$$Y_1 = \frac{\dfrac{(\alpha_{22})(\alpha_{31}\alpha_{12} - 1) + (\alpha_{12})}{(-\alpha_{31}\alpha_{22} - \alpha_{21})}\; C + A + \dfrac{(\alpha_{31}\alpha_{12} - 1)}{(-\alpha_{31}\alpha_{22} - \alpha_{21})}\; B}{\dfrac{(-\alpha_{32}\alpha_{22} + 1)(\alpha_{31}\alpha_{12} - 1)}{(-\alpha_{31}\alpha_{22} - \alpha_{21})} + (-\alpha_{32}\alpha_{12} - \alpha_{11})}$$

To solve for Y_2

1. Multiply (2") by α_{21} and add to (3")

$$0 + (-\alpha_{11}\alpha_{21} + 1)Y_1 + (-\alpha_{12}\alpha_{21} - \alpha_{22})Y_2 = \alpha_{21} A + B \qquad (5")$$

2. Multiply (2") by α_{31} and add to (4")

$$0 + (-\alpha_{11}\alpha_{31} - \alpha_{32})\, Y_1 + (-\alpha_{12}\alpha_{31} + 1)\, Y_2 = \alpha_{31} A + C \qquad (6")$$

3. Multiply (6") by

$$- \frac{(-\alpha_{11}\alpha_{21} + 1)}{(-\alpha_{11}\alpha_{31} - \alpha_{32})} \quad \text{and add to (5")}$$

$$\frac{(-\alpha_{12}\alpha_{31} + 1)(\alpha_{11}\alpha_{21} - 1)}{(-\alpha_{11}\alpha_{31} - \alpha_{32}} + (-\alpha_{12}\alpha_{21} - \alpha_{22})\, Y_2$$

$$\frac{(\alpha_{31} A + C)(\alpha_{11}\alpha_{21} - 1)}{(-\alpha_{11}\alpha_{31} - \alpha_{32})} + (\alpha_{21} A + B)$$

4. Solving for Y_2

$$\frac{\dfrac{(\alpha_{31} A + C)(\alpha_{11}\alpha_{21} - 1)}{(\alpha_{11}\alpha_{31} - \alpha_{32}} + \alpha_{21} A + B}{\dfrac{(\alpha_{12}\alpha_{31} + 1)(\alpha_{11}\alpha_{21} - 1)}{(-\alpha_{11}\alpha_{31} - \alpha_{32})} + (\alpha_{12}\alpha_{21} - \alpha_{22})}$$

5. Simplifying somewhat yields

$$Y_2 = \frac{\dfrac{(\alpha_{31})(\alpha_{11}\alpha_{21} - 1)}{(-\alpha_{11}\alpha_{31} - \alpha_{32})} + (\alpha_{21}) \quad A + B + \dfrac{(\alpha_{11}\alpha_{21} - 1)}{(-\alpha_{11}\alpha_{31} - \alpha_{32})} \; C}{\begin{array}{l}\underline{(-\alpha_{12}\alpha_{31} + 1)(\alpha_{11}\alpha_{21} - 1)} + (\alpha_{12}\alpha_{21} - \alpha_{22}) \\ \underline{(-\alpha_{12}\alpha_{31} + 1)(\alpha_{11}\alpha_{21} - 1)} \\ \quad (-\alpha_{11}\alpha_{31} - \alpha_{32})\end{array}}$$

The reduced form of the structural equation model is

Equation (1R): $X_4 = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon_1$

Equation (2R): $$X_8 = \frac{\dfrac{(\alpha_{32})(\alpha_{22}\alpha_{11} + \alpha_{12})}{(-\alpha_{22}\alpha_{32} + 1)} + (\alpha_{11}) \; B + A + \dfrac{(\alpha_{22}\alpha_{11} + \alpha_{12})}{(-\alpha_{22}\alpha_{32} + 1)} \; C}{\dfrac{(-\alpha_{11}\alpha_{32} - \alpha_{31})(\alpha_{22}\alpha_{11} + \alpha_{12})}{(-\alpha_{22}\alpha_{32} + 1)} + (-\alpha_{21}\alpha_{11} + 1)}$$

Equation (3R): $$Y_1 = \frac{\dfrac{(\alpha_{22})(\alpha_{31}\alpha_{12} - 1)}{(-\alpha_{31}\alpha_{22} - \alpha_{21})} + (\alpha_{12}) \; C + A + \dfrac{(\alpha_{31}\alpha_{12} - 1)}{(-\alpha_{31}\alpha_{22} - \alpha_{21})} \; B}{\dfrac{(-\alpha_{32}\alpha_{22} + 1)(\alpha_{31}\alpha_{12} - 1)}{(-\alpha_{31}\alpha_{22} - \alpha_{21})} + (-\alpha_{32}\alpha_{12} - \alpha_{11})}$$

Equation (4R): $$Y_2 = \frac{\dfrac{(\alpha_{31})(\alpha_{11}\alpha_{21} - 1)}{(-\alpha_{11}\alpha_{31} - \alpha_{32})} + (\alpha_{21}) \quad A + B + \dfrac{(\alpha_{11}\alpha_{21} - 1)}{(-\alpha_{11}\alpha_{31} - \alpha_{32})} \; C}{\dfrac{(-\alpha_{12}\alpha_{31} + 1)(\alpha_{11}\alpha_{21} - 1)}{(-\alpha_{11}\alpha_{31} - \alpha_{32})} + (\alpha_{12}\alpha_{21} - \alpha_{22})}$$

Where $A = (\beta_{10} + \beta_{11} X_5 + \beta_{12} X_6 + \beta_{13} X_7 + \varepsilon_2)$

$B = (\beta_{20} + \beta_{21} X_9 + \beta_{22} X_{10} + \beta_{23} X_{11} + \varepsilon_3)$

$C = (\beta_{30} + \beta_{31} X_{12} + \beta_{32} X_{13} + \beta_{33} X_{14} + \varepsilon_4)$

REFERENCES

REFERENCES

Achen, Christopher H. "The Statistical Analysis of Quasi-Experiments." Unpublished manuscript, 1980.

Anastasi, A. "Intelligence and Family Size." Psychological Bulletin 53 (1956): 187-209.

Andersen, J. G. "Causal Models in Educational Research: Nonrecursive models." American Educational Research Journal 15 (1978): 81-97.

Barnow, B. S. "The Effects of Head Start with Socio-Economic Status on Cognitive Development of Disadvantaged Children." Ph.D. dissertation, University of Wisconsin, Madison, 1973.

Bentler, Peter M., and Woodward, Arthur J. "A Head Start Reevaluation: Positive Effects Are Not Yet Demonstrable." Evaluation Quarterly 2 (1978): 493-550.

Bereiter, Carl, and Englemann, Sigfried. Teaching Disadvantaged Children in the Preschool. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1966.

Bigge, Morris L., and Hunt, Maurice P. Psychologyical Foundations of Education. 3rd ed. New York: Harper and Row, 1980.

Campbell, Donald T. "Reforms as Experiments." Urban Affairs Quarterly 7 (1971): 133-171.

Campbell, Donald T., and Stanley, Julian C. Experimental and Quasi-Experimental Designs for Research. Chicago: Rand McNally College Publishing Co., 1963.

Cicirelli, Victor. "The Impact of Head Start: Executive Summary." In Readings in Evaluation Research, pp. 343-347. 2nd ed. Edited by Francis Caro. New York: Russell Sage Foundation, 1977.

Cicirelli, Victor G. "The Relationship of Sibling Structure to Intellectual Abilities and Achievement." Review of Educational Research (1978): 365-379.

Cohen, David K. "Politics and Research: Evaluation of Social Action Programs in Education." In Evaluating Action Programs, pp. 137-165. Edited by Carol H. Weiss. Boston: Allyn and Bacon, Inc., 1976.

Cook, Thomas D., and Campbell, Donald T. Quasi Experimentation: Design and Analysis Issues for Field Settings. Chicago: Rand McNally College Publishing Co., 1979.

Datta, Lois-ellen, et al. The Effects of the Head Start Classroom Experience on Some Aspects of Child Development: A Summary Report of National Evaluations, 1966-1969. Washington, D.C.: U. S. Department of Health, Education and Welfare, 1973.

Duncan, Otis Dudley. "Occupational Mobility in the United States." American Sociological Review 30 (1965): 491-498.

Gujarati, Damodar. "Use of Dummy Variables in Testing for Equality Between Sets of Coefficients in Linear Regressions: A Generalization." The American Statistician 24 (1970): 18-21.

Iverson, Barbara K., and Walberg, Herbert J. "Home Environment and Learning: A Quantitative Synthesis." Paper presented to the American Educational Research Association Annual Meeting, Toronto, Canada, 1978.

Jenson, Arthur. "How Much Can We Boost IQ and Scholastic Achievement." Harvard Educational Review 39 (1969): 1-123.

Kerlinger, Fred. Foundations of Behavioral Research. 2nd ed. Chicago: Holt, Rinehart and Winston, 1973.

Kmenta, Jan. Elements of Econometrics. New York: MacMillan, 1971.

Lazar, Irving. Lasting Effects After Preschool. Washington, D.C.: U. S. Department of Health, Education and Welfare, 1978.

Magidson, Jay. "Toward a Causal Model Approach for Adjusting for Preexisting Differences in the Nonequivalent Control Group Situation: A General Alternative to ANCOVA." Evaluation Quarterly 1 (1977): 399-420.

Miller, Harry L. Social Foundations of Education: An Urban Focus. 3rd ed. Chicago: Holt, Rinehart and Winston, 1978.

Rao, Potluri, and Miller, Roger. Applied Econometrics. Belmont, Calif.: Wadsworth, 1971.

Salmon, Wesley C. The Foundations of Scientific Inference. Pittsburgh: University of Pittsburgh Press, 1967.

Scheirer, Mary Ann, and Kraut, Robert E. "Increasing Educational Achievement via Self-Concept Change." Review of Educational Research 49 (1974): 131-150.

Shapiro, Jonathan. "Note on Anderson's 'Causal Models in Educational Research: Nonrecursive Models.'" American Educational Research Journal 16 (1979): 347-350.

Smith, Marshall, and Bissell, Joan. "Report Analysis: The Impact of Head Start." Harvard Educational Review 40 (1970): 51-104.

Suchman, Edward. Evaluative Research. New York: Russell Sage Foundation, 1967.

Suchman, Edward. "Evaluating Educational Programs." In Readings in Evaluation Research, pp. 48-53. 2nd ed. Edited by Francis Caro. New York: Russell Sage Foundation, 1977.

Uguroglu, Margaret, and Walberg, Herbert J. "Motivation and Achievement: A Quantitative Synthesis." Paper delivered at the Annual Meeting of the American Educational Research Association, Toronto, Canada, 1978.

Weiner, Bernard. Achievement Motivation and Attribution Theory. Morristown, N.J.: General Learning Press, 1974.

Weiss, Carol H. Evaluation Research. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1972.

Weiss, Carol H. "Utilization of Evaluations Toward Comparative Study." In Evaluating Action Programs, pp. 318-327. Edited by Carol Weiss. Boston: Allyn and Bacon, Inc., 1976.