DOCUMENT RESUME

ED 079 848 EA 005 323

AUTHOR Marcus, Alfred C.; And Others
TITLE An Analytical Review of Longitudinal and Related Studies as They Apply to the Educational Process. Methodological Foundations for the Study of School Effects, Volume III.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY National Center for Educational Statistics (DHEW/OE), Washington, D.C.
PUB DATE 72
NOTE 291p.

EDRS PRICE MF-$0.65 HC-$9.87
DESCRIPTORS Child Development; Data Collection; Educational Planning; Educational Policy; *Educational Research; *Evaluation Methods; Followup Studies; Higher Education; *Longitudinal Studies; Measurement Techniques; Program Evaluation; Research Design; *Research Methodology; Research Reviews (Publications); School Environment; Student College Relationship; *Student Development; Surveys

ABSTRACT

This document is the third volume in a 5-part series that reports the results of a project undertaken to critically review and analyze major longitudinal studies of child and student development. These studies were conducted to discover the variables, techniques, methodologies, and problems pertinent to evaluative studies of the effects of schools and colleges on the growth and development of children and young adults. It was anticipated that study results would provide guidelines for the future research needed to enhance educational program planning, implementation, and evaluation. This volume is primarily devoted to assessing the application of survey methodology to educational research and to the studies reviewed in the project. The authors first discuss the purpose or orientation of the study and the validity of the inferences drawn from the data, and they describe the logic of the survey research and the major data collection technique used in the studies. The success with which the researcher was able to demonstrate or infer causal relationships and the degree to which the observed relationships were explicated are then assessed. The discussion moves on to treat of impact analyses, including the problem of measuring change and some of the problems common to survey research. Conclusions and recommendations for future research methodology conclude the presentation. Related documents are EA 005 321-322 and EA 005 324-325. (Author/DN)

AN ANALYTICAL REVIEW OF LONGITUDINAL AND RELATED STUDIES AS THEY APPLY TO THE EDUCATIONAL PROCESS

VOLUME III

METHODOLOGICAL FOUNDATIONS FOR THE STUDY OF SCHOOL EFFECTS

by

Alfred C. Marcus
with
J. Ward Keesling, Clare Rose and James W. Trent

1972

AN ANALYTICAL REVIEW OF LONGITUDINAL AND RELATED STUDIES AS THEY APPLY TO THE EDUCATIONAL PROCESS

VOLUME III

METHODOLOGICAL FOUNDATIONS FOR THE STUDY OF SCHOOL EFFECTS

by

Alfred C. Marcus

with

J. Ward Keesling,
Clare Rose
and
James W. Trent

1972

PREFACE

The present monograph is the third of a series of four volumes designed to present the results of the Analytical Review of Longitudinal Studies, sponsored by the National Center for Educational Statistics of the U.S. Office of Education. The project is designed to analyze selected major longitudinal studies in order to discover variables, techniques, methodologies and problems pertinent to evaluative studies of the effects of schools and colleges on the growth and development of children and young adults.

Volume I contains the theoretical framework of the project and its applicability to research and development together with highlights of substantive and methodological issues raised by the studies selected for analysis. Volume II contains the dynamics of the development of the abstracting process underlying the project, the elements of the process, the typology of the variables included in the research, an overview of major issues suggested by the research reviewed as well as by the review process and the abstracts themselves.

The present volume emphasizes the research methodology, techniques and instrumentation used in the research in reference to ideal research norms. The synthesis and implications of the findings, including matrices of the findings and variables derived from the studies comprise Volume IV.

The enormous and difficult task of sifting and resifting through seemingly countless documents, forms, and data has been made possible through the exceptional cooperation of involved individuals. Those researchers whose studies are under review have been particularly helpful. Judd Adams contributed to the preliminary conceptualization of the present volume and Blair Sullivan reviewed the statistical techniques of the studies considered. Finally, appreciation is extended to Mrs. Betty Martin and Mrs. Mary McCullum of the Student Service Center of West Los Angeles for their professional expertise in typing this volume.

James W. Trent
Principal Investigator

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Research methodology refers to the specific procedures or techniques that are used by an investigator to collect and analyze or manipulate empirical observations. These techniques represent the tools or methods which enable the researcher to formulate, test and refine statements about "reality." Thus, research methodology determines both the constraints and the possibilities of empirical investigations, or what might be termed the limitations, generalizability and applicability of research. Consequently, before confidence can be placed in the findings of a research study a critical evaluation of that study must be undertaken. Such an evaluation necessitates rigorous assessment of its methodology.

Many methodological criteria, however, cannot be derived through logical deduction or validated by mathematical proofs. Moreover, current methodological standards do not represent absolute or revealed truth, but rest on a consensus of opinion of what should be done given the purpose or goal of the study. As a result, the body of opinions, beliefs, and recommended procedures that fall under the rubric of "research methodology" represent a living or dynamic system of ideas that will change or be modified as new opinions and perspectives are introduced or when new methodological practices are discovered.

The problems of methodology transcend those found in any one discipline. Groups of disciplines may share common methodological problems and some problems of methodology are common

to all scientific research. Consequently, it is important to recognize that the methodological issues discussed in this volume are not necessarily restricted to education or uniquely characteristic of educational research in general, or to survey research in particular.

This volume is primarily devoted to assessing the application of survey methodology to educational research, particularly its application to the studies reviewed by the Analytical Review project, and is organized as follows:

Two major points of reference have guided the attempt to evaluate the success with which survey research has been employed in the studies under review: (1) the purpose or orientation of the study, and (2) the validity of the inferences drawn from the data. These two principal dimensions of the review, which are discussed in Chapter I, provide the frame of reference for Chapters II, III, and IV, and the framework in which the Analytical Review studies are evaluated.

Chapter II describes the logic of survey research, the major data collection technique used in the studies. The limitations as well as the potential advantages of survey methodology are included in this chapter.

Since most of the studies under review incorporated a causal research orientation, the success with which the researcher was able to demonstrate or infer causal relationships and the degree to which the observed relationships were explicated are assessed in Chapter III. Chapter IV treats impact analyses, including the problem of measuring change. Problems common to survey

research are discussed in Chapter V. Conclusions and recommendations for future research methodology are presented in Chapter VI.

CHAPTER I

DIMENSIONS OF EVALUATION

Systematic evaluation of empirical research must be based upon two essential elements: the purpose or objectives of the study, and the extent to which the objectives were attained. In other words, what was the researcher attempting to accomplish and how valid were his conclusions and interpretations? These two principal dimensions of methodological evaluation form the major focus of this volume.

Although each particular investigation has its own unique justification or purpose, the major goals of empirical research can be grouped into a number of analytically distinct categories. Thus, research can be conducted in order to:

1. Explore the dynamics of phenomena for the purpose of gathering initial or preliminary information about the phenomena.
2. Describe the characteristics of phenomena.
3. Test hypotheses concerning the causes of phenomena.
4. Elaborate upon or explicate the dynamics of causal relationships. In other words, the purpose of research can be explorative, descriptive, causal or explicative. A brief description of each of these orientations is presented below.

Exploratory Analysis

Exploratory or "pilot" studies are usually undertaken in order to gain the information necessary to formulate a research

problem more precisely or for developing specific hypotheses. However, an exploratory study may have other functions such as gathering preliminary information about a target population of particular interest, clarifying concepts and establishing priorities for future research.

Many large-scale research projects have associated with them a number of exploratory studies out of which ' ' .er project has evolved. In fact, most elaborate studies should begin with exploratory studies since they make a unique contribution to the more sophisticated designs used in comprehensive research studies.

An exploratory investigation can take many forms. Three basic methods identified by Selltiz, Jahoda, Deutsch and Cook (1959, pp. 53-63) are:

1. Surveying the relevant literature.
2. Interviewing respondents familiar with the phenomenon or situation.
3. Investigating "insight stimulating" examples such as deviant or typical cases.

Exploratory studies are especially important in the field of education where a generally accepted theory of human development is lacking and where learning theories are either too narrow in scope or too global to provide definitive bases for empirical investigation or hypothesis testing. It is important to remember, however, that while exploratory studies can identify suggestive relationships which may merit further investigation, they do not verify the existence of these relationships or test particular hypotheses.

Descriptive Analysis

The focus or goal of a descriptive study is the precise measurement of one or more phenomena in a population. For example, the researcher may administer an achievement test to a defined population of students in order to describe the distribution of scores. Thus, the principal concern in a descriptive survey is to accurately describe the distribution of a particular phenomenon rather than to identify specific factors responsible for the shape of that distribution.

The research questions asked in a descriptive study presuppose a considerable amount of prior knowledge about the phenomenon or problem. The investigator must not only be able to conceptualize and define what he wants to measure but he must select appropriate instruments for valid and reliable measurement so that the necessary data may be collected to accurately describe the phenomenon of interest.

Causal Analysis

When the investigator is interested in determining if one or more independent variables leads to or produces a specific outcome or value in a dependent variable, he pursues a causal investigation. The "common sense" notion of causation suggests that a single event (the cause) always leads to another event or result (the effect). Relationships between variables, however, are rarely this straightforward. Consequently, the doctrine of multiple causation has been adopted by most researchers in every field of scientific inquiry. Briefly stated, the doctrine of multiple causation proposes that a multiplicity of determining

or contributory conditions act separately and in conjunction with one another to influence the probability that a subsequent event will occur.

No scientific method or procedure, however, will permit the investigator to demonstrate with absolute certainty that one variable has caused another. It is impossible, for example, to "prove" that a particular school characteristic, such as the average size of the classroom in freshman English, directly determines another phenomenon, such as verbal ability. Although the existence of a particular relationship cannot be demonstrated with absolute certainty, inferences can be made concerning the likelihood of a particular relationship. The evidence used to infer causality is based on three requirements or propositions that necessarily characterize a causal relationship:[1] co-variation; proper time sequence of variables; nonspurious relationships.

Co-variation. If variable X is a cause of variable Y, then there should be some form of statistical association between the two variables. This does not mean that X and Y need to be related to one another in a linear fashion; a curvilinear relationship can also suggest causation. Furthermore, the lack of statistical

---

[1] H. Hyman (1955) suggests a fourth criterion; the existence of intervening variables which link the independent and dependent variables together. Knowing the processes through which X influences Y is certainly desirable, nonetheless, this criterion goes beyond the minimum requirements for demonstrating a causal relationship. Turning the ignition key, for example, will cause the automobile to start even if we cannot describe the intervening processes.

co-variation does not necessarily indicate that two variables are not related to one another; the apparent nonrelationship could be due to the influence of a third variable which is having the effect of obscuring or suppressing a relationship between the causal or independent and the dependent variable. The basic proposition remains, however; if X is a cause of Y, then knowledge of X should help in predicting Y. This can only occur when a statistical association exists between X and Y.

Proper time sequence of variables. One event cannot cause another if it occurs after the other event. The occurrence of a causal factor may precede or be simultaneous with the occurrence of an event but it cannot be posterior in temporal sequence. Of course, it is possible for each variable in a relationship to be considered both the cause and the consequence of the other. Thistlethwaite (1965) for example, purports to be investigating factors in college environments which motivate a student to seek graduate training. Since Thistlethwaite is defining environmental factors in terms of the subjective judgments of the students, the question of direction of causality becomes crucial; does the type of environment described by students affect their values or do their values affect their perception of that environment?

Relationships of this type are called symmetrical and are frequently observed in social research. When symmetrical causal relationships are identified and when the analyst believes reciprocal interaction is not an adequate or valid representation of the relationship in question, evidence should be provided which suggests the most likely or primary direction of causation.

Nonspurious relationships. Even when the data indicate that a particular variable is statistically associated with the criterion or dependent variable and is antecedent in temporal sequence, it still may be incorrect to conclude that a causal relationship exists. That is, there still exists the possibility that another antecedent causal factor accounts for the variation in both the independent and dependent variables. When this situation occurs, the two variables are said to be descriptively related but not causally related, that is, the apparent causal relationship is judged to be spurious. As Hyman (1955) states:

> Spuriousness applies to situations where a variable other than the apparent explanation was found to have produced the observed effect . . . (p. 256).

For example, the observed positive relationship between per-pupil expenditure and mathematical achievement may be judged spurious because students who perform well in mathematics are also likely to come from middle or high socio-economic status families which tend to reward and encourage academic achievement. Thus, it could be that the value orientation of the student's family, rather than the effects of school expenditures, is actually the factor producing high mathematical achievement in the above relationship.

Explicative Analysis

Although a causal analysis may indicate that a particular relationship probably exists, it does not necessarily identify the reasons why a particular variable produces a certain effect. Thus, the researcher may determine that small classrooms

contribute to higher verbal achievement scores, yet be uncertain as to the underlying processes responsible for this relationship. Explicative studies attempt to answer this type of research question. The key word in explicative analysis is elaboration. Exploratory studies are designed to test specific hypotheses and thereby to elaborate the relationships between variables, including the processes, precipitating factors or events, and the structural contexts which mediate the relationships.

Each of the four research orientations described above has a distinctive set of characteristics that result in a unique contribution to scientific investigation. Exploratory studies are highly flexible because their major contribution is the discovery and identification of new insights that will direct future research. Descriptive and causal analyses on the other hand, are more concerned with accuracy and reducing bias than with flexibility, and thus employ more structured and controlled data collecting techniques.

In addition, the purposes of these research orientations are dynamically related to one another. Exploratory and descriptive studies, for example, frequently raise questions which motivate causal investigations, and explicating a causal relationship, of course, presupposes that a causal relationship has been identified. In short, research should be programmatic beginning with the identification of a suggestive relationship and culminating with a systematic appraisal of the processes and intervening factors responsible for or influencing that relationship.

The Validity of Scientific Research

The concept of validity refers to the degree to which a particular measurement or observation is really measuring what was intended to be measured. In other words, was the researcher successful in measuring or recording the phenomenon of interest, or are the observations invalid, that is, unrepresentative of the intended situation or phenomenon?

The question of validity in reference to the inferences and conclusions contained in the studies under review is of central importance to the objectives of this report since this question necessarily challenges the researcher's definition and analysis of "reality" and thus determines in large part the ultimate value or contribution of a particular investigation.

Validity can be assessed in reference to a particular measuring instrument (e.g., does the Strong Vocational Interest Blank really measure vocational interest?) or to a particular research finding (e.g., do small classrooms really contribute to higher verbal achievement?). In either case, however, the issue of validity considers whether unintended alternative phenomena are inadvertently being measured or observed in sufficient quantity to render problematic the interpretation of the results in terms of the major intention or goal of the particular measurement.

When the validity of an observed relationship is challenged, the real issue being raised is whether the relationship actually demonstrates a true effect or is spurious. Thus, when examining the validity of relationships, invalidity and spuriousness become synonymous.

An observed relationship can be judged spurious or invalid for one of two reasons. First, it may be determined that some temporally antecedent characteristic or experience of the respondent, which is intrinsically part of the process that accounts for the relationship in question, explains the variation in both the independent and dependent variables. Thus, in a previous example it was noted that a relationship between per-pupil expenditure and mathematical achievement may be spurious due to the effects of the value orientations of the student's family. In this case, the effects of the familial value system on mathematical achievement is temporally antecedent to the effects of per-pupil expenditure and can be conceptualized as being intrinsically related to the process contributing to high mathematical achievement.

Spuriousness can also occur because some defect in the research design or some inherent limitation in the method used to collect the data permits a factor or variable that is unrelated or extraneous to the underlying process responsible for a relationship to artifactually produce the relationship in question. Thus, the observed relationship between per-pupil expenditure and mathematical achievement may be invalid because the sample of students in high per-pupil expenditure schools was unrepresentative of the population of students in these institutions in terms of their current mathematical achievement. That is, the factor or condition that may be responsible for the relationship between per-pupil expenditure and mathematical achievement could be the unrepresentativeness of the sample of

students in high per-pupil expenditure schools. If the sample was biased in favor of high mathematical achievement students such that they were overrepresented in the sample, then the finding relating per-pupil expenditure to mathematical achievement is clearly problematic. The results could be entirely artifactual due to this sampling bias rather than to the true effects of school expenditure on mathematical achievement.

In order to avoid possible confusion between extrinsic and intrinsic forms of spuriousness or invalidity, the term "testing for spuriousness" will be used in this report to refer to the investigation of intrinsic spuriousness and the term invalidity (or invalid) will refer to extrinsic spuriousness. This distinction is important to bear in mind. The researcher can be justifiably criticized if an observed relationship is unnecessarily rendered problematic because of some defect in the design of the study. However, when the relationship is judged spurious because of some antecedent variable or condition that is intrinsically part of the process responsible for that relationship, then the criticism is inappropriate. Indeed, the major goal of a study may very well be to ferret out the effects of these intrinsically related factors in order to test a suspected causal relationship. Thus, the researcher who systematically tests for "intrinsic spuriousness" should be commended since this is a crucial test for demonstrating a causal relationship.

CHAPTER II

THE LOGIC OF SURVEY RESEARCH

The term survey research often connotes public opinion polls such as those designed to collect information about political beliefs or consumer attitudes. However, the domain of survey research as an instrument of scientific investigation is neither limited to opinions or attitudes nor to describing the frequency with which a particular phenomenon or characteristic is manifest in a specific target population. To the contrary, survey methodology can be an effective tool for collecting data to analyze cause and effect relationships that concern the dynamics of human behavior.

For a number of reasons, the sample survey has become the most ubiquitous technique for collecting data in educational research. One reason is that the structured questionnaire, which has almost become synonymous with survey research, is a relatively economical method of collecting data from a sample or population of respondents. Another reason for its appeal to educational researchers is the ease with which the mass survey can be administered to school populations since students are not only literate but are used to taking tests and answering questions.

Although the popularity of survey research has been increasing over the years, many educational researchers remain unfamiliar with the underlying logic of this methodology. After reviewing the literature in 1966, Trow (1967) concluded that:

> While the practical and methodological problems faced by researchers in other areas have stimulated critical thought and increased sophistication in the use of survey research, within education much survey research has been carried on in substantial innocence of these developments . . . which substantially reduce both the practical and the scientific value of the research thus conducted. . . .
>
> There is little consideration of the logic of survey research--of the larger problems of design, of analytical strategies, or of the interdependence of the elements in the research. Small wonder, then, that these latter concerns are so rarely in evidence in the published research. . . (pp. 319-320).

Even though many of the studies reviewed in this monograph were not available to Trow at the time of his evaluation, and while a greater sophistication in the use of survey methodology is discernible in the interim, the same criticisms posited half a decade ago still apply today. Therefore, the logic of survey methodology will be briefly discussed, and hopefully this will clarify for the reader some of the basic concepts, terminologies, and strategies of survey research.

Descriptive, Causal, and Explicative Analyses

The four major research orientations outlined in Chapter I can be pursued through survey methodology. That is, surveys can be exploratory, descriptive, causal, and explicative. Since the studies reviewed were major research projects, none were of the exploratory type. As a result, our discussion will be confined to descriptive, causal, and explicative analyses.

The investigator undertaking a descriptive analysis is primarily concerned with estimating the parameters or distribution of one or more characteristics of a phenomenon in a

specified population. An example of a descriptive survey is Coleman's (1966) study, which was designed under the auspices of the U.S. Office of Education to determine the extent of educational opportunity in this country.

In causal and explicative analysis, the researcher is interested in identifying and elaborating upon variables which influence or determine the shape of the distribution. Thus, according to Bachman (1969), Youth in Transition focuses on:

> . . . some major changes in adolescent boys during the high school years . . . (and) . . . is particularly concerned with the ways these changes are affected by aspects of the immediate social environment (p. 1).

The differences in purpose between descriptive, causal and explicative survey analysis pertain to differences in the conceptualization of the study and the type and complexity of data manipulation. While this distinction can be artificial since one study can perform all three forms of analysis, it is important to treat them separately in order to emphasize the most salient methodological issues that relate to the logic of survey research.

Conceptualization of Surveys for Descriptive Analysis

Descriptive analysis usually precedes both causal and explicative analysis since theorists often cannot anticipate which factors influence the distribution of a particular variable until they know the shape of that distribution. Consequently, it is not unusual for analysts involved in descriptive work to find themselves investigating phenomena which

have not been adequately conceptualized.

Conceptualization of phenomena to be investigated is never an easy task. Moreover, the problem is exacerbated when there is a paucity of supporting research which has previously attempted to identify and organize the salient elements and dimensions of the phenomena. It is imperative, therefore, that efforts to conceptualize the phenomena be systematic. Several steps which may help the analyst develop a heuristic conceptual scheme are enumerated below.

1. The conceptual model should begin with a theoretical definition of the research problem. This is necessary for two reasons. First, a clear definition of the purpose of the investigation will separate the phenomena under investigation from other related interests. Second, it will form the major proposition in a deductive system that provides the analyst with the constructs, concepts, and operational definitions necessary for systematic and integrated analysis.

   A properly formulated theoretical definition should contain the following properties:

   a. Delineation of the domain of the phenomena to be described. It is important in this context to identify whether the phenomena is affective, cognitive, or behavioral since this will have a direct bearing on the nature of the conceptual system and the operational definitions.

   b. Specification of the location of the phenomenon or target population in time and space.

   c. Specification of the major unit of analysis, that is, will the analysis focus upon individuals or groups.

2. The conceptual model should identify the major dimensions or constructs of the most salient aspects of the phenomena identified in the theoretical definition. Coleman, for example, identified five major dimensions of educational equality: quality of the physical resources; characteristics of teachers; characteristics of the student body; academic practices; and the variety of curricular offerings.
3. The conceptual scheme should specify all of the major concepts or important elements within each of these primary dimensions. In Coleman's study the major characteristics of the student body included socio-economic background, the educational level of their parents, self-concepts and academic aspirations.
4. A sophisticated conceptual scheme should also describe the concepts in terms of specific criteria relevant to educational theory, on which the phenomena may be stratified for comparative purposes. A global description of the phenomena is open to multiple interpretation and thus too abstract for rigorous, systematic investigation. Consequently, the analyst may decide to enrich the analysis by breaking the data down on the basis of some relevant criteria. Thus, Coleman not only presents estimates of the different measures of educational opportunity for the nation as a whole, but also for different levels of socio-economic class, for major geographical regions of the United States, and for metropolitan and nonmetropolitan areas.

Once the analyst has conceptualized the phenomena, he must then decide how these concepts are to be operationalized or defined in such a way that they can be empirically measured. This is a critical step in scientific research since the validity of a study is necessarily dependent upon the researcher's ability to devise a set of operational definitions that will permit him to construct appropriate measuring instruments designed to obtain information that adequately represents the concept under investigation. In fact, several alternative operational definitions for each concept should be generated and, if possible, multiple measures should be included in the questionnaire so that internal checks of their validity can be performed.

The extent to which the specific operational definitions adequately represent or reflect the phenomena identified in the theoretical definition will ultimately determine whether the researcher is actually describing the phenomena or problems of interest. In other words, the theoretical focus of the study can change if the analyst is not really measuring the selected phenomena. Thus, the appropriateness of the specific operational definitions used in a study determines whether the researcher is measuring the constructs and concepts that were articulated in the conceptual model.

For example, a researcher may state that the purpose of his investigation is to estimate the level or distribution of mathematical achievement in a particular school district, but actually collect data on mathematical aptitude rather than

mathematical achievement. Thus, the researcher's conceptualization of the problem is no longer appropriate since it does not support or correspond to the data that was collected. Consequently, while a strict deductive system is desired, it is important to consider the extent to which the conceptual model is also inductive due to the possible lack of correspondence between the operational definitions and the theoretical concepts and constructs contained in the conceptual system. This sequence is illustrated in Figure 1. A diagram of a partial conceptualization of Coleman's study is provided in Figure 2.

Data Analysis Strategies for Descriptive Analysis

Data analysis procedures in descriptive surveys are relatively simple and straightforward. Tabulating and summarizing the data obtained from the sample and various subgroups in statistical or quantitative terms is of primary importance in descriptive work. Usually this information is presented in the form of frequency or percentile distributions or average scores. Coleman, for example, summarized the estimated distribution of majority and minority students attending college as illustrated in Table 1.

A descriptive report will also stratify the phenomena of interest on selected criteria for comparative purposes. Thus, as illustrated in Table 1, Coleman not only provides estimates of college enrollment for the nation as a whole but for the major geographic regions as well. Breaking the data down on

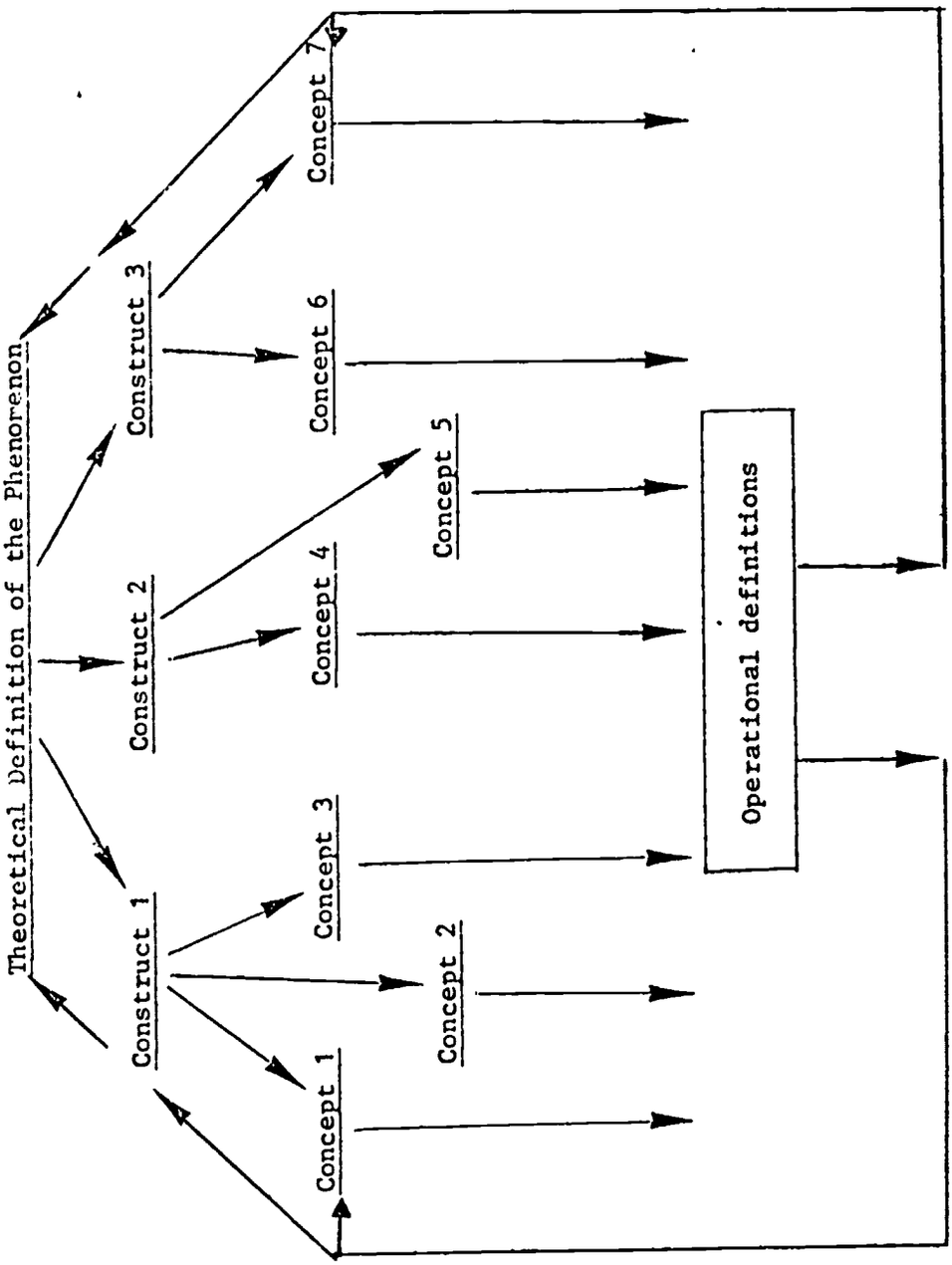


Figure 1. A Conceptual Model for Descriptive Surveys

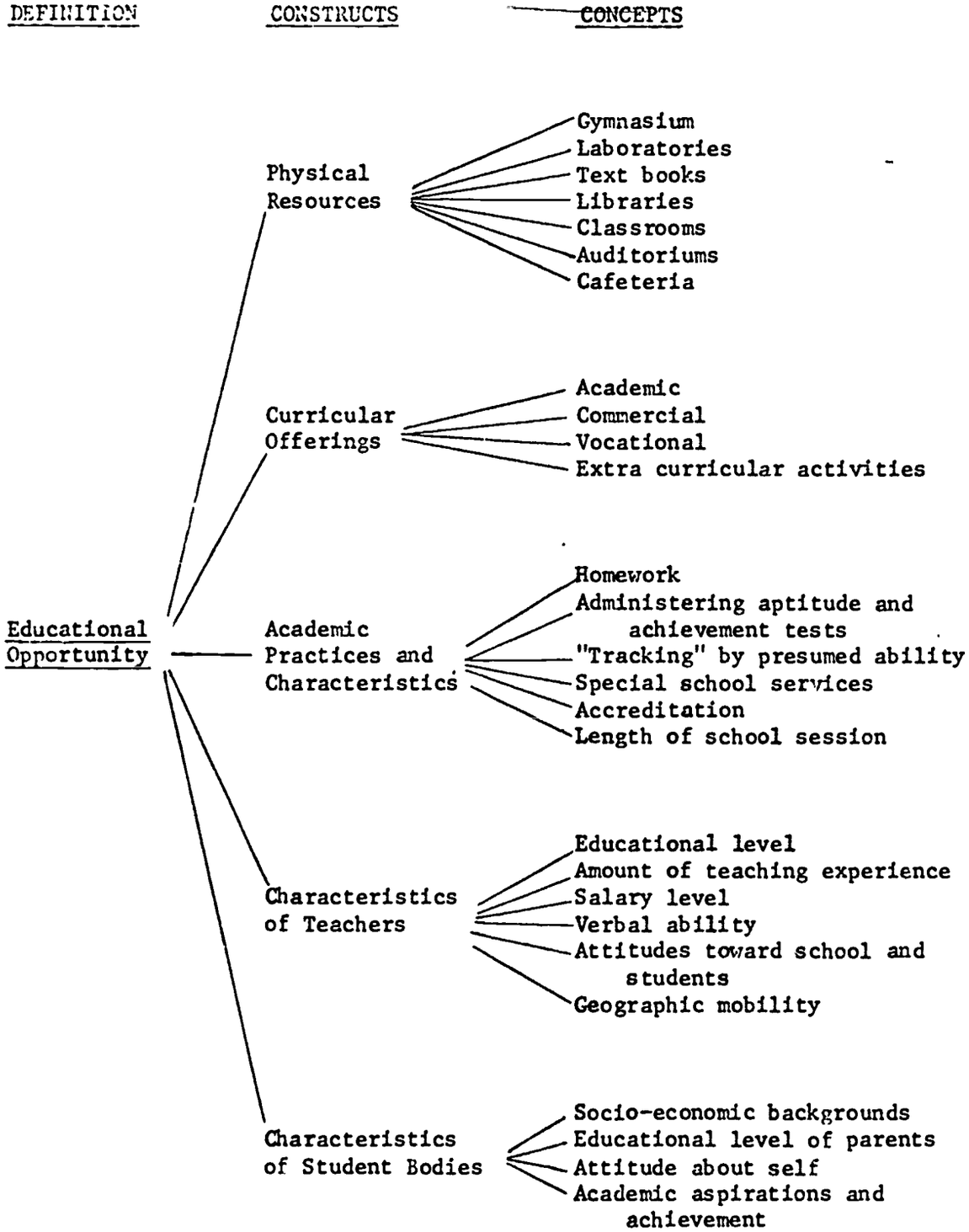


Figure 2. A Partial Conceptualization of Coleman's (1966) Study

Table 1

Illustrating the Presentation of Data in Descriptive Surveys*

Estimated Number of College Students by Race and Region, Fall 1965[a]

	New England	Midwest	Great Lakes	Plains	South	South-west	Rocky Mountains	Far West	Total
Majority	313,514	781,112	821,999	375,043	778,472	434,005	175,000	552,153	4,232,098
Negro	2,216	30,226	30,870	8,500	101,648	20,620	1,605	11,631	207,316
Other minority	1,538	6,542	10,822	2,885	4,996	7,012	1,968	16,092	51,855
Total	317,268	817,880	863,691	386,428	885,116	461,637	179,373	579,867	4,491,269

[a]Based on reports received on 2,013 institutions from among a total of 2,183.

*Reproduced from Coleman (1966, p. 25, Table 11).

comparative criteria serves two major functions. First, it makes the description less abstract for the reader and thus more amenable to interpretation and, when appropriate, formulation of administrative policy and action programs. When a descriptive survey has policy implications, as they usually do in education, the analyst should identify in concrete terms where the findings are more or less applicable. Only by knowing the specific identification of a relevant finding can an administrator be expected to make enlightened policy decisions.

Secondly, comparative description can frequently offer clues or suggest which factors are significantly influencing the phenomena under investigation, although such comparisons, by themselves, do not provide definitive tests of causality. In fact, unexpected or anomalous differences between subgroups can frequently raise questions which will motivate causal and explicative investigations. Consequently, these comparative criteria should not be selected arbitrarily. They are an integral part of the descriptive analysis and can increase the ultimate value or contribution of a descriptive survey.

Data Analysis Strategies for Causal Analysis[1]

In the experimental laboratory causal relationships can be inferred by observing a significant effect between the hypothesized

[1]Although the problem of conceptualization logically and temporally precedes the analysis of the data, data analysis strategies for causal and explicative surveys will be discussed first since many concepts related to the discussion of both conceptualization and data analysis are more easily introduced in the discussion of data analysis strategies.

independent variable and the dependent or criterion variable. The design of the laboratory experiment, which usually includes random assignment to a control and treatment group and purposeful manipulation of the independent variable, greatly reduces the probability that the observed relationship is spurious.

Strict control over the test situation usually is not possible in survey research. Consequently, one of the important limitations of survey analysis is the inability on the part of the investigator to rely exclusively on observation during the test situation to infer causality. In contrast to the procedures of experimental research, most of the controls in survey analysis are of a statistical nature and thus administered after the fact in order to determine whether the observed relationship between variables is actually spurious. That is, the survey analyst is forced, by the limitations of his methodology, to look for variables which could theoretically account for the observed relationship that are antecedent in temporal sequence to the independent and dependent variables. Thus, the researcher tests for spuriousness by statistically controlling for the antecedent variables which could theoretically account for the original relationship.

For example, as illustrated in Table 2, Trent and Medsker (1968) found that college students, when compared with their consistently employed peers, were significantly more likely to change between 1959 and 1963 on:

(1) The Thinking Introversion Scale for females (ψ 4.19) but not for males (ψ .88).

Table 2

Illustrating a Basic Relationship[a]

Standard Mean Thinking Introversion and Complexity Scores of College Persisters and the Consistently Employed, 1959 and 1963

	(N)[b]		Scales and Groups					
			Thinking Introversion		Complexity		Ψ values of College vs. Employed[c]	
	College	Employed	College	Employed	College	Employed	TI	Co
Men	(723)	(444)						
1959			48.62	41.31	50.69	50.95		
1963			51.76	43.57	51.28	48.03		
Difference			3.14	2.26	0.59	- 2.92		
(t)							.88+	3.51**
Women	(578)	(478)						
1959			50.24	43.57	48.79	46.44		
1963			53.74	42.88	50.61	44.43		
Difference			3.50	- 0.69	1.82	- 2.01		
(t)							4.19**	3.83**

[a]Reproduced from Trent and Medsker (1968, p. 133, Table 36).

[b]Samples are composed of those who persisted in college full time during the four-year period of the study and those who persisted in employment full time.

[c]Ψ values constitute the differences in mean differences between the scores of the groups being compared in 1959 and 1963. The computation of the statistical significance of these values may be found in Appendix E.

+p = not significant
*p < .05
**p < .01

(2) The Complexity Scale for females (ψ 3.83) and for males (ψ 3.51).

It could be presumed that the greater positive change observed in the college student population compared to the employed population was due to the impact of the college. However, an alternative hypothesis might be offered. For example, systematic differences in socio-economic status between the student population and the consistently employed population may have been an underlying antecedent variable responsible for the differential changes on the Thinking Introversion and Complexity scales. That is, the college students may have come from a higher socio-economic class than their consistently employed peers and it may be this antecedent difference between the two groups rather than the differential exposure to the college environment which accounted for the relationship in question. To test this rival hypothesis, the investigator needs to control statistically for socio-economic status and observe what effects this third variable has on the original two variable relationship.

According to Glock (1967), the outcome of this procedure can be either explanation or replication.

> Explanation occurs when an antecedent third variable explains away the original relation. Replication occurs when the original relation is repeated when the third variable is taken into account (p. 18).

In other words, explanation occurs when the original relationship disappears after the effects of the antecedent factor have been statistically reduced; confirmation or replication of the original relationship results when the initial effect is sustained after

the antecedent variable has been introduced into the analysis.

When the changes given in the previous example were analyzed according to socio-economic status, Trent and Medsker demonstrated that socio-economic status was able to explain only a portion of the original relationship. For high socio-economic status males, college attendance made no significant difference in change scores on the Complexity scale (i.e., explanation was obtained). Socio-economic status also reduced the original association for high and low socio-economic status females on the Complexity scale and for high and middle socio-economic status males on the Thinking Introversion scale. However, for the remaining subgroups the initial effect of the school was sustained (replication).[2] (See Table 3.)

Thus, socio-economic status did not significantly reduce the original relationship between college attendance and Thinking Introversion scores for females. Social class did explain away the effects of college attendance for high socio-economic status males and high and low socio-economic status females on the Complexity scale. Consequently, Trent and Medsker could feel relatively safe in concluding that social class did not significantly influence the relationship between college attendance and scores

---

[2]Notice that a partial relationship was unmasked when the effects of social class were statistically reduced. In the bivariate condition, scores on the Thinking Introversion scale were unrelated to college attendance for males while a significant relationship between college attendance and Thinking Introversion was observed for low socio-economic class males. The importance of unmasking partial relationships will be discussed later in this section.

Table 3

Illustrating a Test for Spuriousness[a]

Standard Mean Scores on Thinking Introversion and Complexity, 1959 and 1963, of College Students and the Consistently Employed, by Socio-economic Status

SES level & year	(N) College	(N) Employed	Scales and pursuit group: Thinking Introversion College	Thinking Introversion Employed	Complexity College	Complexity Employed	Ψ values of college vs. employed: TI	Ψ values of college vs. employed: Co
MEN								
High SES	(199)	(26)						
1959			50.28	41.71	50.89	51.65		
1963			53.45	44.06	51.91	49.24		
diff.			3.17	2.35	1.02	-2.41		
(t)							.82+	3.43+
Middle SES	(424)	(261)						
1959			48.10	41.23	50.74	50.61		
1963			50.10	43.83	51.11	48.24		
diff.			3.00	2.60	0.37	-2.37		
(t)							.40+	2.74**
Low SES	(73)	(115)						
1959			47.78	41.55	50.07	51.04		
1963			52.06	42.78	50.02	46.77		
diff.			4.28	1.23	-0.05	-4.27		
(t)							3.05*	4.22**

[a]Reproduced from Trent and Medsker (1968, p. 303, Table E-5).

Table 3 (continued)

SES level & year	(N) College	(N) Employed	Scales and pursuit group: Thinking Introversion College	Thinking Introversion Employed	Complexity College	Complexity Employed	ψ values of college vs. employed: TI	ψ values of college vs. employed: Co
WOMEN								
High SES	(183)	(28)						
1959			51.28	47.39	50.22	43.69		
1963			54.97	47.80	51.17	44.69		
diff.			2.69	0.41	+0.95	1.00		
(t)							3.28*	.05+
Middle SES	(316)	(291)						
1959			50.04	43.95	48.07	46.46		
1963			53.66	43.34	50.46	44.30		
diff.			3.62	-0.61	+2.39	-2.16		
(t)							4.23**	4.55**
Low SES	(46)	(119)						
1959			48.02	42.31	46.79	46.70		
1963			50.52	41.68	48.33	44.92		
diff.			2.50	-0.63	1.54	-1.78		
(t)							3.13*	3.32+

+p = not significant
*p < .05
**p < .01

on the Complexity scale for the middle and lower socio-economic class males and middle socio-economic class females.

These tests for spuriousness are essential to the investigator in his search for causality, and the validity of any causal statement in survey analysis rests squarely on the analyst's ability to use this procedure. There is no assurance, however, that two variables are causally related even when replication is the consistent result of testing for spuriousness. The possibility always exists that antecedent variables which were not measured or considered relevant to the analysis will account for the observed relationship. However, if replication is the consistent result of testing for spuriousness, then the probability increases that the original relationship demonstrated a true effect.

Data Analysis Strategies for Explicative Analysis

While bivariate hypothesis testing is of great value to the investigator, it is not the only or even the most important technique for securing information in survey research. Much information can be learned in survey analysis when supplementary research strategies are employed which elaborate or clarify the relationship between two variables by statistically introducing additional variables or test factors into the analysis.[3]

---

[3]Testing for spuriousness is technically part of statistical elaboration since it involves the introduction of a third variable into the analysis to clarify the dynamics of an observed relation. However, to remain consistent with the distinction between causal and explicative data analysis, testing for spuriousness was discussed with reference to causal analysis rather than explicative analysis.

The various statistical elaboration procedures are called different names in the literature.[4] Glock's typology of strategies will be followed in this report. The five basic elaboration strategies identified by Glock are: interpretation, specification, accounting, implication, and accounting-implication, the latter of which Glock labels "phenomenon studies."

1. Interpretation Analysis. Interpretation analysis is undertaken to clarify the processes which produce a relationship between two variables. When a true effect has been observed, the question arises as to how this result should be interpreted. That is, what characteristics of the independent variable are responsible for the observed relationship? Thus, to interpret a relationship the analyst must identify the bonds which link the independent and dependent variables together.

To perform interpretation analysis upon a two variable relationship the researcher statistically controls or reduces the effects of an interpretation test factor--a third variable that intervenes in temporal sequence between the independent and dependent variable. The result of successful interpretation is the same as the result of explanation, that is, the original relationship is sharply reduced or disappears when the effects

---

[4]See, for example, Paul F. Lazarsfeld and Morris Rosenberg, The Language of Social Research. New York: The Free Press, 1955, Section II, pp. 115-125; Herbert H. Hyman, Survey Design and Analysis. New York: The Free Press, 1955, Ch. 7, pp. 275-327; and Morris Rosenberg, The Logic of Survey Analysis. New York: Basic Books, Inc., 1968.

of the interpretation test factor are statistically controlled. Although the statistical outcomes are the same for successful interpretation analysis and tests for spuriousness, the meanings are almost antithetical. When explanation occurs, the original relationship is thought to be spurious because an antecedent variable "explains" the variation in both the independent and dependent variables, as illustrated in the earlier example from Trent and Medsker. With interpretation, however, the validity of the original relationship is confirmed by identifying a third variable related to both the independent and dependent variables which intervenes between the two and at least partially accounts for the original relationship.

Bachman (1970), for example, found a strong relationship between race and scores on the Ammons Quick Test of general intelligence; Black students, as a group, scored significantly lower than White students. However, the researcher observed that the variance of the Ammons Quick Test scores for Blacks was a good deal larger than it was for Whites. This suggested that other factors, imperfectly but nonetheless highly associated with race, were influencing the original relationship. Earlier analyses revealed that Black students were heterogeneous in terms of socio-economic status and exposure to integrated schools, both of which were related to the Quick Test and conceptualized to be intervening in temporal sequence. When the effects of these two variables were statistically reduced, the original relationship between race and scores on the Quick Test all but disappeared. According to Bachman:

> . . . the data on test scores and race add evidence to the view that so called "racial differences" are primarily, if not exclusively--differences in cultural and educational opportunity (p. 84).

In other words, Bachman uncovered the link between race and his particular measure of intelligence. Thus, while the relationship between race and Quick Test scores appears to be real, that is, as a group Blacks tended to score lower than Whites, differences in socio-economic status seems to be the factor which accounted for this relationship.

2. Specification Analysis. A second operation which can be performed on a two-variable relationship involves specifying the conditions which maximize or minimize the strength of the relationship. As Glock (1967) states: "Here, the introduction of a test factor is motivated by the expectation that the strength of the original relation will not be uniform under all conditions" (p. 30).

Specification analysis, like tests for spuriousness and interpretation analysis, involves statistically controlling the effects of a third or fourth variable on a two variable relationship. Unlike tests for spuriousness or interpretation analysis, however, successful specification analysis does not result in the disappearance of the original relationship. Rather, the original relationship is strengthened or weakened, or the direction of the relationship is changed in the partial relationships. Thus, successful specification analysis identifies a third variable (i.e., a specification test factor) that statistically interacts with the original relationship by specifying the

conditions which make the original relationship more or less probable. For example, Newcomb and associates (1967) found that the attrition rate at Bennington College was higher for students who deviated from the norms of the dominant student community. However, these deviants were somewhat less likely to drop out of school if they associated with one another as members of a deviant peer subculture.

There are no time constraints on specification test factors. They can be antecedent, posterior, or simultaneous with the independent or dependent variables in temporal sequence. In addition, successful specification analysis can frequently stimulate further elaboration. Thus, when specification has been observed, the analyst will usually ask why the original relationship was found to be conditional. This, in turn, could lead to an interpretation analysis.

It should also be noted that the use of specification analysis is not restricted to observed two variable relationships. Specification can also be used to identify partial relationships or determine why a predicted two variable relationship was not observed. As previously stated, Trent and Medsker initially found that college attendance was unrelated to scores on the Thinking Introversion scale for males. However, further analysis revealed that college attendance was associated with the Thinking Introversion scale for low socio-economic class males. Consequently, the relationship between college attendance and Thinking Introversion for males was obscured in the bivariate condition because the sample did not contain a sufficient number

of low socio-economic class males to make the relationship observable. In other words, the effect of this partial or conditional relationship was diluted to the point of being undetectable in the bivariate condition because the members of the subgroups in which the relationship was not operating (i.e., the high and middle socio-economic class males) significantly outnumbered the members of the subgroup in which the relationship was operating (i.e., the low socio-economic status males).

In the parlance of survey research, variables which hide or obscure a relationship are called "suppressor variables." It is imperative that the researcher identify the suppressor variables operating in his analysis since they can drastically alter the interpretation of a nonrelationship and result in misleading or erroneous conclusions.

As Kagan and Moss (1962) point out, the variable sex can often conceal or suppress a relationship:

> It may be unwise to pool data for males and females without first examining the data for sex differences . . . if the data had been pooled, many of the relationships between child and adult behavior would have been negligible. For the positive correlation for one sex would have been diluted by the zero order relationship for the other. It is likely that many studies in the literature or in a file drawer would have led the investigator to draw different conclusions if separate analyses had been made for males and females (pp. 275-276).

Since a bivariate relationship can either be real, spurious or suppressed, no two variable relationship ever speaks for itself. A large or significant relationship may be spurious and a small or insignificant relationship may result from the effects

of a suppressor variable. Consequently, it is the responsibility of the analyst to carefully examine both observed relationships and nonrelationships to determine if they are adequate representations of empirical reality.

3. Accounting Analysis. Although accounting, implication, and accounting-implication procedures are especially well suited for educational research, only the former has been used with any frequency. Accounting analysis identifies a number of variables which maximally account for the variation in a criterion variable. Any study which relies on multiple regression analysis performs, to a greater or lesser extent, an accounting investigation.

The works of Coleman (1966), Astin and Panos (1969) and Thistlethwaite (1965) are illustrations of this strategy. Coleman, for example, regressed six and eight background characteristics on the criterion variable verbal achievement as illustrated in Table 4.

At the present time there are no established guidelines to assist the investigator in this type of analysis. However, according to Glock (1967) the most successful accounting studies seem to be those which skillfully incorporate all forms of elaboration into the analysis. In addition, accounting studies should be performed in conjunction with a theoretical model that explicates the interrelationships between the test factors.

4. Implication Analysis. Implication analysis focuses on the independent variable in much the same way as accounting analysis concentrates on the dependent variable. Instead of

Table 4

Illustrating the Results of Multiple Regression Analysis, an Accounting Strategy[a]

Percent of Variance in Verbal Achievement Accounted for at Grades 12, 9, and 6 by Six and Eight Background Factors

	Grade 12			Grade 9			Grade 6		
	Six	Eight	Eight-Six	Six	Eight	Eight-Six	Six	Eight	Eight-Six
Puerto Ricans	3.64	4.69	1.05	3.89	6.18	2.29	23.71	25.51	1.80
Indian Americans	18.89	22.07	3.18	13.92	16.30	2.38	18.40	19.65	1.25
Mexican Americans	7.92	10.23	2.31	12.79	14.25	1.46	21.82	23.07	1.25
Negro, South	14.41	15.79	1.38	12.27	15.69	3.42	14.66	15.44	.78
Negro, North	7.53	10.96	3.43	7.68	11.41	3.73	9.51	10.25	.74
Oriental Americans	11.81	19.45	7.64	12.75	22.81	10.06	34.77	36.16	1.39
White, South	14.75	20.13	5.38	18.40	23.12	4.72	18.14	19.91	1.77
White, North	14.28	24.56	10.28	16.49	22.78	6.29	14.10	15.57	1.47
Negroes, total	13.48	15.14	1.66	12.15	14.99	2.84	14.01	14.62	.61
Whites, total	14.71	23.03	8.32	17.81	23.28	5.47	16.20	17.64	1.44

[a]Reproduced from Coleman (1966, p. 300, Table 3.221.3).

identifying those variables which maximally account for the variance in a single criterion variable, implication analysis attempts to trace the results or consequences of one independent variable on a number of dependent variables. Kagan and Moss, for example, found that the degree of passive and dependent behavior observed in girls during ages six to ten was statistically associated with:

a passive and dependent relationship with their husband or boybriend;

dependency on their parents during their adult life;

withdrawal as adult women;

concern for husbands' job security;

concern for their own job security.

Like accounting analysis, the most useful implication studies are those which incorporate both interpretation and specification test factors into the analysis.

5. Accounting-Implication Analysis. The analyst may also combine both accounting and implication studies into one analysis so that the causes and consequences of a given variable can be investigated. Such studies, according to Glock (1967), "involve seeking to account for the distribution of a variable which is first treated as dependent, then redefined as the independent variable" (p. 38). Accounting-implication studies are most appropriate when the analyst is able to identify a variable which acts as the link between a number of antecedent and posterior variables. This is probably the most sophisticated strategy in survey research for it is theoretically possible to combine all previous operations into one analysis. Consequently,

whether this type of investigation can be hueristically applied to educational research depends on the ability of the analyst to conceptualize multivariable relationships.

Conceptualization of Surveys for Causal and Explicative Analysis

Conceptual models designed for causal and explicative surveys differ from the conceptual models appropriate for descriptive investigations. In addition to focusing upon the manifestations of the phenomena, the investigator is interested in analyzing the dynamics of the phenomena. The major implication of this difference is that the investigator must not only conceptualize the major or salient dimensions of the phenomena but he must also conceptualize interrelationships.

As in descriptive surveys, the conceptual model for causal and explicative analysis should begin with a broad definition of the research problem. In contrast to descriptive surveys, however, this definition will explicitly identify two broad classes of phenomena: the class of criterion or dependent variables and the class of influencing variables that will be called initial independent variables. From these two classes of variables constructs will be formalized and concepts identified.

For example, as mentioned previously, Bachman (1969) defined his research problem in terms of "some major changes in adolescent boys . . . (and) . . . the way these changes are affected by aspects of the immediate social environment" (p. 1). Thus, two major areas of concern were identified: the social

environment and major changes in adolescent boys. Bachman then identified some major constructs and concepts of the dependent variable:

> Much of our interest is focused on dimensions of the person or "personality." Such dimensions include:
> (a) affective states, such as general happiness, anxiety, depression, guilt, and satisfaction with life;
> (b) aspects of the self-concept, including perception of abilities, interests and self-evaluation; and
> (c) values and attitudes, such as social responsibility, attitudes toward jobs, and the perception that one can control his own destiny.
> Our interests also include important plans and behaviors, particularly those relating to educational and occupational aspirations and achievements (p. 9).

About the major dimensions of the independent variable, Bachman says:

> The characteristics of the home, of the peer groups, and of the larger community are all involved, and many will be studied. Of special importance to us, however, are the effects of two environments available to adolescent boys: high school and work (p. 10).

The concepts for these dimensions, according to Bachman, are defined in terms commensurate with person characteristics:

> Ideally, the dimensions we use to measure and characterize environments should be conceptually identical or logically related to the dimensions we apply to people. . . . Accordingly we will measure general ability and aptitude dimensions (such as arithmetic and reading skill) in adolescents and also measure requirements for use of these skills in different school and work environments (p. 12).

It is also necessary in all explicative surveys to reformulate many of the initial independent variables into test

factors that can be used to elaborate bivariate relationships. The analyst should not be solely concerned with identifying causal relationships. He should also explicate the dynamics of these relationships through specification and interpretation analyses. Therefore, many of the initial independent variables should be reconceptualized as specification and interpretation test factors.

This process should be guided by theoretical deduction or intuition since the reconceptualization of variables into interpretation and specification test factors depend upon the logical or theoretical connection between the variables or concepts in question. Interpretation and specification analyses, therefore, represent more than statistical techniques for analyzing data. They represent conceptualizations of the interrelationships between multiple variables which explicate the dynamics of these relationships. Consequently, if theoretical conceptualization is to guide empirical research, the conceptual model should contain specification and interpretation relationships. This process is illustrated in Figure 3.

None of the Analytical Review studies reported this type of conceptual model. However, Bachman did acknowledge the importance of interpretation and specification analysis when he redefined the "person" characteristics as independent variables so that their interactions with environmental characteristics on the criterion variables could be analyzed, as illustrated in Figure 4.

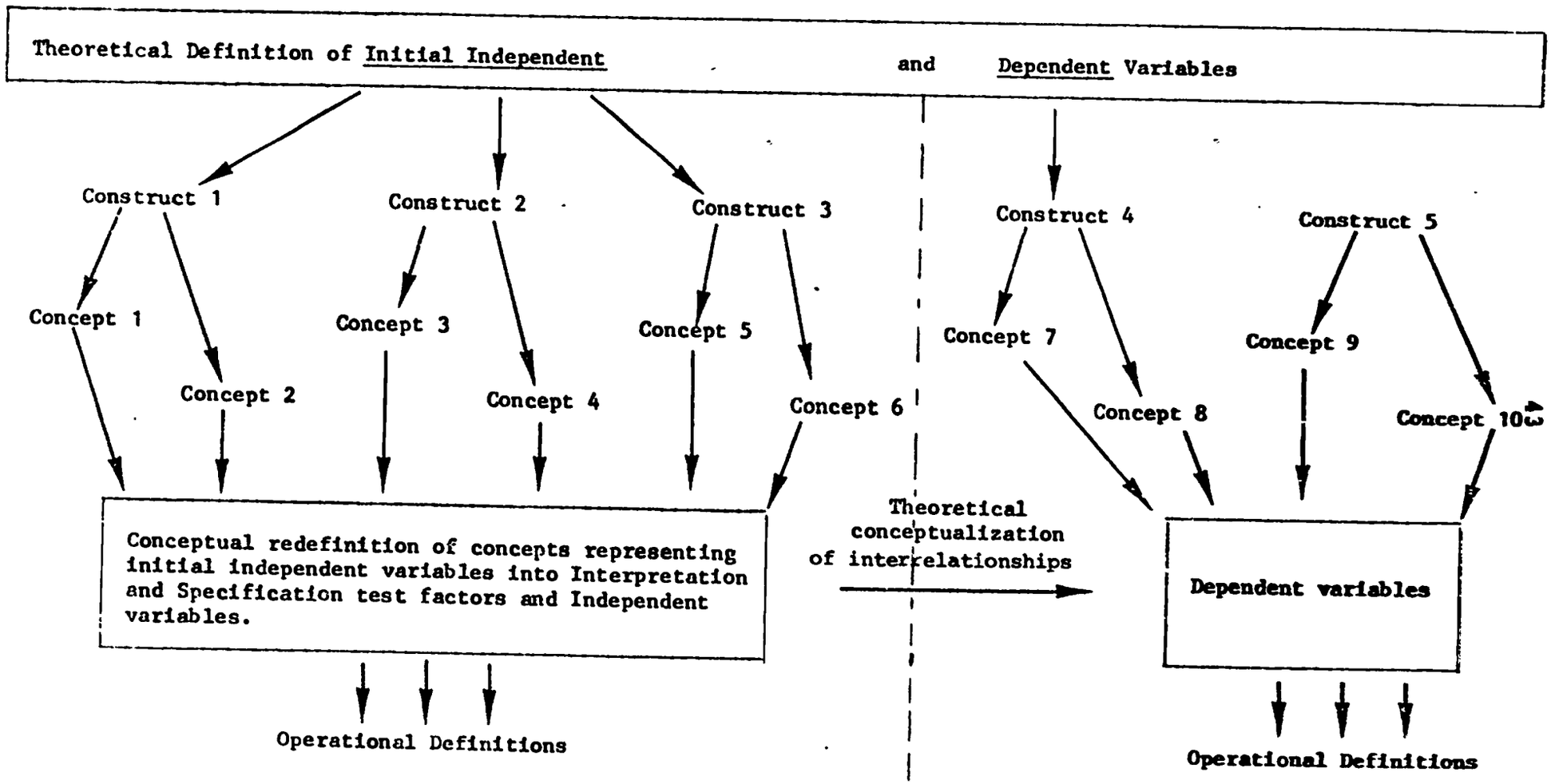


Figure 3. Diagram of a Model for Conceptualization of a Causal and Explicative Survey.

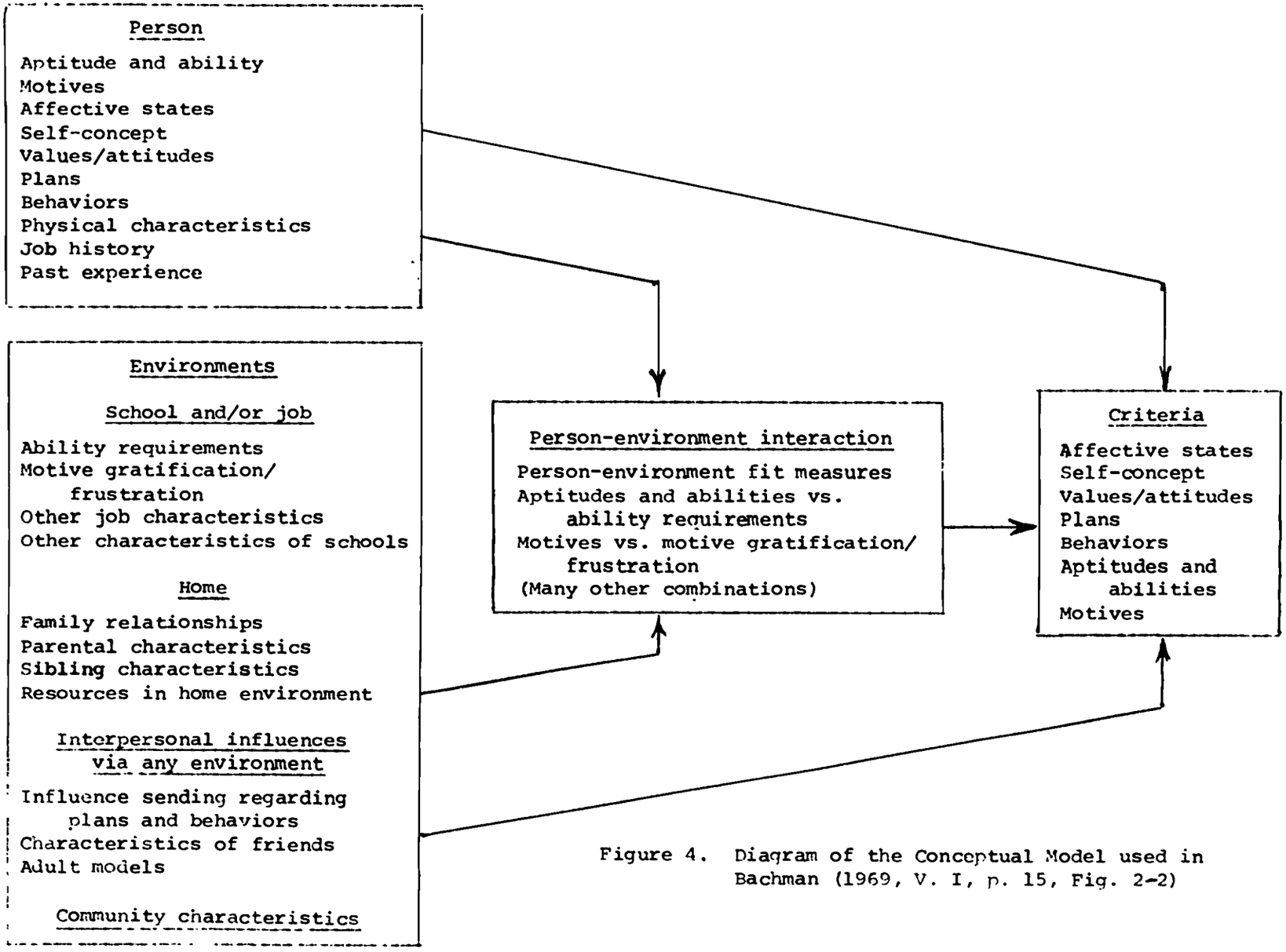


Figure 4. Diagram of the Conceptual Model used in Bachman (1969, V. I, p. 15, Fig. 2-2)

Static and Dynamic Surveys

In survey analysis, as in other methodologies, it is possible for the analyst to collect data at more than one point in time. Survey designs which collect data at one point in time are called static while designs which collect data at more than one point in time are called dynamic. This difference has important implications for the analysis of change and causal relationships.

Cross-sectional Designs. The most frequently used design in survey research provides for the collection of data at one point in time. This strategy is commonly referred to as the cross-sectional survey and is best suited for descriptive analysis. Causal and explicative analyses require, among other things, that the proper time sequence of variables be established. If the temporal sequence of variables is not clearly established by theoretical deduction before the data collection begins in cross-sectional surveys (which is often not possible in educational research), indirect and often unreliable measures of the temporal sequence of variables must be used to provide this information. In other words, collecting data at one point in time does not allow the analyst to observe emergent relationships or to determine on the basis of observation which variables are temporally antecedent to other variables. Furthermore, in cross-sectional designs the inference of change is often problematic. The analyst cannot observe the process of change since at least two measurements from initially equivalent samples are necessary to demonstrate that change has occurred. Thus, cross-sectional

survey designs are not recommended for causal or explicative analyses or for estimating change.

Trend and Panel Designs. Two basic types of dynamic research designs are used in survey analysis: the trend, and the panel. Although both strategies are longitudinal, that is, they employ the same or similar measurements at different points in time and thus are extended over time, they differ in that different people are sampled in trend designs while the same individuals are re-interviewed or re-tested in a panel design. In other words, trend designs collect information from different samples at different points in time while the panel design observes or measures the same individuals over time.

Before the analyst can demonstrate that change in a specific variable has been observed in a longitudinal study, he must assume or demonstrate that the different samples at each time period were initially equivalent. If the samples are not initially equivalent it will be difficult to determine whether the differences observed between the groups are the result of changes associated with the passage of time or are due to differences which initially existed between the samples. In trend designs the analyst must assume or demonstrate that the samples drawn from the different time periods are reasonably similar or equivalent. In panel studies the analyst must assume or demonstrate that the sample has remained essentially intact over time.

In addition to permitting the analyst to investigate change over time, longitudinal designs have a further advantage over

cross-sectional studies in that certain variables can be measured more accurately. Attitude inventories, for example, which ask the respondent to relate how he felt at a previous time may be seriously biased by the subjects' current feelings. If the data is collected in time sequence, this source of spurious association is removed.

A third advantage is that certain phenomena are simply too difficult to examine in the cross-sectional study. For example, in order to systematically investigate the causes of student attrition, a longitudinal study would have to be undertaken in which students were examined at various points in their school career. Those who dropped out at various stages could then be identified and their background characteristics examined for causative influence.

Although trend designs can be used to describe change when the assumption of equivalent samples is justified, without at least two measurements of the same individuals, the analyst is unable to identify with any degree of certainty the individuals who change and the degree to which they change. Thus, trend designs are not recommended for analyzing the dynamics of change.

To illustrate this point, assume that the following cross-tabulations represent trend data from a hypothetical study that analyzes the effects of sex on mathematical ability for elementary school children. The conceptual hypothesis is that the higher the class year, the greater the tendency for male students to score above the 90th percentile than female students.

Table 5

Percentage of Students at or Below or Above the 90th Percentile on Hypothetical Mathematics Achievement Test, by Sex, and by School Year

	Third Grade		Sixth Grade	
	≤90 percentile	>90 percentile	≤90 percentile	>90 percentile
Males	89	11	87	13
Females	87	13	92	8

An examination of Table 5 reveals that in the third grade the female students performed slightly better than their male peers, but in the sixth grade the relationship was reversed: more males than females scored above the 90th percentile. However, the analyst would be unable to describe the dynamics of change for an individual student. There is no way of knowing, for example, whether the 2 percent increase among the male students was due to a simple increase in the number of males scoring above the 90th percentile or the result of a more complex pattern of change-slippage wherein some unknown percentage of males slipped below the 90th percentile and a different, 2 percent-larger, group now scored above the 90th percentile. Conversely, the investigator could not determine if the slippage among female students was simply a 5 percent loss or the loss of the entire group originally above the 90th percentile that was partially counterbalanced by an 8 percent increase in other girls scoring beyond the 90th percentile.

Many of the problems which are associated with cross-sectional and trend designs can be circumvented by the use of panel studies. By re-interviewing or "retesting" the same respondents, turnover tables can be constructed which identify the direction and frequency of specific change patterns. Table 6 depicts the turnover tables the investigator might have constructed if a panel study had been performed in the previous example.

By repeating measurements on the same students the analyst can determine the dynamics of change at the individual level. Thus, the data in Table 6 indicate that 5/87 of the girls and 5/89 of the boys moved into the top 10 percent of their class whereas 10/13 of the girls but only 3/11 of the boys dropped below the 90th percentile. Consequently, the girls were less stable than the boys with the dominant change pattern representing a rather dramatic shift in the number of girls who originally scored in the 90th percentile as third graders but dropped below this mark by the sixth grade.

Once the analyst has identified specific change patterns, more sophisticated analyses can be performed to determine the causes of these changes. For example, the researcher might isolate the changers from the nonchangers, or those students who moved up from those students who moved down in order to determine what factors, if any, discriminate between these comparison groups. Thus, in panel analysis the researcher can not only observe and describe change, he can analyze the

Table 6

Percentage of Girls and Boys at or Below and Above the 90th Percentile on Hypothetical Mathematics Achievement Test by School Year

GIRLS

Sixth Grade

Third Grade	≤ 90th percentile	> 90th percentile	
≤ 90th percentile	82	5	87
> 90th percentile	10	3	13
	92	8	100

BOYS

Sixth Grade

Third Grade	≤ 90th percentile	> 90th percentile	
≤ 90th percentile	84	5	89
> 90th percentile	3	8	11
	87	13	100

dynamics of change, a process not possible in either cross-sectional or tend studies.

Panel studies, however, are considerably more expensive to conduct than cross-sectional and trend studies. Nevertheless, tney are the best means available for analyzing causal relationships and the dynamics of change in survey research. Since most studies in education reflect at least one of these concerns, it is apparent that panel surveys can make a meaningful contribution to educational research.

CHAPTER III

THE USE OF CAUSAL AND EXPLICATIVE ANALYSIS IN THE ANALYTICAL REVIEW STUDIES

Almost without exception, the search for causal relationships has become the primary concern of educational research. As a result, the statistical elaboration strategies set forth in Chapter II are described in this chapter as they pertain to the Analytical Review studies. In particular, the researchers' success in inferring the temporal sequence of variables and testing for spuriousness is evaluated. In addition, the use of interpretation and specification analysis is also assessed. Specific procedures are recommended to assist researchers in dealing with each of these issues.

Causal Analysis in Education

When the analyst employs a statistical technique which assumes an asymmetrical relationship between two variables, such as regression analysis or analysis of variance, or when one of the variables in a specific test represents the criterion for assessing the impact of the educational institution, he is at least implicitly searching for causal relationships.

Before the analyst can infer that one variable has "caused" another, however, it must be established that (1) the variables concomitantly vary with one another; (2) the independent variable precedes the dependent variable in temporal sequence; and (3) the observed relationship is not due to other antecedent factors. The first requirement, demonstrating a statistical

association between the independent and dependent variables, is discussed in Chapter IV. In the following two sections, the problems of establishing the proper temporal sequence of variables and testing for spurious relationships are examined.

Temporal Sequence of Variables. The difficulty in determining the temporal sequence of variables in social research stems from three basic sources. First, the investigator frequently must study social phenomena in their natural setting which usually means that the independent variable cannot be manipulated and thus the researcher lacks control over the timing of exposure to the variable in question. In addition, researchers usually cannot observe the complete life cycle of social phenomena; instead they must often make do with one or more observations over a limited period of time. Thus, the researcher may not be able to determine the temporal sequence of variables on the basis of empirical observation because the effect or impact of a suspected independent variable was experienced by the target population prior to measurement. Finally, the variables of social science often form a discursive system, that is, they tend to be symmetrically or mutually related to one another so that one variable can reasonably be considered the cause as well as the consequence of another variable.

Educational research is, of course, one form of social research. Hence, it is to be expected that educational researchers will experience the same methodological problems as other social scientists. Determining the temporal sequence of variables is one such problem. Even in longitudinal studies, which greatly

reduce the problem of determining the temporal sequence of variables, it is often difficult to determine which variable occurred first when variables from approximately the same time period are analyzed. Thus, the analyst may find it difficult to determine the temporal sequence of variables when student background characteristics are correlated with one another or when a relationship exists between particular school characteristics.

Hilton (1971), for example, found that 70 percent of the high school students taking college preparatory classes reported that 60 percent or more of their friends planned to attend a four-year college while only 28 percent of the students not taking college preparatory classes reported such a high percentage. In other words, students taking college preparatory courses were more likely to have friends who planned to attend college than were students who did not enroll in college preparatory courses. Assuming for the moment that the relationship is not spurious, this finding can be interpreted in one of two ways. The first interpretation assumes that the effects of the peer group are antecedent to the student's educational aspirations. Thus, in deciding whether to enroll at a four-year institution or take college preparatory classes, the student may give serious consideration to his peer group's value orientations regarding education and their post-high school education plans. However, an equally plausible interpretation is that students' educational aspirations are antecedent to the peer group. In other words, the decision to take college preparatory courses or to attend a four-year college may cause the student to identify with a particular peer group. The student may thus select his

friends on the basis of their common educational interests or values, or as Hilton points out, a particular high school curriculum can structure and facilitate social interaction which will often contribute to the development of friendship associations. It is apparent that no definitive conclusion can be reached concerning the temporal sequence of variables in this relationship. Therefore, without additional information the analyst cannot make a reasonable inference as to the most likely temporal sequence of variables.

Another example of the problem of identifying the temporal sequence of variables is provided by Bachman in his study of high school students. Bachman (1970) found a substantial positive correlation between the Crowne Marlowe Social Approval Scale and students' self-reports of good family relations. This might suggest that good family relations are a **cause** of high scores on the Crowne Marlowe scale. However, as Bachman notes:

> An alternative explanation is to consider the family relations measure as reflecting rather than causing the need for social approval. If a boy has a strong need to portray himself in a favorable light, perhaps he will for the same reasons describe his family relations in very favorable terms (pp. 116-117).

Therefore, the need for social approval could be temporally antecedent to the student's family relations and thus partially responsible for the scores on the family relations index. If this is the case, then the validity of the family relations index must be questioned. That is, the index may be measuring the student's need for social approval rather than his actual family

relations. In fact, Bachman seems to believe that the desire for social approval was the antecedent factor in this relationship. According to Bachman:

> We have noted before that the family relations measure is highly subjective; now, given its substantial correlation with the Crowne Marlowe scale, we must be even more suspicious about the extent of its validity as a measure of the actual relationship between a boy and his parents (p. 119).

Thistlethwaite's study provides several additional illustrations of this problem. The data from this study strongly suggest that at least some of the college experiences and college press investigated caused students to change their motivation or desire to seek advanced educational training. Thistlethwaite raises the possibility that the types of college press and experiences reported by the students may have been the effects, rather than the cause, of changes in the disposition to seek advanced training. For example, the students' disposition to seek graduate training was found to be strengthened by undergraduate participation in honors programs, graduate-level courses, research programs and projects. According to Thistlethwaite:

> It would be a mistake to interpret these correlations as proving that each of the experiences significantly related to the residuals caused changes in the disposition to seek advanced training. . . . These experiences were most likely the consequences, rather than the cause, of changes in disposition to seek higher educational attainments. Obviously, we cannot determine the direction of causation from such correlations (p. 94).

The longitudinal study by Kagan and Moss provides another illustration of the problem of inferring the temporal sequence of variables. The investigators report a high correlation between the desire for social recognition and achievement behavior for both men and women. It is not clear, however, whether successful achievement behavior or goal attainment resulted in or was caused by the desire for social recognition.

Kagan and Moss suggest that individuals' early success in achieving personal goals stimulates them to seek recognition for these accomplishments. According to Kagan and Moss:

> The similarities in the pattern of correlations for achievement and recognition behavior are more striking than the differences. This congruence suggests an intimate relationship between behavior aimed at satisfying internal standards of excellence, particularly intellectual pursuits, and the search for social recognition for this competence (p. 134).

Again, an equally plausible interpretation is that the desire for social recognition motivated the individual to engage in socially approved achievement behaviors in which they had a hign probability of successful goal attainment.

As a final example, Tillery et al. (1972) observed that students who did not attend college and short-term college attenders had similar grades and aspirations for grades at grade nine and ten. However, from grade ten to grade eleven the number of nonattenders aspiring to only a high school education dramatically increased and was paralleled by a general decrease in the distribution of the grades they achieved and the grades to which they aspired. For the

short-term college students, on the other hand, a similar shift was observed, but as a group their achievement and aspirations were higher than the nonattenders at grade eleven. Thus, for both nonattenders and short-term college students a decrease in grades achieved was associated with a decrease in educational aspirations between the tenth and eleventh grades. It is not clear how this relationship should be interpreted. That is, the decrease in aspirations may have contributed to lower grades or vice versa.

Problematic cause and effect relationships often deprive administrators and policy-makers of important information. Hilton would have increased the contribution of his investigation if he had been able to determine if the peer group significantly influenced the students' educational aspirations or if the students' educational aspirations determined the composition of his peer-reference group. Similarly, Thistlethwaite's analysis was limited by his inability to ascertain whether certain school experiences, which could be manipulated by administrators, were the causes or the consequences of high educational aspirations. Kagan and Moss could have contributed valuable information to educational theory by determining whether social recognition significantly increases achievement behavior. Finally, Tillery's analysis could have provided administrators and teachers with important information regarding the relationship between aspiration and grades. Unfortunately, however, no attempt was made to identify the most probable direction of causation.

Clearly, the contribution of a study is greatly increased when the analyst is able to determine the probable temporal sequence of variables. Four basic methods or strategies which can be used in conjunction with a longitudinal design to infer the most likely temporal sequence of variables are described in the following section. These strategies are: Theoretical or Logical Deduction; Retrospective Analysis; Cross-lagged Panel Correlation; and Conceptualizing and Testing Differential Outcomes.

1. Theoretical or Logical Deduction

In many relationships the analyst can determine the most reasonable temporal sequence of variables on the basis of theoretical deduction or intuition. This technique works best, however, when the analyst is guided in his research by an appropriate theoretical model. Many conceptual models described in the literature identify at least three basic groups of variables which are arranged according to temporal sequence: student background or input characteristics, process or mediating factors which include school characteristics, and educational or student outcomes. This conceptualization is illustrated below:

student background or input characteristics ⟶ process or mediating factors ⟶ educational or student outcomes

Student inputs are those factors which the student brings with him as he enters school. Process or mediating variables are usually the characteristics of the school, including the quality of the plant site, teaching staff, instructional

material and procedures. Student input factors interact with process variables producing certain outcomes such as student verbal or mathematical achievement, or personality and attitude change.

Most of the student input characteristics such as race, sex, aptitude, size of family, parents' social class, etc. are antecedent in temporal sequence to process variables or educational outcomes. However, some of the student attributes that are conceptualized as input characteristics, such as the student's self-concept or the degree to which parents encourage academic success, can be the result as well as the cause of specific educational experiences and outcomes. Consequently, the researcher must carefully evaluate each observed relationship to determine if theoretical deduction or intuition provides a sound basis for selecting a particular temporal sequence.

Many examples can be cited where common sense, intuition, or theoretical deduction was appropriately used in the Analytical Review studies to infer the temporal sequence of variables. For example, Lehmann and Dressel (1962) found that student's critical thinking ability was related to father's education; the higher the critical thinking score the greater the tendency for the student's father to have attained a high level of formal education. In this case, the researchers reasonably assumed that father's education was the causal variable in this relationship and thus temporally antecedent to the student's ability to think critically.

Bachman observed that the higher the social class position of the high school student's family the more likely he was to have high occupational aspirations. Socialization and child development theories emphasize the important role that social class plays in determining the educational and occupational aspirations of the student. Moreover, it would not make sense to conclude that the occupational aspirations of the student determine his parents' social class position. Thus, on the basis of theoretical deduction the best estimate of temporal sequence is that social class is the antecedent variable in this relationship.

Husen (1967) also used theoretical deduction. He reports that father's occupational level was positively related to mathematics interest scores. Father's occupation is often used as a measure of social class and is undoubtedly related to socialization and childrearing practices that in turn influence the educational aspirations and interests of the student. Consequently, the most reasonable interpretation is that father's occupation is temporally antecedent to the student's interest in mathematics.

As a final example, Coleman reports that race is strongly associated with educational achievement; the average minority student (with the exception of the Oriental American) scores distinctly lower than the average White pupil on standardized achievement tests at every grade level examined. Clearly, race or ethnicity is antecedent in temporal sequence to educational achievement.

2. Retrospective Analysis

When the problem of identifying the temporal sequence of variables is anticipated by the researcher before fieldwork begins, it can be circumvented by modifying the measuring instrument to include retrospective questions which ask the respondent to clarify specific temporal sequences.

Thistlethwaite, for example, was able to identify through retrospective analysis the dominant temporal sequence of variables in the relationship between exposure to college and plans to seek advanced educational training. He asked the sample of students at the completion of their senior years if their decision to seek or forego advanced training was made at the beginning of college, after the second year of college, or after the completion of college. The data from these retrospective reports indicated that approximately half of all the graduates sampled planned to do graduate study before they entered college. After two years of college, however, over 60 percent reported plans for advanced training. At the time of college graduation, about 85 percent of the graduates indicated that they expected to obtain advanced degrees. Thus, although there was a strong positive association between length of exposure to college and the proportion of students opting for graduate training, there was nonetheless a substantial percentage of students who had decided to attend graduate school before they entered college as undergraduates.

Earlier, it was noted that Thistlethwaite was unable to conclude whether undergraduate participation in honors programs,

graduate level courses or research projects occurred before or after the student developed an interest in attending graduate school. If Thistlethwaite had been able to anticipate this problem too, he could have included additional questions which asked the students to clarify these temporal sequences. Similarly, Hilton could have unraveled the temporal sequence in the relationship between the educational aspirations of the student and the educational aspirations of his friends by asking the students if their friendship associations existed prior to their decision to take college preparatory courses or to attend a four-year college.

Lehmann and Dressel (1963) also used retrospective analysis to establish the temporal ordering between various attitudes and opinions and amount of college attendance. The students were asked to indicate whether they considered themselves basically vocational, academic, collegiate or nonconformist in educational orientation both when they were seniors and as they recalled their orientation as freshmen. All students, with the exception of junior-year male withdrawals, felt that as freshmen they were more concerned with attending college for vocational preparation. An increase in the concern for academic interests was also observed between the freshmen and senior years along with a reduction in the percentage of students who could be described as collegiate. In addition most of the students felt that they had become more flexible, less authoritarian, more open-minded and understanding of others between their freshmen and senior years. Many students changed their ideas about

behavior standards, were better able to define their life goals and became more confident in their ability to handle new problems.

Although retrospective analysis can enhance the information obtained in a study, the analyst should be aware of two major problems. First, respondents may not be able to recall the situation in question, or if they do, the response may be unreliable or invalid due to distortions in memory or selective perception. Thus, when using retrospective questions the analyst is advised to include checks on response error. Internal and external checks for response error which can be adapted to retrospective questions are described in Chapter V. A second major problem stems from the fact that researchers often lack the necessary prescience to include retrospective questions in their questionnaire. A systematic evaluation of the need for retrospective questions can greatly reduce these unfortunate oversights.

3. Cross-lag Panel Correlation

The only temporal restriction on the independent variable in a causal relationship is that it does not occur after the dependent variable. Consequently, it may be impossible for the analyst to determine the proper temporal ordering because the independent and dependent variables occur simultaneously or so close to one another in temporal sequence that empirical observation cannot determine which variable occurred first.

When the independent and dependent variables occur simultaneously, or when the previous methods cannot be used, the

researcher can still identify the most probable direction of causation between two variables by comparing the relative strength with which both variables predict each other at a later date (cf. Lipset, Lazarsfeld, Barton and Linz, 1954). This method is based upon the analysis of a sixteenfold or sixteen-cell table which requires that the researcher have measures of both the independent and dependent variables at two points in time.[1]

Earlier in this chapter it was pointed out that Tillery observed a concomitant reduction in the grades and aspirations of certain high school students between their sophomore and junior years. Tillery was unable to clarify the temporal sequence of variables because both variables changed between measurement periods. However, he did have measurements of the grades and educational aspirations of the students at two points in time, the tenth and eleventh grade. Consequently, Tillery could have combined both variables forming a composite measure of grades and aspirations for both time periods and constructed a simple turnover table. When both variables are dichotomized into high-low, plus-minus, etc. and then combined into the composite variable, the result is a sixteenfold table. Table 7 illustrates a sixteenfold table using hypothetical data representing interrelationships between student grades and educational aspirations at grade ten and eleven.

---

[1] A statistical model for the sixteenfold table analysis may be found in J. R. Murray (1971).

Table 7

A Sixteenfold Table with Hypothetical Data[a]

GRADE 10							
grades	aspirations	grades aspirations	high high	high low	low high	low low	Marginals
high	high		148	49	89	93	379
high	low		151	243	27	74	495
low	high		57	16	203	191	467
low	low		6	14	21	453	494
	marginals		362	322	340	811	1835

[a]A statistical model for the sixteenfold table analysis may be found in J. R. Murray (1971).

The marginals in Table 7 indicate that student grades and educational aspirations decreased between the tenth and eleventh grade. At grade ten, 494 students had both low educational aspirations and grades while at grade eleven 811 students had the same status. The question that remains is whether school grades are primarily responsible for educational aspirations or whether aspirations are primarily responsible for grades.

The answer to this question can be determined on the basis of which variable has the greatest influence on the other. The values of the composite variable at grade ten that are inconsistent, that is, high grades-low aspirations and low grades-high aspirations, can be compared with the scores at grade eleven to identify the dominant direction of change. If the inconsistent scores on the composite variable at grade ten change in a direction that results in greater consistency between the two variables forming the composite variable at grade eleven, then the variable that has changed to produce this increased consistency is presumed to be the major dependent variable. Thus, if high educational aspirations at grade ten tend to be reduced at grade eleven when the student has received low grades at grade ten, while low grades at grade ten do not rise appreciably at grade eleven when the student has high educational aspirations at grade ten, the educational aspiration of the student is considered to be the dominant dependent variable and school grades the major independent variable.

The hypothetical data in Table 8 indicate that the aspirations of the students are more likely to change than their grades. In other words, there is a greater tendency for the

Table 8

Hypothetical Data Illustrating the Use of a Sixteenfold Table

Patterns of Students with Inconsistent Values on the Composite Variable Grades-Aspirations				
Grade 10		Grade 11		
grades	aspirations	grades aspirations	high high	low low
high	low		151	74
low	high		57	191

educational aspirations of the student to change in a direction that is consistent with his prior grades than for the student's grades to move toward greater consistency with his earlier educational aspirations. Thus, student grades received at grade ten predict educational aspirations at grade eleven better than educational aspirations at grade ten predict student grades at grade eleven. Clearly, the statistical association should be greater between the effect or dependent variable (educational aspirations at grade eleven) and its prior cause (student grades at grade ten) than between a "prior effect" (educational aspirations at grade ten) and its "subsequent cause" (student grades at grade eleven).

The utility of the sixteenfold table is restricted, however, in that dichotomous data are required and large sample sizes are usually necessary to perform the analysis. Nevertheless, Campbell (1962) and Pelz and Andrews (1964) have independently extended the generality of this technique by proposing that correlation analysis be used. This technique has come to be known as the cross-lagged panel correlation. As illustrated in Figure 5, the underlying assumption of the cross-lagged panel correlation is that if

$$P_{X_1Y_2} > P_{Y_1X_2}$$

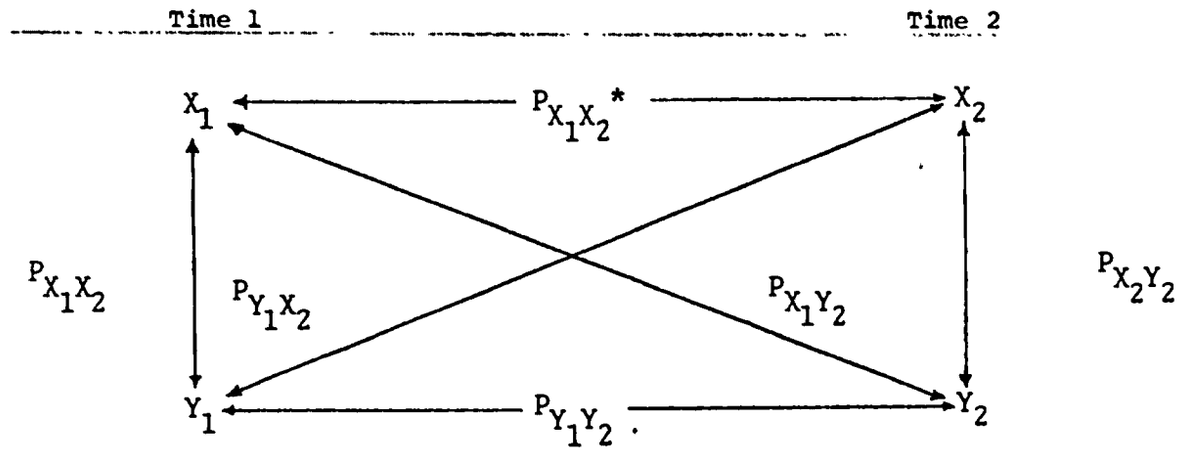
then X is the dominant cause of Y, and if

$$P_{Y_1X_2} > P_{X_1Y_2}$$

then Y is presumed the dominant cause of X.[2]

---

[2]Statistical models for the cross-lagged panel correlaton are discussed in A. S. Goldberger (1971); J. W. Keesling and D. E. Wiley (1969); J. R. Murray, D. E. Wiley and R. G. Wolfe (1971); D. E. Wiley and J. A. Wiley (1970).



P = correlation coefficient

Figure 5. Illustrating Cross-lagged Panel Correlation.

Bohrnstedt (1969) notes that the cross-lagged panel correlation technique fails to take into account that the best predictor of X and Y at time two is likely to be the same variable measured at time one. In addition, Bohrnstedt points out tnat the correlation coefficient and the standardized regression coefficient are sensitive to the magnitude of the standard deviation and thus will vary across subsamples with different variances. Consequently, the researcher is advised to compare unstandardized regression coefficients rather than correlation coefficients and to partial out the effects of initial position when using the cross-lagged panel technique.

The cross-lagged panel correlation is a relatively new methodological innovation which has great potential for identifying the dominant direction of causation in educational research. To use this technique it is necessary that identical questions be used to collect the longitudinal data. Furthermore, the data representing the time one measurements should be collected from all respondents during the same time period; likewise, the time two measurements should be collected together. These restrictions require that the analyst anticipate the problem of causal direction before data collection begins so that the necessary changes in research design and the measuring instrument can be made which will satisfy these requirements.

Cross-lagged panel correlations were not utilized in any of tne Analytical Review studies yet many could have benefited from this form of analysis. When the analyst fails to anticipate

the problem of causal direction and consequently does not collect the necessary data to perform a cross-lagged panel correlation, the outcome can diminish the contribution of the study. As Thistlethwaite writes:

> Unfortunately, the press scales and instructions successively administered in this study were not identical. Moreover, the reports of college press had different time referents than the questions about degree aspirations. Consequently, this method (cross-lagged panel correlation) could not be used to analyze the present panel data, and it was not possible to rule out this second rival hypothesis (1965, p. 106).

4. Conceptualized Differential Outcomes

Another method for determining temporal sequences consists of conceptualizing different theoretical models that vary with respect to which variables are presumed to influence others. Different predictions or outcomes are deduced from the competing models and a choice among the models is made on the basis of which conceptual model comes closest to fitting the data.[3]

Newcomb (1943, 1947), used this technique to illustrate the "liberalizing" effect of Bennington College on student attitudes. The tendency for students at Bennington College to be liberal could have been due primarily to the impact of the school. This relationship could also have resulted from a

---

[3]Thus, the cross-lagged panel correlation represents a conceptualization and a test of differential outcomes since it involves the selection of competing models on the basis of the existing data. However, the cross-lagged panel correlation was discussed separately because it requires very little theoretical conceptualization by the analyst.

selection bias; Bennington College students may have initially been more liberal and as a result decided to attend Bennington College because of its reputation as a liberal institution. In other words, did Bennington College "cause" liberal attitudes or were these liberal attitudes primarily antecedent in temporal sequence to the college experience and thus a possible causal factor in attending Bennington College?

To answer this question Newcomb reasoned that if the school was the principal causal variable in this relationship students who had attended the school for longer periods of time would display more liberal attitudes than students who had attended for a shorter period of time. If, on the other hand, it was selection bias (i.e., possession of liberal attitudes that led to attending the college), that was primarily responsible for the relationship in question, then there should be little difference in the attitudes of students with different rates of exposure to the college environment. In other words, it was hypothesized that if the school was having an impact on student attitudes, then the longer the exposure the greater the impact.

Newcomb tested what might be called "the cumulative effect hypothesis"; if X is the cause of Y and thus temporally antecedent to Y, then the longer the exposure to X, the greater effect on Y. Using a variety of tests, he was able to demonstrate that length of exposure to the Bennington College environment was indeed associated with increased liberalism--that college impact was the principal causal factor in this relationship.

Another test for identifying the most probable temporal sequence of variables requires that information be collected on a variable that is antecedent to and statistically associated with both X and Y. In Chapter II it was pointed out that an interpretation test factor intervenes in temporal sequence between the independent and dependent variables. Thus, if it can be demonstrated that X or Y operates as an interpretation test factor in a relationship between either X or Y and a third variable that is clearly antecedent to both variables, then this interpretation test factor will be the antecedent variable in the relationship between X and Y. In other words, assume that X and Y are statistically associated with each other, and that antecedent variable W is statistically associated with both variables and can be considered the independent variable for either X or Y. A reasonable inference can then be made concerning the temporal sequence operating in the original relationship if either of the two variables function as an interpretation test factor in the relationship between W and X or W and Y. If X statistically interprets the relationship between W and Y, then X is temporally antecedent to Y. If on the other hand, Y statistically interprets the relationship between W and X, then Y will be temporally antecedent to X.

Blau (1955) illustrates this technique in his study of elderly people and their self-concepts. Blau found that an individual's self-concept with respect to age was related to his idea of how significant others perceived him. People who considered themselves old or elderly were more likely to

believe that others thought of them as old than were people who considered themselves to be middle-aged. As Blau notes, two opposite interpretations could be introduced to explain this relationship:

> First, if an individual with advancing years starts to conceive of himself as old and to act as an old person, others will treat him as such, but if he continues to identify himself with middle-aged people, his associates will usually not think of him as old. . . . Second, if and only if, his significant others begin to treat an individual who is getting old as an old man or woman, his self-image will change from that of a middle-aged to that of an old person (pp. 101-102).

In other words, does the individual's self-image of his age determine his perception of how significant others view his age or does the individual's perception of how significant others view his age determine his own conception of his age?

The existence of data on a common antecedent variable, chronological age, allowed Blau to infer that self-image of age was the antecedent variable in this relationship. Chronological age correlated with both self-image and the respondent's perception of how significant others viewed his age; older people (70 years and older) were much more likely to define themselves as old and indicate that their friends also considered them to be elderly than were younger respondents (between 60 and 70 years old). Furthermore, chronological age could not explain the relationship between self-image and significant others' conception of age. That is, statistically controlling for chronological age did not produce a spurious relationship between self-image and the respondent's perception

of how significant others viewed his age. Since it is likely that chronological age is the independent variable regardless of whether self-image or significant others' conception of age is defined as the dependent variable, two alternative hypotheses were advanced.

In the first hypothesis Blau proposed that self-image of age interprets the relationship between chronological and significant others' conception of age and was thus temporally antecedent to the respondent's perception of how significant others viewed him. In the second hypothesis the individual's perception of how significant others viewed his age was believed to be antecedent to the respondent's self-image and, to a large degree, responsible for the relationship between chronological age and self-image of his age. By performing an interpretation analysis to test both hypotheses, Blau was able to demonstrate that self-image of age interpreted the relationship between chronological and significant others' conception of age and was thus the antecedent variable in the original relationship.

This particular test of conceptualized differential outcomes was not in evidence in the Analytical Review studies. This is unfortunate particularly because the requirements for this test are modest. The only data necessary to perform this analysis is information on a common antecedent variable that can be conceptualized as an independent variable. Bachman, for example, might have been able to use this technique to determine whether the need for social approval was antecedent or subsequent to the student's reported family relations since he also

found that race, a variable that is clearly antecedent to a student's family relations and his need for social approval, was highly correlated with both variables in question. Consequently, Bachman could have formulated and tested two competing hypotheses in which one identified family relations as an interpretation test factor between race and the need for social approval while the other hypothesized that the need for social approval would interpret the relationship between race and family relations. Actually, the neglect in using this test is symptomatic of a much larger problem in educational research, the lack of systematic interpretation analysis which is discussed later in this chapter.

Although the general procedure of testing differential outcomes can be of great value in identifying temporal sequences, a word of caution is in order. There is no magic formula, conceptual model or test that can be applied to every relationship to identify the most probable temporal sequence of variables. A conceptual model that posits differential outcomes depending upon which temporal sequence is operating may yield an appropriate and fair test for one particular relationship, yet be inappropriate for another relationship. The "cumulative effect hypothesis," for example, would not be an appropriate test for identifying the temporal sequence of variables operating in a particular relationship when there is no reason to expect that X will have a cumulative impact on Y.

What the investigator conceptualizes in terms of testing competing temporal sequences, and how this test is conducted,

is dependent upon the particular variables involved and the data that are available for analysis. Consequently, there is almost no limit to the number or type of tests that can be performed to help clarify the temporal ordering of variables. The degree to which testing differential outcomes will illuminate the temporal sequence of variables is therefore ultimately dependent upon the researchers ingenuity in conceptualizing the problem, formulating the test and manipulating the data.

In general, the researchers of the studies under review did not experience a great deal of trouble ferreting out the probable temporal sequence of variables. This was due in large part to the fact that most of the studies were longitudinal. There were instances, however, where this problem did confound the interpretation of a relationship. Many of these situations could have been avoided if the analyst had collected the necessary data to perform a retrospective analysis, a cross-lag panel correlation or a test of conceptualized differential outcomes. It is important, therefore, that researchers anticipate the problem of determining temporal sequence before data collection begins and gather the information necessary for proper causal analysis.

The importance of determining the temporal sequence of variables, however, is clearly not restricted to causal analysis. It is also an important consideration in interpretation analysis. As previously mentioned in Chapter II, successful tests for spuriousness and interpretation have the same

statistical outcome. Both tests result in the disappearance of the original relationship when the effects of the third variable are statistically reduced. The critical difference between the two procedures involves the temporal sequence of the test factors. In tests for spuriousness, the test factor is antecedent to the independent variable; in tests of interpretation, the test factor intervenes between the independent and dependent variable. Consequently, there is no way to determine if the analyst is testing for spuriousness or interpretation when the temporal sequence of the test factor is unknown. When this problem arises the analyst may be able to use one of the aforementioned techniques to determine the most probable temporal sequence of the independent variable and the third variable test factor.

Tests for Spuriousness. Tests for spuriousness are of critical importance in survey research because they represent the most available means of controlling factors which can confound the interpretation of an observed relationship. Although Astin and Panos, Coleman, Thistlethwaite, Trent and Medsker, and Bachman usually tested for spuriousness when analyzing the impact of selected student background and school characteristics upon various criterion variables, the use of this testing procedure in the Analytical Review studies was inconsistent.

For example, Bachman observed that Southern Blacks attending segregated high schools, had lower self-concepts of their academic abilities than did Whites. The question, of course, was whether this difference reflected contrasting school

experiences or different social and psychological histories. To determine whether this relationship was in fact spurious, Bachman controlled for certain student background characteristics. According to Bachman:

> Southern segregated Blacks show slightly lower self-concepts of school ability than do Whites; however, once we account for family background and measured intelligence, it no longer appears that they underrate their academic ability - in fact, their self-concepts on this dimension are, if anything, relatively higher than those of Whites (1970, p. 103).

The test for spuriousness, then, not only resulted in the disappearance of the original relationship, but it produced a slight (insignificant) relationship in the opposite direction.

Astin and Panos' (1969) study provides another illustration. In this investigation academic ability, a student background characteristic, was found to predict career choices in certain fields. Students with superior academic records in high school were more likely to make stable choices or to change their choice to college professor, lawyer, physician and physical scientist. Consequently, in order to avoid a possible spurious relationship between selected college characteristics and career choice, the effects of these student background characteristics had to be taken into account. Thus, Astin and Panos statistically controlled for numerous student input variables when looking at the relationship between college characteristics and career choice and were able to conclude that liberal arts colleges diminished the students' interest in becoming lawyers and engineers while increasing their desire to become physical

scientists, social scientists, physicians and college professors.

Both Trent and Medsker and Flanagan and associates (1971) determined the concomitant relationship between levels of academic aptitude and socio-economic status and college attendance while holding sex constant, thereby learning that all three independent variables were related to college attendance independently of one another. They used the same procedure in determining the correlates of a variety of other criterion variables.

Coleman's analysis of the impact of school integration upon the educational achievements of the minority student also included tests for spuriousness. Coleman observed that in grades six, nine and twelve, the highest average test scores for Blacks were obtained from students who attended integrated schools. These results could reflect the impact of school integration upon the educational development of minority students. They could also reflect a selection bias; the Black students in this sample may be unrepresentative of Black students in general. That is, Black students attending integrated schools may come from a higher socio-economic class than Black students who attend segregated schools, and it may be this difference, rather than the effects of school integration, which accounts for their higher achievement scores.

Anticipating this rival interpretation, Coleman repeated the analysis controlling for socio-economic status. When the potentially confounding effects of social class were taken into account the differences remained; Black students in integrated

schools still performed better than Black students attending segregated schools. Consequently, greater confidence can be placed in the interpretation that school integration has a desirable impact upon the educational achievements of minority students.

Unfortunately, there were many other situations where this testing procedure was not consistently performed throughout the analysis. Flanagan et al. (1971) and Trent and Medsker, for example, found a strong relationship between students who took a college preparatory curriculum and subsequent college entrance. In neither case, however, were the input characteristics of the students statistically controlled so that the possible influence of the college preparatory program on actual attendance could be determined independently of the student background characteristics. In addition, Flanagan et al. found that students taking college preparatory courses were underrepresented among college "drop outs." On the basis of this information the authors conclude that:

> This offers some evidence that college preparatory programs in high school are preparing their students for college or are at least attracting those students who are most likely to be admitted to college and do well once they are enrolled (1971, pp. 8-10).

In terms of evaluating the impact of college preparatory programs it makes a big difference whether these programs are actually increasing the probability of college enrollment and success at college or are merely attracting students who would do well in college regardless of their exposure to a college preparatory

curriculum. Neither of these competing hypotheses can be supported or rejected without further analysis of input characteristics, eliminating the possible spurious relationship between high school program and college attendance.

In their study of Michigan State students, Lehmann and Dressel (1963) found that college attendance was statistically associated with certain attitude and value changes. Comparing the changes in value orientations of students who had completed four years of college with students who had withdrawn, the researchers concluded that college attendance was associated with the development of an "emergent value orientation" for males. In addition, college attendance was found to be related to changes in stereotypic beliefs for women; the longer a woman attended college the greater the probability that she would become less prejudiced and authoritarian.

These two relationships suggest that the impact of the college experience was a causal factor responsible for the value changes. However, before confidence can be placed in this interpretation, it must be demonstrated that the college persisters were not in some way predisposed to change independently of their college experience. In other words, tests for spuriousness should have been performed to ascertain whether the observed relationships between college attendance and value orientations were spurious due to the confounding effects of antecedent student characteristics which predisposed the college persisters to change their value orientations. Although Lehmann and Dressel had a considerable amount of data on student characteristics,

including income, parents' attitudes toward education, religious affiliation, size of home community, father's occupation and parents' education, none of these variables were used in a test for spuriousness. Consequently, the extent to which the original relationships are in fact spurious is open to debate.

Bachman (1970) missed several opportunities to test for spuriousness in his study. Earlier in this chapter it was noted that he discovered a strong positive relationship between students' reports of good family relations and the Crowne Marlowe scale of social approval. At that time it was pointed out that the desire for social approval was probably a causal variable responsible for the scores on the family relations index. In addition, subsequent analyses revealed that the Crowne Marlowe scale was highly correlated with impulse to aggression. Consequently, whenever a relationship was observed between the family relations or impulse to aggression index and a criterion variable, the relationship should have been tested for spuriousness by controlling for the desire for social approval. Bachman also reports a suggestive relationship between the family relations index and self-esteem. It may be, as Bachman suggests, that parents who show relatively high interest in their child's academic performance, friends, mealtime conversations, etc. will have children with favorable self-evaluations. However, the strong statistical association between family relations and self-esteem could also be due to the effects of a common antecedent variable which is highly associated with both family relations and self-esteem--the desire for social approval.

Husen's study provides another illustration of a relationship that should have been tested for spuriousness. Husen hypothesized that the mean level of mathematical achievement observed in a school would be related to the size of the enrollment at that school. As expected, the data indicated that the best performances in mathematics for younger students were found in schools with large enrollments. It is questionable, however, whether one can conclude that the size of the school is actually a causal factor in this relationship because the possible effects of student background characteristics or other school characteristics (e.g., urban vs. rural) on mathematical achievement were not introduced into the analysis.

Although the investigator must look for antecedent variables that are statistically associated with both the independent and dependent variables, it is not necessarily the case that every antecedent variable that is so related will produce a spurious relationship when statistically controlled. When the antecedent variable is sequentially related to the dependent variable (i.e., the antecedent test factor causes the original independent variable which in turn causes the dependent variable), controlling for the antecedent test factor will not produce a spurious relationship. In other words, if X causes Y, Y causes Z and X does not have an affect on Z that is independent of Y (the relationship between X and Z disappears when the effects of Y are statistically reduced) then the original relationship between Y and Z will not disappear in the partial relationships when the effects of X are controlled.

For example, if small classrooms offer greater individual instruction and individual instruction tends to increase the verbal achievement scores of the student, then statistically reducing the effects of small classrooms will not cause the relationship between individual instruction and verbal achievement to disappear. In other words, controlling the effects of the independent variable will not cause the relationship between the interpretation test factor and the dependent variable to disappear. For the same reason Blau was unable to produce a spurious relationship between self-image of age and the individual's conception of how significant others perceived his age when the effects of chronclogical age were statistically reduced; self-image of age was an interpretation variable in the relationship between chronological and significant others' conception of age.

In many of the Analytical Review studies the problem of testing for spuriousness was handled by introducing the variables into a multiple regression or multiple correlation equation on the basis of temporal sequence. That is, variables conceptualized to be antecedent in temporal sequence were introduced into the equation first. Thus, Coleman introduced student background characteristics into a multiple regression equation before looking at the impact of various school characteristics on verbal achievement. In a similar fashion, Thistlethwaite entered precollege characteristics into a multiple regression equation prior to analyzing the relationships between various measures of "college press" and student

motivation to seek advanced training. Astin followed the same basic procedure when he used multiple regression analysis to calculate the expected value of a criterion variable on the basis of student inputs and then subtracted this score from the actual value, thereby yielding a residual score which was presumably statistically independent of the input or antecedent variables.

In a critique of Coleman's study, Cain and Watts (1970) advocate the reverse procedure by recommending that school characteristics be introduced into the multiple regression or multiple correlation equation prior to the student input characteristics.[4] Basically, their recommendation is based upon three arguments. First, school and student characteristics are likely to be statistically associated with one another and with the dependent variable, which is often some measure of academic achievement or intellectual or personal growth in educational research. This is important because when two or more independent variables are associated with one another and with the dependent variable, there is usually a certain proportion of the explainable variance in the dependent variable that is shared. The explainable variance in the dependent variable which is shared or held in common by the different independent variables will be attributed to the first variable that is introduced into the regression equation. Thus, the

---

[4]See S. S. [illegible]s and H. M. Levin (1968a, 1968b) for similar critique [illegible]he Coleman study.

estimated main order or independent effect of the first variable will tend to be inflated while the influence of the remaining independent variables will be underestimated because the joint or common effect will be attributed to the first variable that is entered into the regression equation.

Second, Cain and Watts argue that the principal concern of educational researchers should be to identify variables which influence educational outcomes that can be manipulated by educational administrators. School characteristics, such as average size of classrooms, per pupil expenditures, etc. are easier to manipulate through educational policy than student input characteristics. Consequently, the researcher should focus upon school characteristics.

Therefore, according to Cain and Watts, school characteristics should be introduced into the multiple regression or multiple correlation equation prior to the student input characteristics, otherwise, the effects of school characteristics are likely to remain undetected or underestimated because of the joint or common impact they frequently share with student input characteristics.

Educational researchers have shown good judgment in rejecting the procedure advocated by Cain and Watts because it would unquestionably produce invalid inferences and interpretations. Entering school characteristics into a regression or correlation equation prior to student input characteristics completely ignores the temporal sequence of variables. Student input characteristics are temporally antecedent to the effects

of school characteristics, and consequently should be introduced into the equation first. If the effects of the school were to be analyzed without first considering the effects of student background characteristics, then the probability of observing a spurious relationship is greatly increased. Astin (1969) provides a dramatic illustration of this possibility. Without considering the influence of student input variables, Astin discovered that college characteristics accounted for approximately 20 percent of the variance in social science achievement. When these outcomes were adjusted for student input characteristics, however, the contribution of college characteristics to social science achievement plummeted to about 5 percent.

Of course, there is no reason why the analyst cannot provide both the inflated or maximum estimates and the conservative or minimum estimates of the impact of particular school characteristics on the criterion variables. This information would allow the educational administrator to assess the range of impact that could be expected by manipulating a given school characteristic.

Explicative Analysis in Education

Explicative analysis is the logical extension of causal analysis. Once a causal relationship has been detected, the researcher should attempt to elaborate this relationship by identifying factors which will interpret the relationship and make it more or less probable. Interpretation and specification analysis should therefore be an integral part of the

overall data analysis strategy of the educational researcher. The extent to which observed relationships were statistically interpreted and specified in the Analytical Review studies is assessed in the following section.

Interpretation Analysis. The contribution of a study is enhanced considerably when the analyst can identify the reasons why a particular relationship exists. This is especially true in educational research where knowledge of the underlying processes responsible for a relationship is necessary before sound policy decisions can be made. Statistical interpretation attempts to identify and clarify these underlying processes by introducing additional test factors into the analysis.

Although interpretation analysis can make important substantive contributions to educational research, the Analytical Review indicates that analysts were not systematically testing interpretation hypotheses. Although the studies reviewed contain attempts at statistical interpretation, for the most part they remain isolated attempts lacking the necessary integration with causal analysis to become a viable research strategy for understanding the dynamics of human growth and development.

Examples of interpretation analysis can be found in Coleman (1966), Shaycoft (1967) and Bachman (1970). Coleman found that minority students attending integrated schools scored higher on achievement tests than did minority students attending segregated schools even when the effects of social class were statistically reduced. This important finding

merited further investigation to determine why integrated schools had this effect.

Several plausible interpretations were advanced to explain this observation. Minority students might have achieved higher scores in integrated schools because predominately White schools have better physical facilities, higher quality teaching staffs or more remedial courses. Differences within the peer environments of integrated and segregated schools could also have accounted for the greater achievement of minority students in integrated schools. The peer group in integrated schools compared to segregated schools may have encouraged and rewarded academic excellence to a much greater extent thereby motivating the minority students to perform better scholastically.

To determine which of these competing hypotheses was most credible, Coleman performed an interpretation analysis which has become a classic example of this testing procedure in educational research and clearly illustrates that important information can be obtained by systematically searching for interpretation variables. The results of Coleman's analysis revealed that it was the influence of the peer group, rather than the impact of the physical resources of the school or the quality of the teaching staff, which accounted for the relationship between school integration and academic achievement. In the words of Coleman:

> The higher achievement of all racial and ethnic groups in schools with greater proportions of white students is not accounted for by better facilities and curriculum in their schools. . . . The higher achievement of all racial and ethnic groups in schools with greater proportions of white students is largely, perhaps wholly, related to effects associated with the student body's educational background and aspirations (p. 307).

Thus, when the differential effects of various school resources were statistically controlled, the observed differences in the academic achievement of minority students attending integrated and segregated schools remained. However, when the effects of the educational aspirations and achievements of the student body were statistically reduced, the differences which had previously existed in the educational achievements of these students disappeared. This is strong evidence in support of the interpretation that minority students tend to perform better in integrated schools because of the differential characteristics of the student subculture. Furthermore, as Coleman notes:

> This means that the apparent beneficial effect of a student body with a high proportion of white students comes not from racial composition per se, but from the better educational background and higher educational aspirations that are, on the average found among white students (p. 307).

Shaycoft provides several notable illustrations of interpretation analysis in her report of Project TALENT data. Shaycoft was interested in identifying the magnitude of the direct effect that socio-economic status had on grade twelve achievement. Consequently, the analysis of the influence of social class upon twelfth grade achievement was repeated after

the effects of student aptitude, grade nine achievement, courses in high school and college plans were statistically controlled. According to Shaycoft:

> After the effects of their various causative (or possibly causative) factors . . . have been eliminated from the socio-economic variable statistically, the part correlations of the residuals with grade 12 test scores are negligible (1967, pp. 8-24).

Although this test was conceptualized as an attempt to estimate the direct causal relationship between social class and academic achievement, it also serves as an example of interpretation analysis. The results of this investigation indicated that the original relationship between socio-economic class and academic achievement was due, in large part, to the mediating influence of those variables that were statistically controlled in the analysis.

Bachman's analysis of the racial differences in the responses to a job ambitions index illustrates a systematic attempt to identify an interpretation variable. Bachman found that the scores of Black students, regardless of whether they attended integrated or segregated schools, were lower than Whites on a scale presumed to be measuring the job ambitions of the student. In search of an interpretation, Bachman adjusted the scores on the basis of seven background characteristics and scores to the Ammons Quick Test of Intelligence. Although the racial differences did diminish in this phase of the analysis, the basic relationship remained; Blacks had lower job ambitions than Whites.

The ambitious job attitudes index was composed of two major dimensions; attitudes toward jobs "that pay off," and attitudes toward jobs "that don't bug me." Bachman surmized that a separate analysis of these two components of the original scale might reveal the reason for the racial differences in the scores. Consequently, the analysis was repeated using these two dimensions as separate criterion variables. According to Bachman, the results of this investigation were definitive:

> There are scarcely any racial differences in preferences for "a job that pays off" . . . racial differences do appear when we consider preferences for "a job that doesn't bug me." Along this dimension we find integrated blacks more than one-third standard deviation higher than whites; for northern segregated blacks the difference exceeds one-half standard deviation, and for southern segregated blacks the difference reaches three-quarters of a standard deviation (1970, p. 147).

Thus, the Black students in this sample were as equally ambitious as Whites for "jobs that paid off." However, the Black students were less tolerant than Whites of "jobs that bug me." This led Bachman to conclude that:

> The young black high school student probably knows better than most whites what it means to have "a job that <u>does</u> bug me," and avoiding that sort of job seems more important to him than to the average white. In our view, it is likely that some of the items on the "job that doesn't bug me" scale mean something very special to black respondents, and that this, more than anything else, accounts for the racial differences we have observed here (1970, p. 147).

As previously noted, there are not many examples of interpretation analysis in the Analytical Review studies. In some cases, this was due to the lack of data. In many cases, however, the analyst failed to perform an interpretation analysis when the necessary data were available. The work of Trent and Medsker, Astin and Panos, Lehmann and Dressel, and Hilton are cases in point.

Trent and Medsker observed that socio-economic class was moderately associated with college persistence in the expected direction; the higher the social class standing of the student the greater the probability of persisting in college. In addition, academic achievement in high school and the amount of income earned from part time employment during college were also predictive of college persistence. College students who had low academic achievement in high school or worked part time for over half of their income were more likely to withdraw from school than students who had high academic achievement in high school or did not work for over half of their annual income. Since data regarding two principal interpretation variables were available for each analysis--the academic achievement of the student and his employment--Trent and Medsker may have been able to determine why lower class students withdrew from college at a greater rate than upper and middle class students. Unfortunately, however, an interpretation analysis was not performed and as a result it is uncertain whether lower class students withdrew from college because they had lower academic ability or because they were employed.

Astin and Panos missed several opportunities to statistically interpret observed relationships. For example, they discuss their observations of environmental effects by type of institution (pp. 141-145) and the effects of specific environmental characteristics (pp. 145-147) under separate headings in their report and never attempt to integrate these two groups of findings through a systematic strategy of interpretation analysis. Instead, the researchers speculated that:

> . . . some of the environmental effects observed in particular "types" of institutions may be wholly or in part a consequence of differences among the institutions in some of the (environmental) characteristics described below (p. 145).

The "may be wholly or in part a consequence of" clause, however, represents an empirical question which could have been tested with the data that were available for analysis.

Thus, Astin and Panos report that universities and liberal arts colleges had very different rates. Attrition at the liberal arts colleges was substantially lower than would be expected on the basis of student input characteristics. In contrast, the dropout rate at the university was greater than would be expected. The researchers also report that colleges with cohesive peer environments had much lower dropout rates than were predicted from student input characteristics while colleges with fragmented peer environments showed the reverse pattern. These relationships suggest the hypothesis that liberal arts colleges have lower dropout rates than universities because their peer environments are more cohesive. Clearly,

this is a plausible interpretation that should have been tested. Unfortunately, the investigators failed to do so. Instead, Astin and Panos postulated the followint alternative explanation:

> Although there are many possible explanations for the sharp contrast between liberal arts colleges and universities, one interesting hypothesis is suggested by the fact that these two groups of institutions differ markedly with respect to two environmental factors, Familiarity With The Instructor and Concern For The Individual Student. Perhaps the university professor, who spends relatively little time with his students and much time in pursuing his own scholarly interests, provides a relatively poor role model in comparison with the college teacher, who often takes a more personal interest in his students (p. 142).

Even here, the investigators failed to test their interpretation. They could have statistically controlled for the items "Familiarity With The Instructor" and "Concern For The Individual Student" to determine if these two variables did, in fact, interpret the relationship in question. Since they failed to perform this test, their interpretation remains an unconfirmed speculation.

Lehmann and Dressel (1962) also missed an opportunity to clarify the underlying processes of a major relationship reported in their study of Michigan State students. They found that religion was highly associated with a student's value orientation and dogmatism score. The apparent influence of the student's religious training was manifest in two related observations. First, Catholic students were more stereotypic and dogmatic than Protestants or Jews and were also more

traditional-value orientated. Jewish students, on the other hand, were more emergent in their value orientation than either Catholics or Protestants. Secondly, students who had previously attended public high schools were less dogmatic, stereotypic and more likely to be emergent in their value orientation than students who had attended parochial schools.

The researchers also report that students' social class was associated with their value orientations and dogmatism scores. Students whose parents had a high level of education were less stereotypic, dogmatic and had more emergent value orientations than students whose parents had little formal education. In addition, students whose fathers had a high occupational rank were more emergent in their value orientations, less stereotypic and dogmatic than students whose fathers had a low occupational rank.

Social class can often statistically interpret relationships between religious affiliation and various criterion variables. In this case, social class might have interpreted the relationship between religious affiliation and critical thinking, values, and attitudes. Thus, Lehmann and Dressel could have determined whether the effect of the student's religious background upon his value orientation and attitudes was due primarily to the influence of social class, or the more direct effects of religious training. Although a simple test of interpretation would have clarified this issue, the analysis was not performed and important information was lost.

Evans and Patrick (1971) conducted an analysis of high school dropouts from data derived from Hilton's study. In this investigation the analysts observed the expected relationship between the age of the student in the fifth grade, his scores on various SCAT and STEP tests, and withdrawal from high school. Fifth grade students who were later to drop out of high school tended to be approximately one year older than their classmates and obtain scores on achievement and ability tests which were significantly lower than those students who continued through the eleventh grade.

Certainly a critical question to ask at this point is why the older students in the fifth grade had a much greater probability of withdrawing from high school than the younger students. Age and achievement scores were both highly correlated with the criterion variable. In addition, the achievement scores only explained about 1 percent of the variance in the dependent variable after the age of the student was introduced into the correlation equation. This suggests substantial collinearity between the age of the student and his test scores. Thus, it seems plausible that the academic achievement of the student could serve as an interpretation test factor accounting for the relationship between age and withdrawal from high school. In other words, on the basis of the existing data it could be argued that the students who were older in the fifth grade were less capable of performing well in high school and consequently were more likely to become frustrated, disillusioned and finally to drop out of school altogether.

Evans and Patrick, however, appear uncertain of the association between student age in the fifth grade and high school persistence:

> Dropouts are nearly a year older in fifth grade than their non-dropout peers. The dropouts may have failed one or more grades or may have started school later than their peers. . . . Another possible explanation is that grade retention acted to cause dropping out, rather than simply predicting that a student would eventually drop out of school. The important finding is that the age discrepancy is apparent as early as the fifth grade (p. 131).

Although the relationship between student age and withdrawal from high school was an important finding, it is also important to determine why age is a crucial factor in predicting the criterion variable. Interpretation analysis could have answered this question thus clarifying the underlying processes responsible for the relationship in question.

Educational researchers in general have been negligent in conceptualizing, measuring and testing interpretation variables. As a result, the systematic use of statistical interpretation has not been rigorously pursued in educational research. This is indeed unfortunate because the results of interpretation analysis can be tremendously enlightening to the researcher as well as to the educational policy-maker.

Specification Analysis. Specification analysis is another form of statistical elaboration which identifies situations or conditions that weaken or strengthen the association between two or more variables in a causal relationship. The results of systematic specification analysis can provide information

that is of equal if not greater value than that obtained from an interpretation analysis.

For example, in order to achieve optimal allocation of school resources in terms of maximum impact upon the student population, it is necessary to determine which students will receive the greatest benefit from exposure to a particular school resource. The search for specification [illegible]t factors will often identify variables that strengt[illegible] or weaken the association between a school characteristic and a particular educational outcome. Frequently, these factors are student background variables. Thus, specification analysis can provide the investigator with valuable information concerning the differential impact that certain school characteristics have upon different student subgroups.

Coleman's study illustrates the importance of specification analysis in educational research. Coleman found that the student's academic achievement was only marginally affected by differences between schools. However, after a specification analysis was performed in which comparisons were made between racial and ethnic groups, important inter-school differences emerged. In general, the variation in academic performance that existed between schools was noticeably larger for minority students than for majority students. This finding suggested that the academic performance of minority students was more sensitive to the impact of different school environments than was the academic achievement of majority students. As Coleman writes:

> Indirect evidence suggests that school factors make more difference in achievement for minority group members than for whites; for Negroes, this is especially true in the south. This result suggests that insofar as variations in school factors are related to variations in achievement, they make the most difference for children of minority groups (p. 297).

After performing a test for spuriousness by controlling for a number of student background variables, Coleman introduced additional data to indicate that the academic achievement of minority students was, in fact, more sensitive to the impact of various school characteristics. For example, the magnitude of per-pupil expenditures, the quality of the teaching staff, and numerous characteristics of the school facility, including the size of the school and the presence of laboratories, extracurricular activities and guidance programs were found to have a larger impact upon the academic performance of minority students than majority students. Thus, Coleman uncovered an important specification test factor that strengthened the association between various educational experiences and academic achievement.

In a similar fashion, Astin and Panos tested for specification in their analysis of the quality of undergraduate institutions and the student's intellectual achievement. The researchers state the major hypotheses of their investigation in the following manner:

> Stated in positive terms, the general hypotheses tested in this analysis were as follows;
>
> 1. The academic excellence of the undergraduate institution - as defined by the level of ability of the student body, the level of the institution's financial

> resources, and the degree of academic competitiveness in the college environment - has a positive effect on the undergraduate student's intellectual achievement.
>
> 2. The extent of the positive effect of institutional quality on intellectual achievement is proportional to the student's academic ability (1969, p. 72).

The second hypothesis incorporates a test for specification. That is, Astin and Panos propose that the effects of institutional quality will statistically intereact with the student's academic ability to produce a higher association between the quality of the educational institution and the student's intellectual achievement when the student has high academic ability.

The results of this analysis, however, provided little support for either hypothesis. The academic excellence of the school was not strongly or consistently associated with academic achievement after the effects of student background characteristics were statistically controlled. Furthermore, the academic ability of the student did not significantly influence the strength of the association between the quality of the school and the student's academic performance.

Of course, the value of specification analysis is not restricted to policy orientated research. The results of systematic specification analysis can make an equally important contribution to the development of learning and socialization theory, as illustrated by the work of Kagan and Moss (1962). In their longitudinal study of psychological development, Kagan and Moss consistently tested for specification by comparing the similarity of behavioral patterns between different time periods

for both males and females. Many of their findings strongly support the hypothesis that early sex role identification has important implications for adult behavior. For example, Kagan and Moss report that a passive, in contrast to a retaliatory reaction to frustration, was highly stable for boys and girls during the first ten years of life. However, early passivity in males was essentially unrelated to adult behavior while it was moderately related for females. According to Kagan and Moss:

> The primary reason for this lack of continuity in males is the development of conflict over passive and dependent behavior. A passive orientation to problems is inappropriate for the male role. . . . This conflict, which does not swell to such strong proportions in middle class girls, leads to minimal continuity between childhood and adult dependency for males (1962, p. 58).

Thus, by focusing upon sex as a specification test factor, Kagan and Moss were able to contribute valuable information supporting the notion that differential cultural expectations for dependency and passivity in males and females will influence the pattern of psychological growth and development.

Bachman (1970) also performed specification analysis in his panel study of adolescent boys. He originally found that majority and minority students attending integrated schools had similar self-concepts of school ability, while Blacks in segregated schools had somewhat lower self-concepts. However, the relationship was dramatically reversed when the scores were adjusted for a number of student background characteristics including social class and general intelligence. There was a

pronounced tendency for Blacks attending either segregated or integrated schools to have relatively higher self-concepts of school ability than Whites. Furthermore, this same basic pattern was repeated for self-esteem. Southern segregated Blacks had self-esteem scores that were similar to Whites, but after controlling for background characteristics and general intelligence, their adjusted scores were higher than Whites. According to Bachman:

> It is frequently assumed that Black Americans, as a result of centuries of slavery and discrimination, have lower self-esteem than whites. This may be true of adults, but our data lead us to question this assumption as applied to young men in high school. . . . Our view is that the fairly high self-esteem scores for Black respondents represent a real feeling of self-worth (1970, p. 199).

The theoretical contribution of Bachman's analysis was enhanced by systematically testing for specification. The results of this analysis not only uncovered an important relationship that was suppressed in the original analysis, but it provided valuable information concerning the possible changes in self-concept that may be occurring among minority students today.

Despite its obvious value, researchers have not systematically and uniformly introduced specification test factors into their analyses. Astin and Panos, for example, report that colleges which have a relatively large percentage of students who work for pay have considerably higher dropout rates than were predicted from student input data. The researchers also had extensive information on the college environment; measures of the cohesiveness and competitiveness (vs. cooperativeness)

of the peer environment, various indicators of the classroom environment such as student involvement in the class and severity of grading practices, data on the administrative and physical environment of the school, student's subjective impressions of the college environment, the instructor's concern for the student, flexibility of the curriculum, and the degree of academic competitiveness.

Valuable information could have been obtained if Astin and Panos had determined if these environmental factors specified the relationship between student employment and college persistence. High school guidance counselors, for example, could use this information to advise students who anticipate working during college that they would have a greater probability of graduating from particular types of schools (e.g., those characterized by low academic competitiveness, liberal grading practices, or a cohesive peer environments). Unfortunately, tests were not conducted using environmental data as possible specification test factors.

The study by Trent and Medsker provides a number of similar illustrations. The researchers report, for example, that two-year college transfers had a statistically higher rate of attrition than native students. A considerable amount of data on student background, family, and school environmental variables were available for analysis. Consequently, Trent and Medsker were in a position to identify situations or experiences that would increase the likelihood of transfer students persisting in college. The researchers, however, did not specify

or interpret this relationship, and again valuable information was lost.

The dearth of systematic specification analysis is evident in Project SCOPE. Tillery and associates (1972) report that in a subsample of students drawn from California, Illinois, Massachusetts and North Carolina, minority students (excluding Oriental Americans) tended to have lower educational aspirations than White students. The researchers did not attempt to elaborate this relationship, however, by performing an additional analysis designed to identify specific educational experiences that contributed to the lower aspirations of minority students.

In addition, the data from this subsample indicated that junior colleges attract a similar number of students scoring at all four levels of an intellectual predisposition (IPD) test consisting of items from the Thinking Introversion, Theoretical Orientation, and Autonomy scales of the Omnibus Personality Inventory (Center for the Study of Higher Education, 1962). It would have enhanced the contribution of this analysis if the researchers had identified the factors that increased the probability that a student with a high intellectual predisposition would enroll in a junior college. Again, although the data were available an important test for specification was not conducted.

Opportunities to perform specification analysis were also missed in Project TALENT. Flanagan et al. (1962), for example report that the size of the senior class and the average class

size were not highly associated with high school achievement (correlated below .20 with school achievement). No attempt was made to identify student background variables or other school characteristics that would increase or weaken the magnitude of these relationships. Consequently, no information was provided concerning the type of student who would benefit most from exposure to small classrooms.

As a final example, Thistlethwaite reports that the student's disposition to seek advanced training was:

> . . . strengthened by association with peers having high educational aspirations, favorable teacher evaluations of college performance, winning social recognition for intellectual achievement, participation in Honors Programs and graduate-level courses, and by under-graduate participation in research programs and projects (1965, pp. 91-92).

The above relationships indicate the direct or independent impact of these college experiences upon the criterion variable as determined by a multiple regression analysis. However, these findings pertain to students in general and not to particular subgroups of students. That is, Thistlethwaite determined that there was a tendency for all students to seek advanced training if their peer groups had high academic aspirations or their teachers evaluated their academic performance favorably. Consequently, although Thistlethwaite reports that these college experiences "strengthened" the student's disposition to seek advanced training, he did not perform a true specification analysis.

While it is important to ascertain the independent effects that college experiences have upon the student's desire to seek

graduate training, it is also important to determine which students will benefit most from exposure to these experiences. In other words, what factors strengthen or weaken the association between participating in honors programs or receiving a favorable teacher evaluation and student disposition to seek graduate training? These questions are truly representative of the types of questions asked in a specification analysis. Thistlethwaite did not attempt to answer these questions in his investigation.

Educational researchers in general, and the Analytical Review investigators in particular, have not systematically tested for specification or interpretation. Instead, too often they report nothing more than multiple regression coefficients that illustrate a series of relationships involving only independent and dependent variables.

While it is generally recognized by theoreticians and methodologists alike that the ability to accurately predict a specific educational outcome will yield valuable insights into the dynamics of the phenomenon in question, it is nonetheless important to remember that these causal or predictive relationships are not the only or necessarily even the most significant source of information.

The most valuable scientific contributions are those which not only identify causal or predictive relationships but also explicate these relationships by analyzing additional variables that are not conceptualized as independent variables but as elaborating testing factors. The discovery of a causal relationship does not signify the end of systematic inquiry, but

the beginning. It is critically important to the advancement of educational research that analysts introduce and subsequently test hypotheses which identify (1) the reasons why a specific relationship exists (interpretation) and (2) the conditions that maximize and minimize the strength of the relationship (specification). A serious effort should therefore be made to supplement the current orientation of simply predicting specific educational outcomes with one that emphasizes the need to explicate predictive or causal relationships with interpretation and specification analysis.

CHAPTER IV

STATISTICAL MODELS IN IMPACT ANALYSIS

In Chapter II, causal and explicative analyses were discussed in theoretical terms, and Chapter III examined these forms of data analysis in the Analytical Review studies. Chapter IV extends this theme further by examining impact analysis in education. Two major problems are discussed in this chapter. The first deals with the conceptualization and measurement of impact. This problem is discussed in the first section. Section two deals with the problem of selecting an appropriate statistical model for analyzing impact. The final section discusses multiple regression analysis, the most frequently used statistical technique in the Analytical Review studies. The mathematical model is described, along with relevant topics that relate to the use of the multiple linear regression model. Correlation analysis was also employed in a number of the Analytical Review studies. The use of this statistical test in educational research is briefly discussed at the end of this chapter.

The Conceptualization and Measurement of Impact

Impact analysis may be defined as a form of systematic investigation in which the researcher attempts to determine the impact or effect of a particular program or institution upon a predefined target population. In education, this

type of analysis occurs when causal relationships are identified between particular school characteristics and educational outcomes. Exposure to a particular school characteristic cannot have an impact upon the student, in terms of a specified criterion measure, unless a causal relationship exists between the two variables.

Two major conceptualizations of impact have emerged in educational research. The effects of the educational institution are often conceptualized in terms of change or gain scores. In addition, impact is frequently seen as variation in outcome. Both of these conceptualizations are discussed below.

Impact in Terms of Change

The vocabulary of impact analysis is frequently couched in a conceptual scheme that identifies change as the central phenomenon under investigation. Thus, impact is evaluated on the basis of the amount of change in a specific criterion measure that can be attributed to a particular school characteristic which is conceptualized as an intervening treatment variable that is partially responsible for the observed change.

Trent and Medsker's study can be used to illustrate this conceptualization of impact. To analyze the effects of college persistence on personality development, the researchers used the scores to the Social Maturity scale of the Omnibus Personality Inventory obtained during the freshman (1959) and senior (1963) year to create the following categories of change:

A. Exceptional changers: students with change scores falling three-fourths standard deviation or more above the average change score.

B. Average changers: students with change scores falling within three-fourths standard deviation above or below the average change score.

C. Negative changers: students with change scores falling at least three-fourths standard deviation below the average change score.

Table 9 contains the results of their analysis.

Table 9

Illustrating a Relationship in which Impact Is Conceptualized in Terms of Change[a]

College Persisters and Withdrawals in Each Change Group, in Percentages

Pursuit groups	(N)	Change groups		
		Exceptional	Average	Negative
Men				
Persisters	(723)	37	48	15
Withdrawals	(105)	26	47	27
Women				
Persisters	(578)	40	48	12
Withdrawals	(195)	21	49	30

[a]Reproduced from Trent and Medsker (1968, p. 187, Table 54).

The data in Table 9 appear to indicate that college persistence had an impact upon the amount and direction of change in the scores to the Social Maturity scale. Students who withdrew from college were less likely to be "exceptional changers" and more likely to have their scores regress between 1959

and 1963 than college persisters.

There are a number of serious methodological problems associated with gain scores that weigh heavily against their use in educational research. The most serious problem is that the change score is often artifactually dependent upon the initial score. This tendency is manifested in two related problems; "regression effects," and "ceiling and floor effects." Regression effects refer to the tendency for initially extreme scores to regress toward the mean score on subsequent measurements regardless of the effects of the treatment or independent variable.[1] Thus, Trent and Medsker report the effects of regression toward the mean when they observed that:

> On the Social Maturity scale, the lower students scored on the scale in 1959, the more likely were their scores to change significantly in a positive direction in 1963 (1968, p. 188).

Parenthetically, it should be noted that the college persisters in the Trent and Medsker study initially scored higher on the Social Maturity scale than the withdrawals (p. 189). Consequently, the relationship between college persistence and Social Maturity reported in Table 9 is even more impressive than the researchers acknowledge.

---

[1]The reasons for regression effects may be found in Bereiter (1963), Bohrnstedt (1969), Campbell and Clayton (1961), Campbell and Stanley (1963), Garside (1956), Lord (1956,1958,1963), Maccoby (1956), Maccoby and Hyman (1959), Thomson (1924,1925) and Thorndike (1924).

A number of recommendations have been made to correct for the biasing influence of regression effects. Most involve controlling or adjusting for initial position when analyzing gain scores. However, none of these methods provide a general or satisfactory solution to the problem.[2]

Ceiling and floor effects represent a similar type of measurement distortion. When a change score is calculated by subtracting an earlier score from a later score ($X_2 - X_1$), it is clear that the amount of change is dependent upon the magnitude of the initial score. Thus, a respondent cannot increase his score if he initially scored at the top of the scale (ceiling effect), or lower his score if he initially scored at the bottom of the scale (floor effect). In general, the more extreme the score, the less the probability that the score will become more extreme on a subsequent measurement.

Again, there is no adequate solution to this problem. Hovland, Lumsdaine and Sheffield (1962) have suggested, however, that in addition to calculating the magnitude of a specific change pattern for the entire sample, the analyst recalculate the score by excluding those respondents who were not capable of changing in a specific direction because of ceiling or floor effects. This data would provide an inflated estimate of change which could then be compared to the original

---

[2]See Bereiter (1963), Bohrnstedt (1969), Cronbach and Furby (1970), Gurin and Katz (1966), Hites (1965), Lord (1956,1958,1963), McNemar (1958), Skager, Holland and Braskamp (1966), Thomson (1924,1925), Tucker, Damarin and Messick (1966), Webster (1963,1968), Wiseman and Wrigley (1953), Werts and Linn (1970), and Zieve (1940).

measure to determine the implications of ceiling and floor effects.

In an excellent review and critique of change score measures, Cronbach and Furby (1970) identify four purposes of computing gain scores:

1. To provide a dependent variable in a study of behavior change.
2. To provide a measure of rate of growth.
3. To provide an indicator of deviant development (e.g., underachiever).
4. To provide an indicator of a construct having theoretical importance (e.g., self esteem defined as the difference between ratings of self and ideal self).

In each case they present evidence strongly suggesting that the use of raw gain scores is inappropriate. They present a number of alternatives to the gain score approach, usually involving the regression of the posttest score on the pre-test and other variables in order to estimate a "true gain" score. In short, Cronbach and Furby conclude that:

> investigators who ask questions regarding gain scores would ordinarily be better advised to frame their questions in other ways (1970, p. 80).

Impact in Terms of Variation in Outcome

Methods for analyzing school impact have been suggested which do not directly measure change or stability in terms of calculating gain scores, but look at the amount of residual variation in the criterion variable that can be explained by particular school characteristics as evidence of impact.

The input-output model (discussed in the following section) is the most common application of this conceptualization, and will serve to illustrate the logic of defining impact as variation in outcome.

In using the input-output model, an expected value on the criterion variable is calculated on the basis of student input and other nonschool factors. The expected value is then subtracted (statistically controlled) from the observed or actual measure, leaving a residual variance in the criterion variable that is statistically independent of the nonschool factors included in the analysis. The amount of variance in the residual value that can be explained by a particular school characteristic is used as the criterion for assessing impact. Astin and Panos, for example, initially regressed a criterion measure of social science achievement on a number of student input variables, and then determined the extent to which various college characteristics could explain these residualized scores. The results of their analysis indicated that the college environment had very little impact upon the criterion variable.

The application of the input-output model illustrates the principal advantage of defining impact in terms of variation in outcome. This conceptualization does not require a direct measurement of change. Impact is inferred by the magnitude of the explained residual variance rather than by the magnitude of the explained variance in test-retest scores.

Consequently, this conceptualization avoids the problem of obtaining corrected gain scores, which has been a major problem in impact analysis. The input-output model, however, is not the only available method for analyzing impact in terms of variation in outcome. In the next section, covariance analysis, direct and indirect methods of standardization and path analysis are introduced as alternative strategies. The methodological constraints of each are also discussed.

Statistical Models for Impact Analysis

In general, the Analytical Review studies gathered responses from a large number of subjects and collected data on a wide range of variables. The Coleman study, for example, examined over 50 student variables and more than 100 school variables. The student variables investigated in the Analytical Review studies included: vocational development (Astin and Panos; Hilton), career patterns (Super), educational aspirations (Astin; Thistletwaite; Tillery et al.), intellectual growth (Hilton; Jones et al.; Kagan and Moss; Shaycoft), academic achievement (Husen; Flanagan), and multiple combinations of the above (Bachman; Coleman; Lehman and Dressel; Tillery et al.; Trent and Medsker).

A very general analytical framework was applied to the resultant data. Where there were a priori hypotheses, these were insufficiently precise to permit the application of a stronger statistical model.

The statistical models used in the reviewed studies and the models to be presented here are very general and are applicable in a wide variety of situations. As a consequence, no one of them is necessarily the "best" model to apply in any particular circumstance. The assumptions about the data which are required for the analysis are usually quite minimal: (1) the observations must be stochastically independent in order to apply the usual estimation techniques; and (2) the underlying form of population distributions must be normal to apply the usual tests of hypotheses. (See the following section for amplification of these points with respect to regression analysis.)

In certain disciplines, the data do not conform to these requirements. For example, in econometrics, the observations may be serially correlated. The analysis may proceed, however, if: (1) the data are modified such that the assumptions of the statistical model hold; or (2) a new statistical model is developed to accommodate the known structure of the data.

Path analysis (which will be discussed in more detail below) is an example of a statistical model developed for a specific context. It was originally developed by Wright (1931) to represent the expected outcomes of breeding experiments in animal husbandry research. The genetic models for cross-breeding led to very specific predictions about the values of coefficients in the path diagrams which could then

be verified by appropriate collections and analyses of data. In a sense, sociological researchers who have rediscovered path analysis have stood it on its head by inferring the values of the path coefficients from the analysis of the data and using such values as a basis for modification of their theory.

As educational phenomena become more clearly understood and conceptualized, the models for analysis of data will have to respond to the stronger theoretical framework by losing some of their generality and providing better data analysis. (See the discussion of Olkin's paper "Correlations revisited" in Stanley, 1967, p. 133, in particular.)

The implication is that substantive theory has far to go in this area before it will yield hypotheses which are testable using specifically constructed data collections and statistical models. In the study of "impact" educational research is not yet ready to adopt "strong inference" as a mode of inquiry (Platt, 1964). In this spirit, the present section will focus on analytic models which will help the researcher to isolate variables of importance and to construct theories which will be of greater utility.

Models for Data Collection and Analysis

A major problem in impact analysis in education is separating the effects of the school from the effects of student background characteristics and other nonschool factors. Three basic models which may be employed to

investigate the impact of the school are: (1) the Causal-Comparative Model; (2) the Input-Output Model; and (3) the Process Model.[3] A discussion of the types of data collection which may be used, the types of research questions the model can answer and the types of analysis appropriate for the data will be presented for each of these models.

The Causal-Comparative Model

Clearly, the simplest approach to assessing the impact of a particular phase of the educational process would be to compare students who have had that educational experience with those who have not. Indeed, if it were possible to randomly assign subjects to treatment conditions (e.g., receive a four-year college education as opposed to receiving none), the entire, highly developed machinery of experimental design could be brought to bear on the study of impact and our understanding of the phenomenon would advance apace. Unfortunately, the impossibility of random assignment makes our inferences less secure. Thus, if the Causal-Comparative Model is used, some method will have to be found to control for those differences in student background (e.g., precollege) characteristics.

The structure of the data collection for a causal comparative analysis resembles the experimental design paradigm except for the nonrandom assignment of experimental units

---

[3] Averch et al. (1972) refer to both the Input-Output Model and the Process Model. However, they identify the Process Model with experimental (mostly psychological types) of investigation, whereas it will be used in a much different way in the present context.

to treatments. The treatments of interest are characteristics of the educational experience whose impact upon students is to be assessed: Denominational or Nondenominational College, Public or Private, Large or Small, Liberal Arts or Engineering Emphasis. Within each treatment combination (e.g., denominational, private, small, liberal arts colleges would be one of sixteen possible combinations from the above classifications) a random sample of n units should be taken. In this case, the unit is the college as a whole and a small sample of students within that college serves to define the average values of the student characteristics. The analysis will be performed on the averaged values for each college sampled rather than on the individual students.

In the example comparing bussed to nonbussed students (presented in the next section) the individual would seem to be the appropriate unit of analysis. When entire groups of students are bussed from one school to another and remain together at the receiving school, however, the group is a better unit to be compared to classroom groups or other groups that remain behind.[4] By keeping the sample size (number of colleges) equal, the researcher assures that his estimates of the contrasts or differences between levels within one way of classification (called a "main effect") remain

---

[4]The reader is referred to Glass and Stanley (1970, section 19.11) for an excellent discussion of the appropriate unit of analysis.

independent of the other main effects and of the "interaction effects" (the joint action of levels on two or more ways of classification). If equality of sample size is not possible, then the effects become dependent upon one another and statistical testing of the effects which must be performed in stepwise fashion may yield equivocal results. For example, one may not be able to determine whether it is a size effect or a denominational effect which is responsible for certain features of the data. This becomes a great problem, particularly with respect to survey data based on probability samples where the data cannot be expected to yield equal cell sizes when cross-classified by some arbitrary scheme of interest to the researcher. The researcher must determine whether his arbitrary scheme of classification is sufficiently important to warrant that it dictate the sampling plan. He must realize that, in maximizing the precision of his experiment relative to his scheme, he is limiting the overall utility of the data collection by building in certain imbalances (e.g., the proportion of denominational engineering schools is larger in his sample than it is in the population so that applying other classifiers not used in the selection of the sample will have a built-in bias). The methods of "standardization" discussed below will enable the researcher to circumvent this difficulty to a certain extent.

At this point, the problem of controlling for background characteristics must be examined more closely. One analytic

technique which has been suggested for this purpose is the analysis of covariance. In the true experiment (where experimental units are randomly assigned to treatments) the analysis of covariance serves two purposes: (1) the reduction of error variance which enables the researcher to more precisely estimate the size of his effects and makes his tests of significance more powerful; and (2) the control of bias in the assignment of units to treatments. In the experimental setting, the second feature is considered less important than the first because random assignment itself will tend to control for bias. In the causal comparative study[5] the analysis of covariance is most favored for its bias-reducing property. However, the interpretation of the results of such an analysis is subject to many qualifications.

If a researcher sets out to investigate the impact of bussing on non-White students who are bussed to predominantly White schools, he may use the students who remain at the non-White school as a control group. If the bussed students are volunteers, they may be quite different in background characteristics from those who are not bussed. In particular, their pretest scores on a standardized achievement test administered in the fall may be higher. If their posttest scores are also higher, the problem of determining the impact of bussing can be circumvented by analysis of covariance.

---

[5]The Causal Comparative study is similar to the Static Group Comparison described in Campbell and Stanley (1963).

The analysis of covariance proceeds by first estimating the relationship between the background variable, in this case the pretest, and the dependent variable (the posttest) within each group. Providing that the relationship has the same form from group to group (assessed by the test of homogeneity of regression), then a combined estimate of the relationship is formed and the dependent variable mean for each group is adjusted by using the background variable (or covariate) mean for each group and the estimated relationship of covariate to dependent variable. The contrast or difference between the adjusted means is tested for significance. Figures 6 and 7 indicate how this analysis works for the example of bussing:

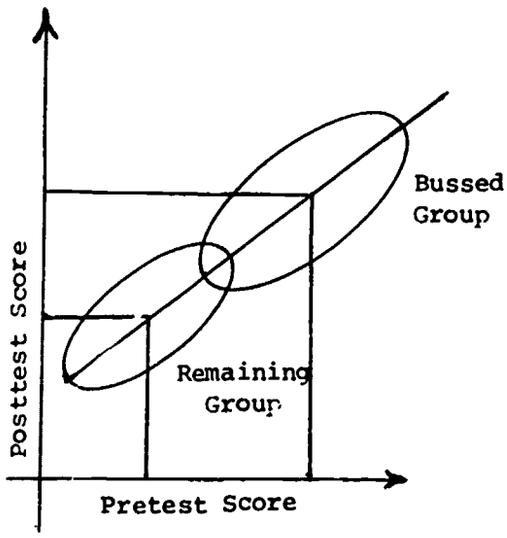


Figure 6. Covariate Adjustment Completelv Eliminates the Effects of Bussing

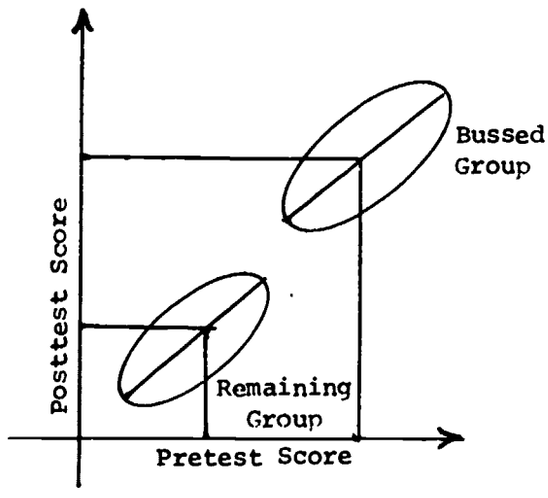


Figure 7. Covariate Adjustment Partially Eliminates the Effects of Bussing

An examination of Figures 6 and 7 reveals two ellipses: one to enclose the pre-post score pairs of the bussed group, the other to enclose the scores of the remaining group. The line drawn through the long axis of each ellipse represents the relationship between pretest and posttest scores for that group. In each figure the lines are parallel so the hypothesis of homogeneity of regression is not rejected.

For each group, in each figure, a vertical line is drawn to show the location of the pretest mean. It is intersected by a horizontal line showing the location of the posttest mean for that group. To remove the biasing effect of the covariable, the analysis attempts to bring together the two pretest means which it does by adjusting the vertical line in each group such that it comes closer to the vertical line in the other group. Horizontal lines are then drawn from the point on the regression line intersected by the relocated pretest mean. This horizontal line crosses the vertical axis at the location of the adjusted posttest mean. In Figure 7 when the two vertical lines are superimposed, the two resultant posttest lines will also coincide. This means the result of the analysis is that: given minority students of equal pretest scores, bussing to predominantly White schools has no effect on their achievement when compared to peers who stayed behind. In Figure 7 when the two vertical lines are superimposed there will still be a gap between the two posttest lines. If this gap is statistically significantly

different from zero, the inference is that: Given minority students of equal pretest scores, the bussed students will outperform those who remain in the local school on the post-test.

Careful attention must be paid to the common phrase in both inferences: "Given minority students of equal pretest scores. . . ." Note that the figures are drawn to show only the slightest overlap in the pretest scores: the inference from either analysis of covariance applies only to the slightest proportion of the true population, it does not generalize. The researcher only knows what to expect for the portion of those who remain or are bussed whose pretest scores are equivalent. Since all the high pretest scores fall into the bussed group and all low pretest scores fall into the remaining group, the differential impact of bussing for these groups cannot be assessed.

A further qualification of the interpretation is that the covariable used may only be one apparent manifestation of a construct which remains uncontrolled. For example, suppose those children who score higher on the pretest (and are bussed) have more positive attitudes towards Whites and towards schooling. The children who are not bussed do not score well on the pretest because their attitudes are negative. Further, suppose that the results of the study resemble Figure 7. The hopeful inference that bussing will improve the scores of all children lying in the overlapping midrange of the two sets

of pretest scores will prove invalid when it is put into practice. Children who do not volunteer for bussing are often too hostile toward school and Whites to benefit from bussing. Thus, even when the analysis indicates a potential difference between treatment conditions, the researcher cannot be sure that this difference will hold true until he implements the preferred treatment. The reader is referred to Lord (1969) for further insight into these interpretation problems.

Technically, the analysis of covariance imposes some constraints on the data to be collected. (1) The covariate should be measured without error. The achievement test used in the above example, of course, is subject to measurement error. When error of measurement is present the analysis of covariance tends to underadjust. (2) The covariate should be linearly related to the dependent variable. Certain forms of curvilinear relationship, however, can be used through suitable transformation of the data.[6] (3) The hypothesis of parallel regression lines across all the comparison groups must not be rejected. The analysis should not proceed when this occurs, as adjustments will be made in the wrong direction for some groups. (4) The covariate should be significantly related to the dependent variable, otherwise degrees

---

[6]The assumption of linearity is made in most of the statistical models discussed in this section. Methods for testing the appropriateness of this assumption are discussed in the section on multiple regression analysis.

of freedom are lost in the analysis. Clearly, the time wasted collecting, recording and analyzing responses which are unrelated to the outcome variable of interest is a great loss to the researcher. When several covariates are used simultaneously, the data base must contain many more units (recall that schools are units in some analyses) in order for the analysis to function properly.[7]

A recent study in the medical literature (Bunker et al., 1969) presents several other methods for controlling the background variables. Of these, direct and indirect "standardization" seems to be most appropriate to the current discussion. In both methods of standardization the goal is to adjust the values which will be contrasted by taking into account the biasing factors in the data.

In direct standardization the biasing factors are controlled by adjusting the comparison group scores to reflect the result to be expected if the biasing variables were evenly distributed across the comparison groups. Application of this method requires that the students be cross-classified with respect to all the biasing variables to be controlled within each of the comparison groups. For example, in the bussing study above, the bussed and nonbussed groups could be cross-classified by sex and pretest score (the latter would have

---

[7]The reader is referred to Elashoff (1969) for further information concerning the technical aspects of analysis of covariance.

to be divided into reasonable intervals). Then the proportion of students of each sex and pretest score combination in the population can be determined by combining frequencies across the bussed and nonbussed groups. These proportions are then used to weight the average scores of the corresponding groups within each of the comparison groups. The result, for each comparison group, is the score to be expected if the proportions of students in the various cross-classifications of controlled variables were like the proportions in the population.

One difficulty with this procedure is that if any of the comparison groups has no representation for one of the control variable cross-classification groups, the adjustment cannot be performed. Clearly, it would not be appropriate to represent that part of the population by a zero score.

One solution to this problem is to use indirect standardization. With this method the average score for each subgroup in the cross-classification of biasing variables is computed, collapsing across comparison groups.[8] Then these subgroup means are weighted by the corresponding proportion in each comparison group to produce a predicted value which represents the value expected due to the biasing factors alone.

---

[8] In some data collections the cross-classification of variables to be controlled will have many cells with small frequencies of occurrence. In this case some grouping can be performed to increase cell size and stabilize the estimated means. One method of doing this, which is too involved technically to describe here, is called "smear and sweep" by its originators, Gentleman, Gilbert and Tukey (in Bunker et al., 1969).

These values are then subtracted from the observed values for each comparison group to show the effect due to the treatment variable. In this case, when a subgroup is missing from a comparison group nothing is lost in making the adjustment. Astin (1963) uses a procedure similar to indirect standardization.

To summarize, the method of direct standardization seeks to replace the observed values for comparison group means with values adjusted to reflect equalizing of population characteristics thought to be biasing the result. Direct standardization replaces the observed means for comparison groups with values reflecting the influence of the biasing variables. The difference between the observed mean and the substituted value is an estimate of the effect of the treatment, corrected for bias. Of course, the researcher should use caution in selecting variables which he feels are responsible for biasing the outcome, although this caution is less imperative here than in the analysis of covariance.

The basic advantage of the standardization methods over analysis of covariance is that they do not entail the restrictive assumptions imposed by the covariance model. The adjustments are made in a "distribution-free" environment in which assumptions of linearity of relationship, homogeneity of variance and homogeneity of regression need not be considered.

The causal comparative model has much to recommend it in terms of simplicity of design. It is, however, subject to

some difficulties in interpretation (which may be alleviated by newer analytic techniques). This model is also rather limited in scope. It cannot be used to provide the researcher with information about reciprocal interactions among variables or about lagged effects of changes in variables. To do so would require that data be collected over a series of time points and the design would then be changed to a "control series" (Campbell, 1969) which might more properly be investigated using the analytic techniques presented for Process Models below.

The Input-Output Model

The need to control for possibly biased inputs is more explicitly recognized in the input-output model than in the causal-comparative model. In the input-output model the influence of student input characteristics on the output (usually an achievement test score) is first controlled. Characteristics of the educational experiences of the students are then used to account for the remaining variation in outputs.

The input-output model focuses on the characteristics of the schools. The researcher who uses the input-output model may be trying to determine the potential value, in terms of output of certain school characteristics. Thus, Astin and Panos used this model to determine the effects of various college characteristics on students' aspiration to the Ph.D.

They also investigated the effects of college characteristics, particularly that of the peer environment, on students' persistence in college.

In order to control for student characteristics, the researcher typically averages them for the sample from each school and includes these averages in the regression analysis as the first set of predictor variables entered. Unless the number of schools sampled is quite large, the number of such student characteristic variables which may be included is very small. Each such variable uses a degree of freedom which might better be used to test the fit of the model. In the extreme, one could fit enough predictors to entirely account for output in the sample, but a replication of the study would yield much different relationships.

One possible solution to the problem of using up too many degrees of freedom in controlling student characteristics is to use the indirect standardization procedure discussed above to create an estimated mean output for each school based upon the proportions in that school of subgroups of a cross-classification by student characteristics. This estimated value will represent the effects of all the student characteristic variables and their interactions and may be entered into the regression as the first variable--using only one degree of freedom![9] This modification of the input-output model

---

[9]David E. Wiley, personal communication.

should prove highly valuable to researchers concerned with the differential impact of school characteristics.

Since the appropriate sampling unit for this level of analysis is the school rather than the individual within the school, the researcher is advised to sample a few students in each of many schools. Husen's (1967) study best exemplifies this strategy. The number of schools sampled in this study ranged from 8 schools in France to 395 schools in the United States.

Typically, the survey method of data collection is used to obtain the data for an application of the input-output model. Stepwise multiple regression analysis is the usual analytic technique employed in which the researcher forces the student characteristic variables to enter the equation first and then tests the school characteristics for significant additional contribution to output. Examples of this type of analysis are found in Astin and Panos (1969) and Coleman et al. (1966).

A basic difficulty in the interpretation of this analysis is that student background characteristics are often related to school characteristics. The socio-economic status of students, for example, is usually related to the per-pupil expenditure of a school. Thus, when the student background characteristics are controlled, the effects of the related school characteristics are also controlled or at least diminished. This form of analysis is, then, self-defeating

to the extent that school characteristics are not independent of the characteristics of the pupils attending them (Bowles and Levin, 1968A, 1968B; Cain and Watts, 1970). Unfortunately, there is no direct solution to this problem. However, Astin (1968) has attempted to circumvent the problem of underestimating the effects of school characteristics by temporarily excluding the input variables from the analysis and calculating the proportion of explained variance in the criterion variable that can be attributed to college environmental factors alone. This technique provides an inflated estimate of the effects of the school variables studied, which can then be compared with the estimates obtained by controlling for student input, to determine the effects of collinearity upon the analysis and interpretation of the data.

Werts (1968) and Werts and Watley (1968) recommend a similar procedure by proposing that both student input and environmental variables be entered into single regression equations using standardized partial regression coefficients. The explained variance in the criterion variable could then be partitioned into three components: (1) the explained variance due to input variables independent of environmental factors, which would also include the joint effects between input and environmental variables; (2) the explained variance attributable to the school environment, including the joint effects between input and environmental factors; and (3) the explained variance due to the joint effects operating between input and environmental variables.[10]

---

[10]For other methods of partitioning the explained variance, see Creager (1969A,1969B), and Creager and Boruch (1969).

If the researcher is willing to forego the notion of accounting for the effect of various presumed causes of output in terms of a percentage of output variance accounted for, there are other methods available. These methods are discussed with respect to the Process Model in the next section since they require a somewhat more sophisticated knowledge of the relationships among cause and effect relationships. As will be seen in that section, when the researcher's theoretical framework permits him to hypothesize a causal structure interrelating his variables, he may use a variety of techniques to estimate the parameters of this structure which can be used to verify his theory and to make inferences about the potential effects of manipulations of the causal variables.

The Process Model

The process approach is probably the most sophisticated from the standpoint of modeling theory about phenomena. This approach assumes that the researcher has in mind one or more, perhaps competing, structural models based upon the theories developed in the field. The researcher collects data which he analyzes with a view to substantiating one or more but, hopefully, not all of the competing structural models.

A structural model may be thought of in two ways. It may be conceived of as a path diagram with arrows drawn from causal variables to effect variables representing the researcher's hypothesized flow of causality. It may also be thought of as the corresponding set of equations, each of

which summarizes the causal influences bearing upon one variable. (Sometimes, equations represent constraints in the model; for example, variable X must equal variable Y plus variable Z).

Figure 8 illustrates the two aspects of a process model.

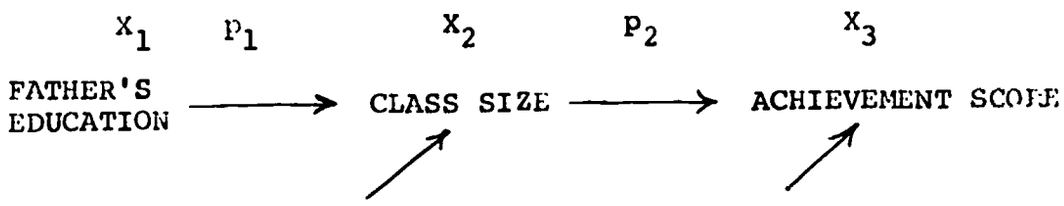


Figure 8. Path Diagram for Model 1.

Examination of Figure 8 indicates that student achievement is a function of class size which is, in turn, a function of father's education. The unlabeled arrows represent unmeasured sources of variation such as measurement error or other, uncontrolled, variables. The equation system, for variables with no error, is:

(1)
$$X_1 = X_1$$
$$X_2 = p_1X_1$$
$$X_3 = p_2X_2 = p_2p_1X_1$$

The p's are the structural parameters of the system (path coefficients) and the researcher's focus is to estimate them. An alternative model is:

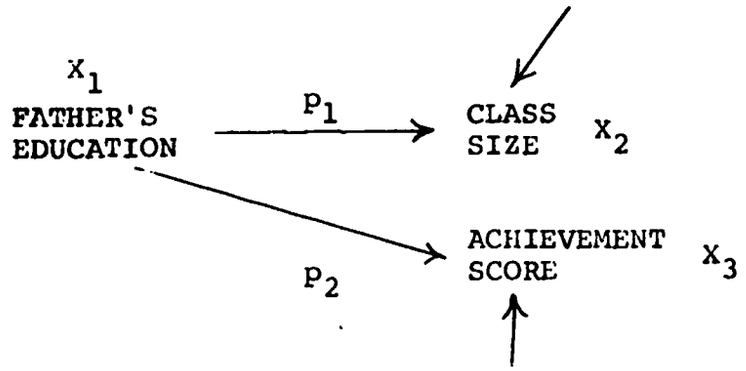


Figure 9. Path Diagram for Model 2.

The corresponding equations are:

(2)
$$X_1 = X_1$$
$$X_2 = p_1 X_1$$
$$X_3 = p_2 X_1$$

Whereas Model 1 hypothesizes a causal chain, Model 2 hypothesizes that class size and achievement score are related merely due to a mutual dependence upon father's education. In the first model the size of the class (i.e., number of students) is the intervening variable through which father's education affects achievement. Thus Model 1 shows a path diagram for what was referred to as interpretation analysis in Chapter III. In the second model, variation in class size is another result of variation in father's education which directly causes student achievement. Thus, Model 2 shows a path diagram for the test of spuriousness described in Chapter III.

It is important to emphasize that the unit of analysis is again the school, so that student characteristics are

averaged for each of a large number of schools. The distinct difference between this model and the input-output model is that differences in student characteristics are no longer considered as something to adjust away or eliminate; they have become an integral part of the causal structure the researcher is modeling.

For the researcher who uses the process model approach, the data collection will yield estimates of the structural parameters (i.e., the p's) and the variances due to unmeasured sources. He uses these to assess the adequacy of the theoretical model and to predict the outcome of certain manipulations. For example, if the value of p_1 is positive indicating that the schools with pupils of more educated fathers have larger classes, the researcher may determine that he needs to include some other controlling variable in the analysis, such as per pupil expenditure by school, in order to see the expected negative value of p_1. (He could, of course, conclude that the theory leading him to expect a negative value was not correct.) A researcher who finds, in the estimation of parameters for Model 1, that the variance in class size attributable to unmeasured variables is zero or very small might then infer that variation in father's education almost completely determines the variation in class size, and thus adopt Model 2. He may then decide to use p_2 from Model 1 to estimate the kind of effect on achievement that would be expected from an experimental manipulation of class size.

The process model would seem to be the most promising technique for analysis of longitudinal data. Complex interactive (in the sense of feedback loops) processes can be adequately represented in the context of this general model. One possible specification of the general model makes possible a cross-lagged panel correlation analysis as described in Campbell and Stanley (1963), Goldberger (1971) and Murray, Wiley and Wolfe (1971). The goal of this analysis is to assess the magnitude and direction of effects when the temporal order of the observations is the only guide to the structural relationship between variables. Murray (1971) and Reynolds (1971) indicate that similar analytic techniques are available for use when the response data are dichotomously scored.

A more detailed exploration of these process models and the techniques they imply for data analysis is beyond the scope of this chapter. However, a few points should be noted. When the path diagram includes no reciprocal influences (feedback loops), and the unmeasured variables throughout the system are uncorrelated, the system of equations is called "recursive" and least squares regression analysis may be used to estimate the parameters of each equation in the system. When there are reciprocal relationships and/or the unmeasured variables are correlated, other methods such as two-stage least squares, an econometric technique (cf. Theil, 1971) become necessary. Blalock (1971) has edited a valuable

collection of papers from sociology, econometrics and other social science fields showing the application of other techniques. In another volume, Blalock (1970) explores the reciprocity between the theory and the data analysis. A recent development is a covariance structure model for causal flow analysis (Keesling, 1972) which eliminates some of the confusion in building and analyzing this process type of model. A special feature of this model is the overall test of goodness of fit which is a useful tool in assessing how well the model represents the data. Finally, it should be pointed out that extensions of these models to cover the case of qualitative rather than quantitative variables are also available (see, e.g., Murray, 1971; Reynolds, 1971).

Three basic approaches to the assessment of educational impact have been explored. Each has special value for the researcher. When the differential impact of a few variations in the educational system is to be investigated, the causal-comparative approach offers simplicity of design. On the other hand, when the potential impact of a large number of school variables is to be explored, the input-output model seems most appropriate. In addition, the input-output model is potentially more flexible than the causal-comparative model because the former permits the researcher to relate more, and more finely measured school characteristics to output than does the latter. The causal-comparative approach, however, has the advantage of making the researcher be very certain

ahead of time of the effects he wishes to investigate, which forces him to closely scrutinize his theoretical foundation. The process model is most applicable when the researcher has a moderately complex causal theory to validate. Finally, it should be remembered that the best assessment of the impact of the school characteristics will be achieved through experimental manipulations of these characteristics and the concomitant random assignment of schools to treatments. The three statistical models presented above may very well serve as exploratory phases in which variables of interest are identified for more controlled experimental investigation.[11]

Multiple Linear Regression Model

1. The Mathematical Model and the Substantive Model

The multiple linear regression model is of the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_p X_p + e$$

where Y is the value of the dependent variable, and X_i is the value of the ith independent variable ($i = 1,\ldots,p$). The β's are the parameters of the model which the researcher wishes to estimate and test for significance. The equation posits a linear and additive model relating the observed values of the independent variables to the observed value of the dependent variable. That is, the model posits a linear relationship between Y and the X's, and the X's are assumed to be related to Y in an additive fashion (i.e., each X affects

[11] The reader is referred to Gilbert and Mosteller (1972) for a discussion of this important point.

Y independently of the values of the other X's). The additivity assumption is indicated by the lack of terms involving products of the X's on the right-hand side of the equation. The statistical procedure associated with this model requires the assumption of an additional term on the right-hand side of the equation, an error term which is taken to have a normal distribution with a mean value equal to 0. For a more complete discussion of this procedure with its concomitant assumptions, the reader is referred to Hays (1963), Kendal and Stuart (1961) and Rao (1965).

The failure to justify the assumptions of this procedure is ubiquitous in the Analytical Review studies; however, various techniques are available which can be used to determine the appropriateness of the multiple linear regression procedure.

Linearity Assumption: The correlation ratio was employed in Bachman's study to test the linearity of the relationship between a dependent and independent variable. This is the only study in which this useful statistic appeared. However, Bachman should have included both the value of the correlation ratio and the product moment correlation coefficient since the test for linearity is based on the difference between the squared values of these two quantities. For example, assume that the analyst is interested in the linear regression of a dependent variable Y on an independent variable X. The square of the true underlying correlation between Y and X is equal

to the square of the correlation ratio of the regression of Y on X if, and only if, X and Y are in a strict linear functional relationship. A much simpler technique for assessing the linearity of the relationship between X and Y involves plotting the values of the two variables against each other and observing whether the graph approximates a straight (i.e., linear) line.

When the researcher suspects that the data will not form a linear relationship, polynomial, multiplicative and exponential models may be used. The reader is referred to Kendal and Stuart for a more detailed discussion of these models.

Additivity Assumption: The inspection and comparison of means is useful in assessing the adequacy of an additive model. For example, in a model containing two categorical predictors with equal numbers of observations at each level, the investigator can compare, for each predictor, the differences between level means taken over all levels of the other predictor. If these differences remain constant, the two independent variables have additive effects. That is, additivity between two or more predictor variables occurs when the relationship between a predictor variable and the dependent variable is the same for different values of the remaining predictor variables. Thus, if differences between levels of one predictor remain constant for all values of a second predictor, then the two predictors form an additive relationship. The table of means presented in Figure 10 is

an example of predictor effects which are additive.

		PREDICTOR A		
	LEVELS	1	2	3
PREDICTOR B	1	12	18	6
	2	20	26	14
	3	17	23	11
	4	23	29	17

Figure 10. Illustrating a Table for Inspection of Interaction Effects

An examination of Figure 10 reveals that the means for the second level of predictor A are 6 more than the means for the first level of A for every level of predictor B. The means for the third level of predictor A are 6 less than those for level 1 and 12 less than those for level 2. Similarly, differences between levels of predictor B remain constant from level to level of A. The case of unequal numbers of observations at each combination of levels is much harder to analyze by inspection.

The detection of interaction in the case of continuous predictor variables is also more difficult. For example, suppose that the model is of the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

The existence of an interaction between the X's indicates that the relationship between X_1 and Y is different for different values of X_2. One of the easiest methods for detecting

interaction is to perform a multiple linear regression and inspect the regression residuals, that is, the differences between the predicted and actual values of the dependent variable. If systematic errors are observed which relate to particular values of the predictors, the interaction between the predictor variables has been identified.

It is possible, of course, that the researcher might know in advance that some of his predictor variables will interact. In the case of categorical predictors, he can then include interactive (i.e., nonadditive) terms in his model and still perform a simple analysis of the data by using a method described in the following section. In the case of interactive continuous predictors, however, there is no simple procedure for including the interactive terms, although there are computer programs that will fit models containing cross-products of the independent variables to the data. Naturally the discovery of interactive effects between variables on the basis of one sample should be verified through further investigation.

To assist in the discovery of interactive terms between categorical independent variables, Bachman used a program developed by Sonquist and Morgan (1964) entitled "Automatic Interaction Detector" (AID). Basically, the program takes a number of categorical predictors, or independent variables, and one dependent variable and follows an iterative procedure of binary splits into a mutually exclusive series of subgroups; at each stage of the procedure the subgroups are selected such

that their means account for more of the total sum of squares than the means of any other equal number of subgroups. Interacting predictors are identified through inspection of the output of the program; the experimenter can then include product terms in his regression model where appropriate.

The AID program seems capable of wide application in survey research studies since the predictor variables are frequently categorical in nature.

Use of Dummy Variables. Dichotomous or "dummy" variables are widely used in linear regression where the predictor variables are categorical rather than continuous. The use of a dummy variable to represent a dichotomy is accomplished by defining a variable which takes the value "1" for all individuals falling into one category and the value "0" for all individuals falling into the other category. For example, if the categorical variable is sex, then males might be assigned a score of "1" and females a score of "0." When variables have three or more categories, the correct procedure is to exclude one of the categories and define variables for the remainder of the categories. For example, suppose the categorical variable is political affiliation. The dummy variables x_1 and x_2 and x_3 might then be defined as follows:[12]

---

[12]Note that only three dummy variables are necessary since an individual's value on a fourth dummy variable is completely determined by his values on the first three.

Political Affiliation	Dummy Variables		
	x_1	x_2	x_3
Democrat	1	0	0
Republican	0	1	0
Independent	0	0	1
Other	0	0	0

Figure 11. Illustrating the Use of Dummy Variables

The dummy variable method allows the analyst to perform analyses of variance using a regression program once the proper recoding has been done. Furthermore, terms representing interactions between the categorical predictors can be included in the regression equation. For example, suppose that the experimenter has found that predictor variables A and B interact. He can define a new predictor variable having values corresponding to each possible combination of values A and B. Thus, if A represents political affiliation and B represents sex, then the variable AB would have eight possible values; female Democrat, male Democrat, female Republican, male Republican, etc. A dummy variable can be constructed corresponding to the predictor AB and included in the regression equation. This procedure is equivalent to the estimation of interaction effects in analysis of variance. However, the benefit of formulating the problem with dummy variables is that it can be handled by a regression program. The dummy variable method is available as a program package entitled "Multiple Classification Analysis" (MCA) and was

used by Bachman. A complete discussion of the use of dummy variables may be found in Lansing and Morgan (1971).

Use of Dichotomous Dependent Variables. When multiple linear regression is performed in the usual manner, and when the dependent variable is dichotomous, the normality assumption is violated and the classical tests of significance do not apply. An example is found in Thistlethwaite's study in which a criterion measure was regressed on a set of student background characteristics. The dichotomous criterion variable was enrollment or nonenrollment in college. Referring to the description of the linear regression model given at the beginning of this section, it can be seen that if y takes on only two values then the error term cannot possibly have a normal distribution.

The problem encountered in the use of a dichotomous dependent variable is discussed by Kmenta (1971) and Lansing and Morgan (1971). Formally, regression of a dichotomous dependent variable on a set of independent variables yields the solution which is obtained when the problem is treated as a discriminant analysis, that is, when the two possible responses on the dependent variable are used to define the two groups and the observations on the independent variables are thought of as observations from each group. This formal identity suggests that problems which have been treated as regression problems involving dichotomous variables can be reformulated as discriminant analyses.

2. Estimating the Parameters of the Model

The method most commonly employed to estimate the β's in the multiple linear regression model is called "ordinary least squares" (OLS). If the data are arranged such that y is a column vector of observations on the dependent variable; x is a matrix consisting of one row (having p column entries--one for each independent variable) for each observation corresponding to the row entry for y; β is a column vector of parameters and e is a column vector of errors, then the model may be expressed in the following form:

$$\underline{y} = \underline{x}\beta + e$$

Ordinary least squares provides the researcher with estimated parameters, $\hat{\beta}$, by solving the following equation:

$$\hat{\beta} = (x'x)^{-1} x'y$$

However, the ordinary least squares method of estimation requires the following assumptions:

a. the values in x are "fixed"; they are measured without error and consist of the range of values to which the results of the study are to be generalized.

b. the entries in e are independently distributed with the means equal to zero and constant variance.

If the model is correctly specified and the assumptions for OLS are met, the parameter estimates are unbiased and consistent, which implies that as the sample size increases

the estimates will approach the true values.

Four problems are readily apparent in the application of this estimation technique in educational research:

1. The model may be incorrectly specified. This has been discussed earlier.

2. The values of x may not be measured without error. A good solution to this problem is not available for general use. However, techniques which can be employed when there are replicate measures of the variables are described by Keesling (1972).

3. The errors may not be distributed with constant variance. Adaptations of OLS are available to handle this contingency (see Graybill, 1961).

4. The errors may not be independently distributed. For example, if the data are measures on the same subjects at several points in time, the errors will be correlated with one another. The econometricians have worked extensively with estimation procedures to circumvent the autocorrelation phenomenon (see Theil, 1971).

Further complications arise when there is more than one equation implicit in the model. Once again, the econometricians have actively investigated this question, which they refer to as "simultaneous equaltion models" (see Theil, 1971).

Testing Hypotheses about Parameters. If, in addition to the foregoing assumptions, it is appropriate to assume that the errors are normally distributed, t tests can be used

to test the statistical significance of $\hat{\beta}$ or to create confidence intervals for the estimated parameters (see Johnston, 1963, for a definitive explanation of this method). Even when the form of the distribution of errors is not known, tests based upon the normal curve may still be used because they have been shown to be robust. As the field of statistics advances, however, techniques are certain to develop to handle alternative error specifications.

Stepwise Regression Analysis. When confronted with a large number of possible independent variables, analysts frequently use stepwise regression analysis to identify the factors which are statistically related to the criterion measure. This technique, however, suffers from the same problems as the input-output model. Specifically, in the case of stepwise regression the amount of variance explained by the variables entered into the equation at any one stage is complicated by the presence of correlated independent variables which have joint as well as independent effects. In the case of the input-output model, the shared portion of the variance in the dependent variable that can be attributed to either input or environmental variables will be attributed to whichever set is controlled initially.

An adequate solution to this problem does not exist. However, the analyst can reverse the ordering of variables into the regression equation to determine the extent to which collinearity is influencing the obtained results.

The Use and Abuse of the Correlation Coefficient

The analytic framework of many of the Analytical Review studies was that of the general linear regression model. Within this context one often finds researchers attempting to "account for" or "explain" the variance in a dependent variable as a function of the explanatory variables. The usual approach to this problem is to compute the coefficient of determination which is the squared multiple correlation coefficient. This statistic presumedly tells the researcher what proportion of the variance in his dependent variable is accounted for by the explanatory variables. Unfortunately, this coefficient does not generalize very well from sample to population or from one population to another population. When a regression is fitted to a sample of data by using a stepwise procedure designed to maximize predictive power, the coefficient of determination is strongly influenced by chance relationships in the data. Thus, cross-validation studies usually show a dramatic drop in the proportion of explained variance. In attempting to generalize the results of a study from one population to another one finds considerable difficulty when the dependent variable is more variable in one group than in another. The form of the relationship may be identical (this may be ascertained by comparing the unstandardized regression weights), while the coefficients of determination will be quite dissimilar.

The researcher is often tempted to try to express the "unique contribution" of each explanatory variable in the model through the use of correlational measures. As Duncan (1970) points out, this is really a hopeless quest when the explanatory variables are themselves correlated. Darlington (1968) has also criticized the attempts to isolate unique contributions of variables. Indeed, Fisher (1946) criticized the practice of generally computing partial or total or multiple correlations by noting: "In no case . . . can we judge whether or not it is profitable to eliminate a certain variate unless we know, or are willing to assume, a qualitative scheme of causation" (p. 191). And, ". . . if . . . we choose a group of social phenomena with no antecedent knowledge of the causation or absence of causation among them, then the calculation of correlation coefficients . . . will not advance us a step towards evaluating the importance of the causes at work" (p. 190).

Even when we are able to specify the causal framework relating the variables, the correlation coefficients (and standardized regression coefficients) are liable to be much less generalizable than the unstandardized regression weights, for the reason given above. (See also, Tukey, 1954.) If the model may be specified in the form of a path diagram, as in the process model above, the causal coefficients are estimated as "structural parameters" which are equivalent to unstandardized regression weights when ordinary least squares regression

is appropriate. These structural parameters are stable as population variances change (unless, of course, a "specification phenomenon" or interaction also occurs in which the form of the relationship changes from population to population) and provide the researcher with all the information he needs to characterize a relationship. Any one coefficient is the unit change in the dependent variable (the one the arrow points to) per unit change in the independent variable (the one the arrow starts from). This provides a direct assessment of the impact of changes in one variable on the outcome of another variable.

Correlational analysis is well established in educational research but it may be time for educational researchers to seriously consider membership in the Society for the Suppression of the Correlation Coefficient described by Tukey (1954). As theory specifies more strongly formulated causal structures, the educational researcher will undoubtedly forsake this correlational analysis for the simultaneous equation models of econometrics (Theil, 1971) and their analogues (Keesling, 1972).

CHAPTER V

COMMON PROBLEMS IN SURVEY RESEARCH

The ultimate contribution of a study depends upon the way in which the analyst treats three major problems common to all survey research--selecting an appropriate sampling design, dealing with nonresponse bias and response error. These problems are discussed in the following chapter particularly as they pertain to the Analytical Review studies. In addition, specific techniques designed to alleviate these problems are recommended.

Sampling Theory

The most precise method of determining the distribution of a particular characteristic or the validity of a specific hypothesis in a predefined group is to collect the necessary information from each group member. For example, the most systematic method of determining the average age of the student body at a particular school would be to ask each student his date of birth and from this information calculate the arithmetic mean. Obviously, this procedure is impractical particularly as the numerical size of the group increases. It would be prohibitively expensive and time consuming, for example, to interview every junior college student in the United States even though the entire population of students attending junior college is the group of principal concern. As a result, survey researchers estimate or infer population values from a subgroup or sample drawn

from the original population. The characteristics of this sample are used as the basis for estimating or inferring the characteristics of the population. The underlying assumption of the sample survey technique, then, is that a fraction of the whole can be used to represent or depict the whole.

The methodological problems that arise from the discrepancy between sampling theory and practice are numerous, usually serious, and often require consultation with sampling experts due to the highly technical and specialized nature of sampling theory. However, some critical concepts in sampling theory and design which can be used to guide the researcher in the selection of sample designs are discussed below.

Sampling Variability. Information gleaned from a sample of respondents is of little value in itself. The data become scientifically important when they can be used to estimate, within the limits of acceptable reliability, corresponding information about some larger group or population. If, for example, 300 students were sampled from a student body of 6,000 in order to estimate the verbal achievement scores of the students at that particular school, the scores of those 300 students would not be of primary concern. Of critical importance would be the degree to which the analyst could feel confident in using the verbal achievement scores obtained from the sample as a criterion for estimating or inferring the distribution of verbal achievement scores for the entire student body.

When a sample of a population is used rather than the entire population, however, a certain amount of deviation or error between the sample value or estimate and the corresponding population value is to be expected. Sampling experts conceptualize this deviation in terms of what is to be expected in the long run. For example, if the data from a particular sample yielded a sample mean, $\bar{X}$, this value would have a certain probability of being observed or selected from among all the sample means that were theoretically possible for that particular sampling design. In other words, if through the same sampling design an infinite number of samples were drawn from a given population and for each sample the mean response to a particular question was calculated, each mean could then be plotted on a graph or histogram. The result would be a distribution consisting of an infinite number of sample means in which each mean value has a certain probability or relative frequency of being selected if a single mean value from this distribution were to be randomly drawn. Distributions of this type are called sampling distributions, the most important being the sampling distribution of sample means.

One of the more significant characteristics of a well designed survey is that the mean of the sampling distribution of sample means (i.e., the arithmetic mean of the sample means) equals or closely approximates the mean of the population. When this is true, the magnitude of the variance of the sampling distribution of sample means for a given sample design

becomes critically important because it indicates the degree of confidence that can be placed in the sample estimates. That is, when the sampling design is used to select each sample that becomes part of the sampling distribution of sample means, the amount of variance in the sampling distribution represents the fluctuation or deviation between the sample estimate and the population value that is due to the specific sampling design. Consequently, the smaller the variance, the greater the probability that any given sample mean approximates the true population mean.

The discussion of sampling variability up to this point has been largely theoretical. The sampling distribution of sample means was initially defined in terms of an infinite number of sample means plotted in the form of a histogram. Clearly, however, it is not possible to select an infinite number of samples. Consequently, it is impossible to obtain the standard error directly from the data. Instead, the analyst must estimate the standard error by using statistical formulas designed specifically for this purpose. Since different formulas exist for different sample designs it is generally wise to consult a sampling expert or one of the available texts on sampling theory in order to determine which formula is appropriate. The formula for estimating the standard error for the most basic sampling design, the random sample is presented below:

$$\text{estimated standard error of the sampling distribution of sample means for random samples} = \frac{S}{\sqrt{N}}$$

where S = the sample variance and N = the size of the sample.

It is apparent from this formula that one method of reducing sampling variability is to increase the size of the sample N. In fact, this is true for all sampling designs; the larger the N, the greater the confidence that can be placed in the sample estimates. The standard error can also be reduced by using a sampling design that yields a smaller sample variance. The differences that exist between sample designs reflect this dual consideration. Some designs allow the analyst to increase the N of the sample at a fixed cost while other designs will minimize the sample variance at a fixed cost. Frequently, a particular sampling design will yield a small sample variance but only allow the investigator to collect a small number of cases at a fixed cost, while other sampling designs will provide a large number of cases at a fixed cost which have a large sample variance. The task facing the analyst, then, is to select the one sampling design that optimally combines the two methods of reducing the standard error for his particular survey. There is no one design which is best for all surveys. The final decision of which sampling design to use must take into account the available resources (including time, convenience and money), the phenomena to be analyzed, and the definition and geographic location of the population.

The optimum sampling design will yield the smallest variance per unit cost. This can be determined in one of two ways: by calculating the cost that would be required for different sampling designs to reach a fixed level of

variance, or by fixing the cost and determining which sampling design will yield the smaller variance. Unfortunately, this is easier said than done. It is virtually impossible for the analyst to obtain accurate estimates of the sample variances and cost factors for all of the different sampling designs. Frequently, past experience with similar surveys and target populations can provide some guidance. Or, pilot surveys can be conducted in order to provide this information. Perhaps the best solution in the long run would be to start a data bank in which sampling information from the many surveys in education would be catalogued. By carefully analyzing these data and periodically summarizing the major conclusions in the form of progress reports the latest information concerning the efficacy of particular sampling designs appropriate to specific research goals and target populations can be made available to the analyst upon request.

Basic Sample Designs

The most essential criterion of a good survey sample is that it be a probability sample. A probability sample is one in which every individual or sampling element has a known probability of being included in the sample. When these probabilities are not known, that is, when the sample is a nonprobability sample, the analyst cannot legitimately use statistical inference. Nevertheless, a number of studies under review employed inferential statistics even though their samples were actually nonrandom, nonprobability samples. For

example, Lehmann and Dressel surveyed the entire freshman class at Michigan State University; Katz and associates surveyed the entire freshman class at Stanford University and two thirds of the class at the University of California, Berkeley; Kagan and Moss investigated those subjects who voluntarily participated in the Fels Institute program.

All of these samples were self-contained, representing no known target population other than the subjects surveyed. Consequently the interpretation of the inferential statistical significance of observed differences and relationships reported by the researchers are problematic since these statistics are not based upon "true" probability samples. In the strictest sense these statistics may be considered illegitimate under the circumstances. However, generally the best of probability samples are only approximations of their target populations; moreover, the researchers just cited were no doubt using some of the best tools at their disposal to obtain an index of differences and relationships among the individuals and groups they were investigating. The point is to recognize the limitations of these tools, particularly when generalizations based upon their application are made to a population beyond that actually investigated.

To calculate the sampling probabilities it is necessary that a listing of all possible population elements be used for the actual selection process. Creating such a list is not easy, although in educational research the problem is usually not as difficult as it is in other areas of empirical investigation.

Even in education, however, there may be no listings of the appropriate population elements or, if there are lists, they may be incomplete or contain duplications or individuals who are no longer considered members of the population. Nevertheless, it is essential to the validity of any probability statement that a reasonably complete list be used to select the sample. When a complete list is not available it may be advisable to redefine the population to conform to the existing list.

There are four basic types of probability samples in survey research: (1) random samples; (2) systematic samples; (3) stratified samples; and (4) cluster samples. Each of these types is briefly described below.

Random Samples. In a random sample each individual in the population has an equal chance of being selected and all combinations of individuals selected are equally probable. Random sampling without replacement (i.e., once selected, the individual's name is removed from the pool of names so that he cannot be selected again) is sometimes referred to as "simple random sampling." Although simple random sampling violates the assumption of independence that usually accompanies statistical tests (i.e., selecting one sampling element does not affect the probability of selecting another sampling element), the problem is not serious when the sample represents a small fraction of the population. However, when the sampling fraction is as high as one-fifth, most sampling experts suggest that correction factors be introduced, if they

exist for the particular statistic used. It is important to remember that most of the statistical tests described in introductory texts assume random or simple random sampling. Consequently, these tests should not be used in conjunction with other sampling designs without first considering the extent and possible consequences of violating this assumption.[1]

Systematic Samples. In a systematic sample every K^{th} individual is sampled instead of selecting each individual independently. For example, Jones et al. selected for their Berkeley Growth sample every third child born in Berkeley, California within the 18-month period between January, 1928 and June, 1929. In a systematic sample, the first choice must be selected through a process of random selection and all sampling elements should be randomly allocated before selection takes place. Of special significance are two deviations from random allocation which can result in serious sampling biases when a systematic sampling design is used. These deviations are called trends and periodicity and will be briefly described below.

1. Trends. Suppose there are 600 students at a particular school and the analyst decides to take every sixth

---

[1]Several books on sampling theory are: W. G. Cochran (1953); W. E. Deming (1950); M. H. Hansen, W. N. Hurwitz, and W. G. Madow (1953); and L. Kish (1965).

student in order to obtain a sample of approximately 100 students. Suppose further that the list of 600 students was not arranged on the basis of random allocation but on the basis of scores to a particular achievement test so that the first student on the list scored the highest mark, the second student the second highest mark, and so forth. If the intervals are six sampling elements wide, one of the six possible samples would contain individuals 1, 7, 13,... 595 while a different sample would contain individuals 6, 12, 18,...600. It should be apparent that each member in the first sample would have a higher test score than the corresponding person in the second sample. In other words, the mean test scores of the two samples would differ significantly from one another even though they were drawn from the same population. This type of variation between samples, or sample variability, must be reduced if the sampling design is to be efficient because it makes the sample estimates overly dependent on the particular sample that happened to be selected rather than on the true population value.

2. Periodicity. Sample variability can also be observed when the listing contains cyclical fluctuations. For example, imagine that 1,000 schools were initially surveyed and ten students were sampled from each school. If a list was then compiled in which the students from the same school were listed one after another in descending order of their scores to an achievement test, this enumeration would contain 1,000 cyclical fluctuations of test scores with a complete cycle occurring every eleventh student. If the analyst

decides to subsample from this list and selects a sampling fraction of one-eleventh, it is apparent that a random start of one would produce a subsample consisting entirely of students who had the highest test scores among their peers. Likewise, if the random start was ten, then the subsample would consist entirely of students who had the lowest test score among their school peers.

One of the easiest and least expensive methods of estimating the extent of sampling variability due to trends or periodicity is to select two or more independent systematic samples that have different random starts. Thus, instead of drawing one large systematic sample the analyst might select two smaller samples. If the sampling variability is small, then the mean of the two samples should be approximately the same.

When the sampling elements on the list are randomly allocated and when the first selection in a systematic sample is a random selection, it is usually safe to conclude that the systematic sample is equivalent to a simple random sample. In addition, systematic random samples are frequently easier and less expensive to select than are simple random samples. Consequently, the systematic sampling procedure, when properly designed, can often reduce the costs associated with drawing a sample or allow the analyst to increase the size of the sample for a fixed cost.

Stratified Samples. To stratify a sample the population is first classified into a number of subpopulations or groups

called strata according to some prespecified criteria such as geographical location, race, age, etc.; then an independent sample is drawn from each stratum. For example, Astin and Panos, Coleman, Flanagan and associates, Husen, Thistlethwaite, and Tillery and associates employed a stratified random sampling design. Coleman stratified his sample on the basis of metropolitanism, geographical location or region, and race. Flanagan and associates stratified their project TALENT sample on the basis of geographical region, type of school (e.g., public, private, parochial), size of senior class and school retention ratio. Husen stratified his sample on the basis of four types of populations: all pupils who were 13.0-13.11 years of age at the date of testing; all pupils at the grade level where the majority of pupils of age 13.0-13.11 were found; all pupils studying mathematics as an integral part of their course; and all pupils studying mathematics as a complementary part of their studies.

Stratified samples can be either proportional or disproportional. In proportional stratified samples the sampling fractions for each stratum are identical so that the sample proportions will correspond to the population proportions. In disproportional stratified samples, on the other hand, the sampling fractions are unequal. Thus, Coleman insured that a proportional stratified sample of minority and majority students would be obtained by allocating a predetermined proportion of the sample to metropolitan and nonmetropolitan areas. Estimates of non-White enrollments in grades one,

three, six, nine and twelve were obtained from metropolitan and nonmetropolitan areas for each of the major geographical regions. It was found that approximately 62 percent of the total non-White enrollment were in metropolitan areas and about 38 percent in nonmetropolitan areas. As the number of non-White students that were to be included in the sample was set at approximately 450,000, the number of non-White students that was allocated to metropolitan areas was set at 279,000 (450,000 x .62) and the number of students in nonmetropolitan areas was set at 171,000 (450,000 x .38).

Tillery and associates used a multistage probability sampling design in which a disproportional sample of schools from California, Illinois, Massachusetts and North Carolina were initially selected. Specifically, the number of schools which participated in this project were:

	Public Schools	Private Schools
California	32	12
Illinois	46	18
Massachusetts	28	21
North Carolina	138	4

However, within each state, counties were grouped on the basis of their similarity in terms of median family income, proportion of white collar workers in the county, racial composition of the county, school size, etc. Counties were then randomly selected from each of these groups presumably with equal sampling fractions.

Similarly, Flanagan and associates used differential sampling ratios for the different school size strata, under-sampling the smallest public schools and oversampling the largest ones.

Proportional stratified samples attempt to reduce sampling variability by guaranteeing a more representative sample than might be expected from chance or random selection. The way in which this operates can be seen more clearly if the total sample variability is partitioned or broken down into two major components; the variation or discrepancy between the relative size of sample strata and the corresponding population strata and the variation or representativeness of the elements within each sample stratum. In other words, there are two basic sources of error in selecting a sample: (1) selecting too many (over-sampling) or not enough (under-sampling) of certain subgroups of the population; and (2) selecting sampling elements that are not representative of the population. The researcher, for example, can over-sample within the senior class and/or select a group of students that are unrepresentative of their peers. Both of these sampling errors contribute to the total sample variability.

When the analyst uses proportional stratification he exerts control over the size of the strata and thus can reduce the total sampling variability by reducing the variation that is due to the discrepancy between the relative size of the sample strata and the corresponding population strata.

Thus, Coleman insured that the sample of minority students drawn from the different geographical regions would be proportional or representative of the true racial composition of these regions.

Stratified samples, however, do not reduce the variation that is due to the representativeness of each sampling element. Consequently, when the variation between the size of the sample and population strata is expected to be large, proportional stratification can be of considerable value in reducing sampling variability. But when the representativeness of each sampling element is the primary source of sampling error, then little is to be gained by stratifying the sample. In other words, if the differences between strata are large compared to the differences within strata, a stratification design will be of value to the investigator. Consequently, the larger the correlation between the stratifying characteristics and the variables to be studied, the greater the efficacy of the stratification procedure. Therefore, when the strata are homogeneous and related to the major criterion variables, the gain from stratification can be considerable.

In addition, when the sample is large, the gain from proportional stratification is usually insignificant because chance factors alone will provide a close approximation of the relative size of the population strata. For large samples the same degree of accuracy or precision in the sample estimates can often be obtained by using a random sampling design.

Thus, it is doubtful that the stratification procedure employed by Flanagan and associates significantly decreased the sampling variability, since the sample contained over 400,000 students. When the sample size is small, however, proportional stratification can significantly reduce sampling variability.

As mentioned previously, in disproportional stratification the strata are sampled unequally. This is done to further increase the efficacy of the sampling design. In general, a disproportional sampling design is best suited for situations in which (1) the focus of the survey centers on investigating specific subgroups rather than the total population, (2) large differences exist in the homogeneity of the strata, or (3) the cost of gathering data differs significantly among the strata. When the survey is conducted to analyze subpopulations, it is recommended that measures be taken to insure the inclusion of these strata in the sample since the possibility exists that a purely random sampling technique would select an insufficient number of cases to support the type of analyses desired by the investigator.

When the analyst has reason to believe that the homogeneity or the variances of the different strata on the criterion or dependent variable differ from one another, this information can frequently be used to improve the design of the sample. Intuitively it should be clear that as the variance of a stratum decreases, the number of cases needed to adequately represent the subgroup also decreases. In

other words, when the variance of the stratum is small, less sampling is required to achieve a given level of accuracy or precision in the sample estimates.

Finally, the cost of collecting the data will often vary from stratum to stratum. Obviously, it would be less expensive to select a large proportion of the sample from the strata that are the least expensive sources of data. It can be demonstrated mathematically, for example, that the optimum allocation of resources will be attained if the sampling fractions are inversely proportional to the square root of the cost factors. This procedure should be followed with caution, however. Strata that differ in data collection costs are also likely to differ in other ways that could impart a serious bias into the sampling design.

Ordinarily it is not recommended that a disproportional stratification sampling design be used unless there are clear advantages for doing so. When the assumption of homoscedasticity or equal variances across strata can be defended, and when the largest population or subgroups of principal interest are sufficiently large in number to make it highly unlikely that an insufficient number of cases will be randomly selected, then there will be little gain in disproportionately stratifying the sample and often considerable disadvantage in terms of increasing the complexity of the sampling process. In short, considerable differences are required in order to justify disproportionate stratification. Furthermore, unequal sampling fractions should be used only after consultation with

a sampling specialist who is familiar with the substantive area under investigation. The optimum allocation of sampling fractions for one phase of the analyses can frequently result in large losses of precision or accuracy in other areas of the investigation.

Cluster Samples. In a cluster sample the population elements are also divided into groups but instead of sampling within groups, as in stratified sampling, entire groups are sampled. For example, the investigator may divide a particular geographical region into school districts, then randomly select one of the school districts and survey all of the students in this cluster.

The principal advantage of sampling among groups rather than among individuals is the reduced expense; cluster samples are usually much less expensive to select and interview than random or stratified samples. This is due to the fact that the clusters are usually grouped on the basis of physical location rather than on the basis of certain attributes or characteristics of the population elements that are related to the dependent variable. Cluster sampling takes advantage of the decrease in cost per interview that comes from collecting data on subjects who are centrally located.

Cluster samples are most effective when the sampling elements in the groups are as diversified or as varied as possible. Unlike stratified samples where it is desirable to have the strata homogeneous, the goal in cluster sampling is to select heterogeneous groups because they will be used

to represent the entire population. However, the very process of selection in cluster sampling makes the attainment of a heterogeneous group problematic. Clustering would not be a liability if all of the elements in the population were randomly distributed throughout all of the clusters. But practical experience as well as intuition suggest that the elements in a particular cluster tend to resemble other elements in the same cluster more than they do elements in different clusters. In other words, clusters tend to be homogeneous, resulting in a considerable amount of sampling variability. In fact, as a general rule cluster samples have more sampling variability than simple random samples of the same size.

Finally, it should be pointed out that serious problems arise from the application of the more common statistical formulas to data derived from clustered samples.[2] The statistical formulas found in the popular texts on statistical analyses cannot be applied to data derived from cluster sampling. Kish (1957) demonstrated, for example, that when the true alpha levels are as high as .50 for clustered sample data,

---

[2]This is so because the number of independent selections is markedly reduced in cluster sampling. A random sample of 400 students implies that 400 independent selections were made, whereas, selecting ten clusters of forty students each indicates that only ten selections were made. This important difference has serious ramifications for many of the popular statistical tests. The standard error for example, is calculated by dividing the sample variance by the square root of the number of independent sampling selections which, in cluster sampling, equals the number of clusters rather than the number of sampling elements.

the alpha level calculated from formulas designed for simple random samples can be as low as .05. In other words, the errors that will be committed when simple random sample formulas are used for clustered data will rarely be conservative. For stratified samples, however, the problem is more tolerable. Stratified samples are frequently more efficient than simple random samples in terms of reducing variability. Thus, the analyst will usually be on the conservative side in estimating the alpha level for data derived from a stratified sampling design.

It should be apparent from the previous examples of Analytical Review studies that the sampling designs discussed in this section are usually used in combination. In fact, almost all of the sampling designs used in the studies reviewed represented such combinations. In addition to the examples previously cited, several other studies should be mentioned. Bachman and associates, for example, employed a three-stage probability sample. In the first stage, the continental United States was divided into eighty-eight clusters. Sixty-two of these clusters corresponded to separate counties; the rest were grouped into twelve major metropolitan areas. In the second stage a single school was randomly selected from each cluster. Finally, a random sample of approximately thirty boys was obtained from each selected school.

Thistlethwaite employed a two-stage probability design in his study by first stratifying his sample and then

selecting respondents within each stratum on the basis of a simple random sampling technique. Similarly, Flanagan and associates used a two-stage probability sample in which the first stage consisted of a stratified random sampling design and the second stage consisted of simple random sampling within strata. Astin and Panos also stratified their sample. However, in this case respondents within each stratum were selected by a systematic random sampling procedure.

It should be noted that the variety of sampling designs available far exceeds those discussed in this section. Consequently, only through careful and knowledgeable evaluation of the available and feasible sampling designs will the optimal design for a particular study be identified.

Specification of Target Population and Sample Representativeness. In designing the sample and determining the manner in which the sample is to be drawn, it is essential that the objectives of the study be considered. An appropriate sampling design may be employed by the analyst while the sample elements selected for analysis are inappropriate in terms of the goals of the investigation. For example, an analyst might appropriately select a cluster sampling design when the groups he wishes to analyze are heterogeneous and yet select inappropriate groups to form these clusters. This situation occurred quite frequently in the Analytical Review studies, as will be illustrated below.

According to Thistlethwaite, the primary objective of his study was to:

> . . . identify types of college environments which facilitate or impede the undergraduates' motivation for advanced training, and to formulate steps which college administrators and faculties might take to encourage more of their talented students to seek graduate or professional training (p. 19).

To investigate this problem, Thistlethwaite stratified a random sample of 30,000 National Merit Scholarship Qualifying Test examinees (NMSQT). It is questionable whether this sample actually dealt with the central objective of the analysis stated above. First, NMSQT examinees are not representative of the entire population of college students. The 480,000 examinees from which the sample was drawn represented only 28 percent of all the high school graduates in the United States. In addition, only half of all United States high schools offered this test during the year the sample was drawn. Furthermore, as Thistlethwaite acknowledges, students from high schools which administered the NMSQT were more likely to plan to go to college, to have parents who encouraged them to do so, and to have enrolled in college preparatory courses, than students in schools which did not give the test. Thus, a selection bias was operating in this sample; many, if not most of the students had high educational aspirations prior to their college enrollment. Consequently, Thistlethwaite was attempting to ascertain the effects of the college environment on student aspirations when student background characteristics had previously played a major role. This, of course, reduced the probability of detecting the impact of the college and provided a relatively limited examination of

the major objective. Moreover, since an important segment of the student population was excluded from analysis a very important theoretical question was left unanswered: to what extent does the college environment influence students who do not have strong educational aspirations prior to college enrollment?

Similarly, Bachman and associates were interested in identifying factors related to high school attrition. Yet, an important segment of the dropout population was underrepresented in their sample. The sampling design used by the investigators was not well suited to the description and comparison of subgroups of minority students because they were located in a small number of schools. The researchers acknowledge this fact by stating:

> Only 256 of our 2,213 respondents are black; more serious from a sampling standpoint is the fact that over two-thirds of them are concentrated in only nine of our sampled schools (with the remaining third scattered in 25 other schools). In short, our ability to generalize accurately from the black subsample is severely limited, and this argued against a strong concentration on racial differences (pp. 25-26).

Again, it is often difficult to obtain the ideal sample. What is important, then is to recognize the limitations of any given sample, to examine the validity of the objectives and generalizations of each study in reference to its sample and to pay close attention to commonality and divergence of findings across studies and samples.

Trent and Medsker's cross-country sample of high school students was also unrepresentative of high school students in

general. Although the researchers assert that they purposely excluded the Northeast and Southern United States, they also excluded the Northwest and the Southwest was only represented by California. In addition, as the authors acknowledge, the selection of high school seniors was not representative due to the lack of representation of Jewish students who tend to cluster in large metropolitan areas and, who proportionately to their total population, have the highest proportion of college attenders. The systematic random sample of infants used in the Jones et al. study was also restricted in its generalizability since the sample was drawn entirely from a list of registered births in Berkeley, California. Berkeley is an atypical community compared to the general population in terms of infant mortality rate, level of parents' education, level of fathers' occupation, per capital income, percent of foreign born, and percent of home ownership. Thus, it is difficult to determine to what population the results from the Berkeley Growth study may be generalized.

Although all of these studies incorporated a random sampling design, important segments of the target population were excluded from analysis. At issue is not whether these analysts drew a sample that was representative nationally, rather, it is to what target populations these findings can be generalized. Random sampling designs should provide a representative sample of the sampling lists from which the sample was drawn, however, they will not provide the analyst with a sample appropriate to the goals of the study if the sampling lists do not conform to

the objectives under investigation. The analyst must carefully specify the target population in relation to the specific objectives of his study and select a sample that is representative of this population so that generalizations can be made to the population of interest.

Nonresponse

The most carefully designed survey will yield accurate data only to the extent that the sample is actually representative of the target population. As a result, analysts are justifiably concerned with nonresponse rate because a significant number of missing respondents can impart serious biases into the data. As Blalock (1960) notes:

> Whenever (sampling) lists are incomplete or whenever a large percentage of persons must be considered as nonrespondents, we have in effect another example of nonprobability sampling . . . even though pains may have been taken initially to obtain a probability sample, certain individuals actually have no probability of being included in the ultimate sample because they have selected themselves out by refusing to answer (p. 411).

A number of factors can contribute to the rate of nonresponse, such as unavailability or refusal of the respondent to cooperate, incorrect mailing lists, and misplacement of questionnaires. Unintentional oversights on the part of the respondent or interviewer can also lead to significant nonresponse rates.

The rate of nonresponse becomes a serious methodological problem when (1) the nonresponse occurs systematically and in large enough numbers so that the responses of certain segments of the target population are not adequately represented, and

(2) when this distortion leads to alternative or rival interpretations of observed relationships. For example, Bowles and Levin (1968A) point out in their critique of the Coleman study:

> Complete sets of survey instruments were returned for only 59 percent (689 out of 1,170) of the high schools. Moreover, there is reason to believe that the pattern of sample nonresponses is not random. One characteristic contributing to this bias is the fact that a disproportionately large number of big cities refused to participate in the sample. Thus, in an analysis of metropolitan data one finds an over-representation of suburban relative to city schools (p. 6).

This bias, as Bowles and Levin acknowledge, casts doubt on the representativeness of the sample and complicates the interpretation of the data.

Astin and Panos uncovered a nonresponse bias that could have posed a problem in their analysis of Ph.D. aspirations among college students. The researchers used response-nonresponse as the criterion variable in a multiple regression analysis and determined that students with high school grade point averages ranging from B to D, whose fathers did not graduate from high school, who aspired to less than a bachelor's degree, and had not published original writing were less likely to complete and return the follow-up questionnaire than students who had A averages, aspired to at least a bachelor's degree, had published original writing and whose fathers graduated from high school. Clearly, if the response bias had not been corrected, any conclusions reached by the analysts concerning the effects of the school

on aspirations for graduate school could have been challenged on the grounds that an important segment of the student population was not included in the analysis.

In addition, systematic response biases in favor of higher academic aptitude and higher socio-economic levels were found in both the Project TALENT and Trent and Medsker studies. In the latter case, a significant chi square was obtained for the college students when the respondents and nonrespondents were compared by level of socio-economic status, but not for the youths who did not enter college. College and noncollege respondents were also significantly higher in level of ability than nonrespondents. However, when the respondents and nonrespondents were compared on the three attitudinal scales related to intellectual and academic motivation, there was no systematic difference between the two groups.

According to Project TALENT's data, respondents to the mailed questionnaire had greater academic aptitude than nonrespondents and came from higher socio-economic level families. Specifically, students whose fathers held professional or technical jobs were more likely to be respondents. Students whose fathers were workmen or laborers and students who did not know their fathers' occupations were more likely to be nonrespondents. As the educational level of the parents increased, the incidence of questionnaire response increased.

There was also somewhat greater mobility found among the nonrespondents in the Project TALENT sample, and a slight

tendency for regularity in school attendance to be associated with questionnaire response. In addition, although length of residence in a community was associated with response to the mailed questionnaire, the type of community as such was not associated.

In spite of these problems, however, the majority of the Analytical Review studies failed to systematically analyze the effects of nonresponse on the resultant data. Hilton, for example, reports that 15 percent of the Growth Study sample was lost between the ninth and eleventh grades. However, no attempt was made to determine if these nonrespondents were in any way different from the respondents. Furthermore, according to Hilton, when a student left a Growth Study school, no effort was made to follow him to his new location. Similarly, Kagan and Moss allude to the problem of nonresponse but do not examine the differences between respondents and nonrespondents. Newcomb distinguished between respondents and nonrespondents, but only in a cursory way. Finally, Katz and associates and Lehmann and Dressel did not report any nonrespondent follow-up procedures whatsoever. In light of these discrepancies between problem and practice, methods for reducing and correcting for nonresponse are discussed in the following section of this chapter.

There are two basic types of nonresponse in survey research:

1. the rate or percentage of nonresponse to the entire survey instrument or "questionnaire nonresponse"; and

2. the nonresponse rate for particular questions or "item nonresponse."

In addition, two related forms of nonresponse occur in panel designs:

1. sample mortality, or the extent to which the primary units of analysis (usually the individual or student) are lost or unavailable for subsequent investigation; and
2. item mortality, or the percentage of nonresponse for a given survey item that occurs after the initial measurement.

Questionnaire Nonresponse. The most common method of reducing questionnaire nonresponse is to increase the number of attempts to contact the respondent. In personal interviews, for example, the interviewer may continue calling on the respondent until X number of attempts have been made.[3] Most of the survey instruments in educational research, however, are self-administered. Moreover, many survey samples such as those investigated by Coleman, Flanagan and associates, Tillery and associates, and Trent and Medsker, are distributed over large geographical areas and the expense involved in maintaining a research staff in the field often makes follow-up contacts prohibitive.

---

[3] For a discussion of the technical considerations in the call-back strategy, see Stephan and McCarthy (1958), Zarkovich (1963), and Kish (1965).

Researchers, therefore, must be most careful in planning their surveys so that the initial nonresponse rate is minimal. This admonition is all the more relevant since common instances such as scheduling interviews or administering questionnaires on a Monday or Friday, during an epidemic, or during periods of bad weather are likely to yield a relatively high questionnaire nonresponse rate because of student absenteeism. Moreover, college students may be less cooperative during examination week, when special extracurricular activities are taking place, or when the climate makes outdoor activities especially attractive.

If the student sample has been drawn from the student body roster, a postcard mailed to the student or his parents explaining the nature of the survey may serve to increase the response rate. Contacting students and faculty prior to the survey date can also help to increase response rates. Several techniques are available to the researcher who uses mailed questionnaires as the principal method of collecting data. For example, there is evidence to indicate that hand-stamped envelopes elicit a higher return rate than business reply envelopes (Gullahorn and Gullahorn, 1963; Price, 1950; and Robinson and Agism, 1951). The amount of postage on the return envelope also appears to be related to the return rate in a positive direction (Gullahorn and Gullahorn, 1959, 1963; Kephart and Bresslar, 1958).

Astin and Panos' research was unique in this respect in that the effectiveness of different follow-up procedures was systematically investigated. A random sample of 665 subjects was drawn from a final pool of 23,673 "hard-core" nonrespondents and assigned to one of twelve treatment cells in a 2 x 2 x 3 design. The effects of type of cover letter, type of outgoing postage, and class of outgoing mail were investigated. Astin and Panos concluded that using first class mail, metered stamps and mimeographed cover or introductory letters reduced the chances of obtaining a response compared to certified or special delivery mail, live stamps and personalized cover letters.

A postcard mailed to the student prior to the questionnaire explaining the objectives of the study and the importance of a high response rate may also increase the rate of response. In the original Bennington College study, for example, Newcomb sent to each student a statement of intent to participate in the follow-up and a letter stressing the importance of the student's cooperation. About 80 percent of the students signed these statements and of that group almost 90 percent returned each subsequent questionnaire.

Further efforts can be made through "reminder postcards," telegrams and telephone calls. Persistent efforts of these kinds to obtain responses from all members of the initial sample were made with positive effects by Bachman, Flanagan and associates, Super, and Trent and Medsker. These studies

obtained greater response rates through personal contact than studies that did not use these techniques. Lowest response rates are characteristically obtained by those studies that make no effort to contact nonrespondents or make only a single effort to do so. In Super's second follow-up, for example, questionnaires, cover letters and $2.00 were sent by certified mail, return receipt requested to the 140 survivors of the original ninth grade group. Fifty-three percent of the 140 subjects responded. A postcard was sent to 57 more subjects and 25 completed questionnaires were returned, making a total of 71 percent. A letter and another copy of the questionnaire were sent to an additional 31 subjects. Seven questionnaires were returned for a total of 76 percent. The names of 32 of the remaining 34 subjects were turned over to a psychologist associated with the Community College in Middletown, who secured the addresses of these subjects and also made personal contacts with friends and relatives of the subjects, former landlords and neighbors. This field follow-up produced 17 more completed questionnaires, for a final total of 88 percent of the surviving original ninth graders.

Finally, mailing a short-form questionnaire containing measures of the most important variables can be used. Astin and Panos employed this technique with success in their study of student aspirations and career plans. In addition, Trent and Medsker report that brief postcard questionnaires elicited much higher response rates than the comprehensive questionnaires which took over an hour to complete.

The potential for nonresponse bias exists in every survey. It is almost inevitable that some respondents will refuse to cooperate or become lost to the analyst. For example, in Lehmann and Dressel's study the nonresponse rate was 32 percent for the four-year group and 40 percent for the control group. Similarly, Flanagan and associates report that substantially greater than 25 percent of the students surveyed in grade nine were lost by the twelfth grade in Project TALENT. In Newcomb's study the nonresponse rate was approximately 37 percent, while the Coleman study contained a 30 percent nonresponse rate among the schools initially sampled. Since a certain percentage of the respondents will undoubtedly be lost regardless of the techniques employed to increase the response rate, analysts should also be concerned with estimating the effects of nonresponse and correcting or adjusting for the biases that arise from a significant nonresponse rate.

One technique that can be used to estimate the effects of questionnaire nonresponse requires the availability of call-back data. If a definite response pattern emerges after plotting the scores obtained during different call-back attempts, a curve can frequently be extrapolated to include the nonrespondents.[4] For example, if 60 percent of the students were interviewed initially, 10 percent in the first

---

[4]For excellent literature reviews of this technique, see Houseman (1953) and Zarkovich (1963).

follow-up and 5 percent in the second and final follow-up, and each succeeding interview attempt revealed a greater percentage of dissatisfaction with the policies of the school administration, as indicated in Table 10, then the analyst can make a reasonable inference that the nonrespondents will also tend to be dissatisfied with the school administration. Unfortunately, however, this technique was not commonly found in the research examined.

Table 10

Hypothetical Data Illustrating a Definite Response Pattern for Different Call-Back Attempts

	Percent of Total Sample	Percent Dissatisfied with School Administration
First Call	60	30
Second Call	10	50
Third Call	5	70

The major disadvantage of extrapolating curves to fit the nonrespondent population is the expense associated with the call-back strategy. It may therefore be advisable to draw a random subsample from the initial population of nonrespondents for future call-back attempts.

Drawing subsamples is often an unwieldy task that complicates the administration and bookkeeping of a survey. To

circumvent the problem of call-backs and subsampling, Politz and Simmons (1949) have suggested an alternative procedure that avoids call-backs altogether. The analyst first collects information from the respondents concerning the probability of their being interviewed during other similar periods. Then, a probability coefficient of participation is calculated for each respondent and their responses are weighted accordingly. For example, each respondent is asked on how many K similar occasions he would be available for the interview; if the answer is S, the interview or questionnaire is weighted by $(K+1)/(S+1)$. Thus, if K=5 and S=3 then the questionnaire is weighted by a factor of 6/4 because it is assumed that only 4/6 of the respondents with similar probabilities will be successfully contacted.

The evidence pertaining to the Politz scheme, however, is not encouraging. Thus, Durbin and Sturat (1954) found that the weighted results resembled the responses of the initial interviews more than the call-back responses. Simmons (1954) also found substantial biases in the weighted responses. In addition, the cost of obtaining weighted responses can often make the Politz scheme less economical than call-backs. Consequently, it appears that the weighted first call procedure has serious problems of practicality and validity and should only be used with discretion. However, appropriate data are needed from large student populations before the Politz scheme can be adequately evaluated from the standpoint of educational research.

Although substituting missing respondents with individuals drawn from an alternate sampling list is often suggested as a solution to nonresponse, this technique is of questionable value and may, in fact, actually exacerbate bias. Imagine, for example, that each element in the sample has a certain probability (1-K) of participating in the survey. If the analyst decides to make X number of calls (X>1), then the initial call will obtain an overrepresentation of people with large probabilities of participating and the follow-up calls will contain an increasingly larger percentage of respondents with lower probabilities. Thus, by substituting respondents on the basis of their availability, the analyst will usually substitute a low probability participant with a high probability participant.

Kish (1965) has suggested a replacement procedure that attempts to alleviate this problem:

> In this plan we include with the new survey addresses some nonresponse addresses from an earlier survey which had similar sampling procedures. Thus interviews from addresses that were nonresponses on former surveys became replacements for nonresponse addresses in the current survey (p. 560).

Whether this replacement strategy can be applied to educational surveys remains to be seen. However, it would appear that this technique has only limited applicability. First, the replacements would have to come from similar schools and be of the same grade or class and possibly of the same academic major as the missing respondents. Furthermore, the

survey that is used to draw the replacements for another study would need to be conducted during the same time period. In addition, for many studies this strategy would be economically infeasible if the replacement survey was conducted in a different geographical location that would require additional traveling and field work expenses.

Nevertheless, when the analyst can avoid the problems associated with substitution, this technique can be of value in reducing nonresponse bias. However, no method of substitution is entirely free from disadvantages and the apparent gains of this procedure can frequently be outweighed by the costs of introducing further bias. Consequently, when a substitution policy is employed, the investigator should acknowledge this fact and report the procedures and the extent to which it was used.

Weighting subsamples inversely to their response rate is another technique that can be used to partially correct for nonresponse. For example, if there is a 40 percent questionnaire nonresponse rate from students with a poor academic record, the analyst could weight the responses from this subgroup by a factor of 10/4. It should be noted, however, that this procedure is useful only to the extent that the criteria used to weigh the responses are in some way associated with the variable to be adjusted. Thus, if a student's academic record is not related to his evaluation of the school administration, it will be of little value to use this

variable as a weighting criterion in an analysis of student attitudes toward the school administration. In addition, weighting procedures are a form of substitution in which the analyst substitutes missing respondents for respondents. Consequently, the same caution recommended in the use of substitution also applies in this context.

Simply preparing tabulations and statistical tests solely on the basis of the existing responses is an unaccepted method of dealing with questionnaire nonresponse. In effect, the analyst assumes that the nonresponses are randomly distributed throughout the sample so that the nonrespondents are identical to the population of respondents. It is extremely difficult, however, to determine on the basis of speculation whether a significant nonresponse bias has been introduced into the data. Such a strategy is justified only when the nonresponse rate is so low that even the most systematic differences between the respondents and nonrespondents would not alter the observed relationships. Because this assurance is so rare in survey analysis, particularly in educational surveys, it is recommended that researchers empirically determine the extent to which nonresponse bias has influenced the distribution of responses.

Finally, it is important to bear in mind that the problem of questionnaire nonresponse is not solved by starting with an excess number of cases to allow for shrinkage. The sample will still contain a disproportionate number of respondents

with low probabilities of participating in the survey. If they differ systematically from respondents with high probabilities of participating, then the nonresponse bias will still exist.

Item Nonresponse. A number of studies evinced considerable fluctuation in rate of response to individual items. However, little attention was given to the import or implications of these fluctuations, although the lack of response to a particular questionnaire item represents a special form of nonresponse. Item nonresponse may be the result of an oversight on the part of the interviewer or a deliberate or inadvertent nonresponse in the case of self-administered surveys. If these omissions are the result of oversight, the analyst can usually reduce their frequency by carefully organizing the questionnaire or the interview session. For example, the investigator can often reduce these oversights by instructing the interviewers to check over the questionnaire before they leave the respondent's house, or asking the respondent to check over his answers before he leaves the interview session or mails in the questionnaire. In addition, all the responses should be coded in the margins of the questionnaire so that the interviewer or the respondent can easily locate any omissions by quickly scanning the page. Using machine-scored test sheets serves the same function. When anonymity is not an issue in the survey, a phone call to the respondent can rectify many of these oversights. When the

survey instrument is to be self-administered in a central location such as the school cafeteria or auditorium, the item nonresponse rate will be lower if one of the field staff checks the questionnaires as they are turned in so that students or faculty can answer any questions they have missed before they leave the premises.

Reducing item nonresponse rates becomes much more delicate when the omissions are deliberate. If the respondent initially refuses to answer a particular question for personal reasons and the researcher unwittingly attempts to force him to respond by calling attention to the oversight, the investigator is likely to get a socially desirable response rather than a true response. In other words, there exists a possible conflict between the effort to decrease item nonresponse and the desire to collect valid information. No adequate solution exists for this problem other than avoiding questions which may be judged overly personal by the respondent. In the event that questions of this type may be included in the questionnaire, the analyst should advise the respondent that he is free to skip any item he considers offensive.

Certain items, such as father's income and education, are frequently beyond the knowledge of young respondents. Consequently, items inquiring into matters such as these must be used with caution. Bowles and Levin (1968A), for example, report that serious item nonresponse occurred in the Coleman study.

> The nonresponse rates for mother's and father's education are particularly important since parents' education represented a prime control for student's social class. Nonresponses on father's education were about 50 percent for first graders, 40 percent for third graders, 41 percent for sixth graders, 21 percent for ninth graders, and 11 percent for twelfth graders. Nonresponse rates for mother's education were as high as those for father's education at grades 1 and 3, and represented 33 percent, 15 percent, and 7 percent at the higher grades. Nonresponse rates for other background variables were also high (p. 7).

Furthermore, there is evidence to suggest that this item nonresponse was not random. According to Bowles and Levin, preliminary analyses indicated that achievement test scores of nonrespondents on these particular items were generally below the mean scores of the respondents.

Since the treatment of item nonresponse in survey analysis is usually a nuisance, most analysts ignore the nonresponses altogether. Whether this is sound methodological practice in educational research is questionable, particularly since systematic differences have been found between respondents who omit items and those who do not (cf. Ferber, 1966; Gergen and Back, 1966). The assumption of no difference appears particularly tenuous when refusals are not a legitimate response category and when the items may appear to be overly personal to some respondents. Consequently, if these conditions exist, or if the item nonresponse is high enough to confound specific interpretations, the analyst should correct for the item nonresponse.

One correction procedure is to assign each omission or nonresponse a value equal to the arithmetic mean of those respondents who answered the particular question and have certain characteristics in common with the nonrespondents. Thus, the investigator might treat the omitted response as a dependent variable in a multiple regression equation. Or, the investigator might assign each [illegible] the mean response value plus a random component so that these omissions will be reasonably distributed throughout the response distribution. Such procedures have two purposes: (1) to remove bias from the sample estimates; and (2) to simplify tabulating procedures. However, since the use of these strategies delays the data processing, it may be best to simply ignore the nonresponses when the item response rate is high. However, as mentioned previously, a better alternative is to have contact officers return to the subjects, or to a random subsample of subjects, to obtain the missing information on selected problematic items.

Sample Mortality. Systematic sample mortality, like the previously discussed forms of nonresponse, can introduce serious bias into the analysis. Moreover, first wave biases often become intensified in later waves. For example, the response bias in favor of higher academic aptitude students in the first phase of Thistlethwaite's study increasingly favored higher ability students in the second and third waves. Such biases cast doubt about the independent impact of the various college press factors studied.

Although some sample mortality is to be expected, careful planning can often reduce the frequency of this form of nonresponse. For example, periodic phone calls or postcards to update the sampling list will impress upon the respondent the importance of his participation. Secondly, by keeping in contact with the respondent, the investigator will be better informed about any of the respondent's plans which might subsequently influence the nonresponse rate (e.g., vacations, school transfers, residential mobility). As a general rule, the longer the delay in contacting missing respondents, the greater is the probability that they will remain lost. Consequently, all large scale longitudinal studies should employ particular methods or have personnel available to maintain contact with the respondents. For example, both Flanagan and associates and Bachman and associates used project newsletters as a means of keeping in touch with their sample. Bachman and associates mailed a project newsletter to each subject at six month intervals along with a fact sheet asking him to answer brief questions about his current educational and occupational status and requesting his current address.

One of the best ways to safeguard against sample mortality is to collect extensive information about the respondents including the names, addresses, and occupations of spouses, siblings, neighbors, and friends. These people can often be instrumental in tracking down missing respondents. In addition, the respondent's social security number, driver's

license and selective service number can also be obtained for this purpose.

The most concerted effort in the Analytical Review studies to increase the response rate was that of Flanagan and associates. Some of the techniques employed in Project TALENT are enumerated below:

1. Posters were distributed in the schools emphasizing the importance of the project.

2. Identification cards certifying membership in Project TALENT were given to each participating student. According to the researchers, the membership cards were intended to engender a feeling of personal identification with the project.

3. Participating students were given an explanation of the importance of the project, its nationwide scope and long range goals.

4. Four mailings of the questionnaire were made approximately one month apart to nonrespondents with reminder postcards between the first and second waves.

5. Intensive follow-up of a random sample of nonrespondents included checks of local telephone directories and information operators, city directories, parents, relatives, employees, former neighbors, teachers, guidance counselors, chairmen of class reunion committees, former classmates living in the community, the Department of Motor Vehicles, banks, finance companies, voter registration records, marriage license bureaus, police records, and income tax or personal property tax bureaus.

These techniques appear especially important since, as the researchers remarked, a special problem contributing to nonresponse was the inaccurate personal information supplied by the students at the time of the original test administration. They recommended that investigators provide for vigorous supervision of the students and allow ample time for students to perform these information-giving tasks. In this context, as indicated above, personal follow-ups, particularly through telephone and individual personal contacts, are most productive in reducing attrition rates.

One of the most common practices used in estimating the biases of sample mortality is to compare the initial measurements of those who were successfully reinterviewed with those who were not. If the two groups are comparable initially, then the investigator has some evidence suggesting that the group may also be comparable at the time of the later measurement. However, when the responses are not similar, weighting procedures are frequently employed to adjust for the nonresponse bias. Astin and Panos used this technique to determine that the less academically able students from relatively less educated families were less likely to return the mailed questionnaire than students with high academic ability from well educated families. Thistlethwaite also compared the initial scores of the respondents and nonrespondents to the National Merit Scholarship Qualifying Test (NMSQT) and concluded that there was a slight nonresponse bias associated

with geographical region (proportionately fewer returns from the South) and a somewhat greater bias associated with aptitude level (proportionately more returns from high-aptitude students).

This technique assumes that the characteristics and experiences of the respondents and nonrespondents will be similar at a later date due to their initial similarity. However, the experiences and characteristics of the individual change through time. Not only do people's aspirations, expectations, and attitudes change, but their social and physical environments also change. Individuals who shared similar attributes at one moment will subsequently be exposed to different stimuli and new situations. Consequently, it is questionable whether the analyst can assume that these subsequent experiences, or the subjects' reactions to these experiences, will be similar because they initially had similar characteristics.

This is of particular importance in educational research since students are constantly being exposed to new people and new ideas. Moreover, the experiences that intervene between measurement periods are likely to differ. Thus, while no systematic differences in the initial measurements may be found between respondents who were successfully restudied and those who were not, these intervening experiences may differ. Consequently, comparing the initial responses of nonrespondents with respondents should not be considered

definitive evidence of similarity. The problem of sample mortality should be confronted more directly. Every effort should be made to reduce this mortality before data analysis begins.

Item Mortality. Item mortality can also be a serious problem in panel studies. Experience gained from the previous survey should allow the investigator to reduce inadvertent oversights by redesigning the questionnaire or rewording the instructions. In addition, examination of the data can assist the analyst in detecting deliberate nonresponse. If the same question or item elicits a nonresponse on succeeding measurements, the analyst can be more certain that the nonresponse is deliberate, especially if he redesigned his questionnaire to reduce inadvertent oversights. The questions themselves, however, should not be reworded or even placed in a different order. It is essential that the instrument used in panel analysis be standardized across measurement periods so that responses will be comparable. Of course, if certain questions do not prove heuristic, that is, the responses are sufficiently skewed or one-directional to make the question unsuitable for analysis, the question should be changed if the analyst feels the same response distribution will occur in succeeding surveys.

Basically, the same techniques for dealing with item nonresponse apply to item mortality. In addition, the analyst can refer to the initial response to develop a weighting

system. As in sample mortality, however, the investigator should not weight the responses if the item mortality is trivial.

One of the major problems of correcting for nonresponse is the dearth of empirical data on nonrespondents. Researchers frequently have little information concerning the parameters of the nonrespondent population. Until these data become available, there may be some gain in weighting responses on the basis of criteria derived from particular studies. In the meantime, more research might well be conducted to ascertain the characteristics of nonrespondents. With the accumulation of this type of data, it may be possible to devise a system of weighting procedures based on continuous, systematic knowledge rather than on the basis of evidence supplied from a single survey, keeping in mind the problems of weighting responses discussed above.

More specifically, on the basis of the review of the present studies and a number of other studies considered, certain items are manifestly problematic in this respect. Consequently, more attention ought to be given to those important items proven to be problematic in previous research so that they can be dealt with more productively in future research. Ways of dealing with these items have been considered in the previous section. In addition, it is obvious that more research should be conducted on the dynamics of nonresponse to selected items so that the problems under discussion can be eliminated to a greater degree than is apparent to date.

Response Error

Even when the nonresponse rate is low, the analyst must still be concerned with possible biases in the data which result from response or measurement error. In discussing the problem of response error, it is useful to conceptualize the existence of a "true value" representing the correct score on some characteristic or attribute measured in the survey. The response error may be defined as the difference between the reported or measured value and the true value.

Response error should be considered in terms of the proportion of the sample reporting erroneously. If a very small proportion of the sample reports large errors, the distribution of responses is not likely to be affected. Furthermore, response error will not be a major problem in the survey if the errors are randomly distributed throughout the sample. Random errors can be expected to cancel one another out, resulting in a mean value approximating zero and having little impact upon the sample estimates. If, on the other hand, a large proportion of the sample reports errors of any magnitude, then the distribution of responses could be significantly altered from the true values. Moreover, if the errors are systematic, then serious bias can be introduced into the data. Since most, if not all, survey items contain some error or imprecision in their measurement, it is crucial that the magnitude of error remains at a manageable level.

There are three primary sources of systematic response error in survey research, each representing a major component of the interview situation: the respondent, the interviewer, and the questionnaire. These sources of systematic error are discussed below.

The Respondent. A major source of measurement error in survey research is due to intentional or unintentional errors in reporting on the part of the respondent. Deliberate or conscious response errors in reporting usually stem from three major conditions: (1) when psychological benefits can be gained from distorting responses; (2) when the respondent is distrustful of the interviewer; and (3) when the presence of a third person makes response distortion rewarding to the subject.

There are two basic types of psychological rewards derived from response distortions; satisfying a need for social approval and a need to conform to the expectations of the interviewer. Distortions due to the desire for social approval stem from the motivation to answer in such a way that is consistent with, or approved by, the respondent's peer or reference group. As a case in point, Parry and Crossley (1950) report that 23 percent of their respondents said they voted in a 1944 election when they actually had not done so. Similarly, Cahalan (1968) reports that 28 percent of the respondents in Denver misrepresented their vote in a mayoralty election. Bell and Buchanan

(1966), Clausen (1968) and Weiss (1968) report similar results in studies of voter registration and behavior.

Socially desirable responses are also found in a number of studies of deviant behavior. Robins (1963) determined that 42 percent of those who had adult nontraffic violations denied them. Twenty-nine percent of those who had been truants as children failed to admit it, 23 percent who had not attended high school claimed they were high school graduates, and 13 percent of those who were divorced denied the fact. Weiss found that 37 percent of the mothers in her sample inaccurately reported whether their children received a failing mark on their last report card. Gould (1969) found very little correspondence between self-reported acts of delinquency and official records. Clark and Tifft (1966) used the threat of a polygraph test as a criterion measure and reported that all respondents underreported the frequency of their misconduct. Furthermore, the researchers learned that in 66 percent of the instances in which individuals did not initially admit a behavior but did later, an act had been committed that was "never permissible" according to the perceived standards of their reference group. Ball (1967) reports similar results in a study of drug addicts.[5]

---

[5]For further evidence and discussion of the socially desirable response set, see K. R. Athey, J. E. Coleman, A. P. Reitman and J. Rang (1960); E. Bryant, I. Gardner and M. Goldman (1966), D. P. Crowne and D. Marlowe (1960), A. L. Edwards (1953,1957,1961,1963); A. L. Edwards, C. I. Diers and J. N. Walker (1962); A. L. Edwards and J. A. Wash (1964); J. B. Taylor (1961); and H. D. Willcock (1951).

Distortions due to self-enhancement in the presence of the interviewer are a similar source of response error. In this situation the respondent attempts to enhance his self-image by responding in a manner that is believed to be approved by the interviewer rather than by the respondent's peer group. A study by Parry and Crossley (1950) illustrates this problem. In this study, 900 interviews were collected from a single community. Analysis of the data indicated that the interviewers did not significantly differ from one another in the responses they collected. However, significant differences were observed between the reported data and the criterion data obtained from the appropriate agency, as illustrated in Table 11.

Table 11

Response Error to Specific Items in Parry and Crossley (1950), in Percent

Item	Percent of Respondents Giving Inaccurate Reports
1. Contributing to Community Chest	40
2. Voting	25
3. Possessing a library card	10
4. Possessing a driver's license	10
5. Owning a home	4
6. Owning an automobile	3
7. Possessing a telephone	2

According to Parry and Crossley, items 1 and 2 were behaviors heavily associated with the desire for social approval. The 40 percent and 25 percent response error rates were entirely in the direction of the socially desirable response. Items 3 through 7, on the other hand, were felt to be related to self-enhancement, since possessing these items would enhance the respondent's self-image in the presence of the interviewer.

The studies by Bachman and associates and Coleman, among others, contain items susceptible to a socially desirable or self-enhancing response set. Bachman and associates, for example, administered a thirteen item "Rebellious Behavior in School" scale and a twenty-six item "Delinquent Behaviors" scale, in which the students were asked to check the frequency with which they had engaged in certain socially unapproved behaviors. It seems clear that indices of this type are prone to underreporting. Coleman asked his students to indicate whether certain items such as encyclopedias, magazines, vacuum cleaners, telephones and television sets were in their homes. In addition, students were asked how often they attended a public library and did homework each night. These questions also tap socially desirable aspects of the student's life and are therefore subject to response error.

Although socially desirable and self-enhancing responses are frequently treated as being synonymous, they are often in conflict with one another. In projecting an acceptable

self-image to the interviewer, for example, the respondent may have to deny or distort the attitudes and behaviors rewarded by his peer group. Conversely, the socially approved response from the standpoint of the respondent's reference group may represent an inappropriate response in terms of what the respondent believes the interviewer defines as appropriate.

Extensive literature on this subject clearly indicates that the probability of obtaining a socially desirable response is largely determined by the racial and social class match between the respondent and interviewer. For a comprehensive bibliography on this problem, see L. Bauman et al. (1970).

There is also evidence to indicate that the respondent will attempt to win approval from the investigator by distorting his responses in a direction consistent with the expectations of the investigator. Rosenberg (1965,1969), for example, has been concerned with "evaluation apprehension," which he defines as:

> an active, anxiety-toned concern that he (the respondent) win a positive evaluation from the experimenter, or at least that he provide no grounds for a negative one (1969, p. 281).

Rosenberg has demonstrated that evaluation apprehension can significantly influence the results of a study. Friedman (1967), Rosenthal (1966) and Rosenthal and Rosnow (1969)

investigated the extent to which the laboratory experimenter can unconsciously transmit to the subject his own expectations of the experimental results. The data from these analyses indicate that the expectations of the experimenter can have a large impact upon the results of an experiment.

Unfortunately, there is a dearth of information concerning the effects of interviewer expectations. Rice (1929), however, reports a study which investigated the perceived causes of destitution. One interviewer, a prohibitionist, obtained three times as many responses indicating alcohol was the perceived cause of destitution as did a second interviewer, who was an acknowledged socialist. The socialist, on the other hand, obtained significantly more responses blaming social conditions as the cause of destitution.

Katz (1942) reports similar results in a study of attitudes toward sitdown strikes. The interviewers of working class origin obtained responses from 40 percent of their respondents favoring a law against sitdown strikes, while middle class interviewers obtained favorable responses from nearly 60 percent of their respondents. Ferber and Wales (1952) also found evidence linking interviewer attitudes to differences in obtained response patterns. In their study, the investigators found that interviewers who favored prefabricated housing were more likely to collect responses favoring this type of construction than were interviewers who had unfavorable attitudes toward prefabricated housing.

Distrust of the interviewer can occur for a number of reasons. The respondent may be suspicious of the agency which the interviewer represents, or feel the interviewer has certain undesirable attitudes, prejudices and expectations. In addition, as will be seen in the following section, the personal characteristics of the interviewer can cause the respondent to be apprehensive. Racial and social class distrust are probably the most familiar and common forms of this type of response distortion.

The respondent can act in one of several ways when he distrusts the interviewer. He may feel timid and insecure and attempt to answer in a noncontroversial and self-enhancing manner. The respondent may also feel he has no reason to respond correctly and give little consideration or thought to his responses. Or, the respondent may use the interview as an outlet for hostility and lie systematically so that the data are rendered useless.

The presence of a third person during the interview can also result in response error. This is especially true when the third person represents the respondent's reference or peer group. The chances of obtaining a socially desirable response will naturally increase if a representative of the respondent's peer group is present at the interview. However, the presence of a third person can also reduce error. (See Philip Taietz, 1962, for an analysis of this possibility.)

The most common and serious form of unintentional response error is selective perception or faulty memory. According to Neely (1937), failing to remember an event or experience is likely to occur when there is an unconscious motivation to suppress the experience, the event is not salient for the respondent, or the event changes over time. That is, unconscious response errors frequently occur when the respondent is asked to recall a situation or experience that is emotionally painful or threatening, unimportant, or continually changing.

Cannel (1961), for example, reports that threatening health histories are significantly less likely to be remembered than nonthreatening histories. Neely (1937) found that a less important experience (time lost from school due to an automobile accident) is less easily remembered than a more important experience (time lost from work due to an automobile accident), presumably because the costs incurred in lost work are more salient to the individual. Weiss et al. (1960) found considerable response error in recalling job histories over a five year period that involved at least three job changes, while Cannell (1961) reports that the tendency to forget health histories increases as the time between the incident and the attempted recall is extended.

In addition, the sex and age of the respondent can influence the accuracy of the reported data. Sex differences

in accuracy of information are a function of sex role. Respondents answer questions more accurately when the items refer to their own experiences. Finally, older respondents are more likely to forget or unintentionally distort their recall of previous experiences than are younger respondents.

Acquiescence is another form of response distortion in which the respondent tends to agree with an item regardless of its content. Bass (1955) has argued that the F Scale used to measure authoritarianism is actually a measure of acquiescence. Similarly, Messick and Jackson (1961) conclude that an acquiescence response set is largely responsible for the scores on many psychological instruments.[6]

Many psychometricians believe acquiescence to be a personality characteristic, while others maintain it is a function of the interview situation.[7] When it is conceptualized as an attribute of the respondent's personality, then it leads to unintentional response error. However, it is equally probable that acquiescence also arises from the

---

[6]For additional evidence concerning acquiescence, see I. A. Berg and G. M. Rapapert (1954); L. Bauman et al. (1970); L. J. Cronbach (1946); C. W. Gray and H. Crisp (1961); D. S. Jackson and S. Messick (1962); and R. K. McGee (1962).

[7]See A. Couch and K. Keniston (1960,1961); A. L. Edwards (1963); A. L. Edwards and J. N. Walker (1961A,1961B); D. S. Jackson and S. Messick (1958,1961); I. Mahler (1962); S. Messick (1962); D. R. Miklich (1965); D. Peabody (1966); L. G. Rorer (1965); R. E. Schutz and R. J. Foster (1963); L. J. Stricker (1963); and J. B. Taylor (1961).

interview situation. In this case, acquiescence could be used by the respondent to intentionally distort responses or to expedite the interview. This is especially likely when biases exist due to racial distrust.

There are four basic methods of dealing with response errors that arise from the characteristics or attributes of the respondents: (1) improving the administration of the survey questionnaire; (2) modifying the questionnaire; (3) detecting and discarding subjects who respond erroneously; and (4) correcting the scores of all subjects in proportion to the amount of their estimated response error. Each method is briefly described below.

1. Improving the administration of the questionnaire: It is important in every survey that the respondent be placed in a nonthreatening situation. In addition, an appeal is usually made to convince the respondent of the importance of his participation and to assure him that his responses will be kept in strict confidence. To reduce distortions due to self-enhancement or social desirability, it is wise to stress the importance of valid responses. It is also desirable to get the respondent to identify with the survey and feel that he is making an important contribution which is very much dependent upon the validity of his responses. To increase the respondent's feeling of identification and participation in the project, for example, he could be told that the questionnaire is in preliminary form and any comments he has

concerning its clarity or relevance would be greatly appreciated. Flanagan and associates gave each member of their Project TALENT sample an identification card signifying their participation in the study in order to increase the subject's feeling of involvement.

In order to reduce any distrust the respondent may have toward the interviewer, it is essential that the interview be task oriented and that the task be made acceptable to the respondent. Endorsements from popular community or national leaders can help establish the legitimacy and importance of the survey. Thus, Bachman and associates, Flanagan and associates, and Trent and Medsker solicited the cooperation of community leaders, school administrators and newspapers to help establish the importance of their studies.

In dealing with the problem of interviewer expectations, the respondent should be told that there are no right or wrong answers and that the survey is essentially exploratory. Biases resulting from faulty memory, on the other hand, are largely unconscious errors and thus more difficult to confront. However, careful probing in conjunction with a supportive environment can often help trigger a respondent's memory.

2. Modifying the questionnaire: Frequently, efforts can be made to reduce response error by designing the questionnaire, or specific items, in a particular manner. For example, the analyst may be able to "legitimize" a socially

undesirable response by introducing questions with such statements as: "Some people believe . . ., while others feel that . . .; what do you think?"

The acquiescence response set may be counterbalanced by using equal numbers of direct worded and reverse worded questions. The forced-choice questionnaire has been used by Edwards (1957) to reduce the effects of a socially desirable response set. Subjects are asked to select one item of a pair in which both items have been equated for "desirability." Concealing or disguising the purpose of the question may also reduce conscious error response (cf. Campbell, 1950; Cook and Selltiz, 1964).

3. Detecting response error: The analyst can also attempt to identify and then exclude from analysis subjects who have erred in their responses to a particular item. Hyman (1954) identifies two general classes of methods for detecting response error: methods involving internal checks and those involving external checks.

Internal checks are based on the assumption that the validity of one response can be inferred from the response to some other question or item in the questionnaire. For example, an internal check on the question "How old are you?" would be the respondent's reply to "When were you born?"

One of the most common internal checks on response error is the open-ended question that requires the respondent either through a series of interlocking or contingent

questions or through carefully designed interviewer probes to elaborate on his initial response. The information gleaned from these questions can often provide the analyst with valuable insights into the validity of the original answer and the degree to which the overall response was consistent or confused. Some surveys do not include open-ended questions because they are both difficult and expensive to code. However, when response error due to a lapse in memory is suspected, or when questions refer to objects or situations that the respondent has probably not thought a great deal about, the data from these questions can usually be of great value to the analyst in ascertaining the more subtle nuances of the response referent as well as the response.

Using open-ended questions will not be an effective error detecting device, however, when the errors are deliberate because the responses will not be independent of one another. One possible solution to this problem is to separate contingency or matched questions from one another in the questionnaire. A related method suitable for longitudinal surveys involves repeating the item in future surveys. Thus, while the student's estimate of his grade point average should not change between surveys, depending upon the time lapse between them, the analyst may decide to ask the student this question again so that the frequency of response error in self-estimates of grade point average can be determined.

Even this technique, however, will not usually detect the response errors made by clever respondents or individuals who consistently distort answers for reasons of ego defense, self-esteem, or mischief. The analyst can, of course, employ methods which are more covert in hopes that the responses will truly be independent of one another. Yet, this practice is limited by the fact that the theoretical connection between the two items may be more apparent than real. In other words, according to Hyman, ". . . the less apparent the connection to the respondent, the more the possibility that the expected relationship between replies is truly indeterminant or tenuous" (1954, p. 154).

For example, if the researcher suspects that one norm of the dominant student subculture at a particular school sanctions liberal political activism to the extent that his political ideology scale will elicit a large percentage of normative responses from the students rather than true political beliefs, the analyst could introduce an internal check by asking the students a series of forced choice or fixed alternative questions that depict a number of legal cases which the student must litigate. If carefully designed, these hypothetical cases could serve as a projective test in which the student's political ideology will influence his response. This particular check will likely be effective against motivated response error. It is covert in nature and the respondent is not likely to see the

connection between the courtroom scenes and the political ideology scale. However, it is apparent that the link between an individual's political beliefs and the way he adjudicates certain legal cases is somewhat unclear since the correlation between the two is undoubtedly far from perfect.

Another internal check that partially circumvents this problem involves checking the responses of subgroups which have been differentiated on the basis of some reliable factual characteristic that should influence the respondent's answer to the particular item or scale. Thus, the analyst would expect that respondents who are known to be active members of liberal political groups on campus will score significantly higher in political liberalism than the student body at large. A particularly covert method of identifying motivated response error would be to include an item in the questionnaire that is designed in such a way that a specific response will necessarily be invalid. Thus, an investigator measuring student political activism may ask the respondents if they participated in a certain "large and well-known political demonstration" that never existed.

A similar type of check has been used in identifying the socially desirable response. The response categories to an item are dichotomized into agree-disagree, yes-no, etc. One response is presumed to be factually true for everyone, but undesirable, while the other response is

highly desirable, but contrary to fact. This technique provides a relatively straightforward, yet covert, method for detecting the socially desirable response set.

Acquiescence can also be checked in this manner. By repeating the same or a similar question with a reverse wording, the analyst can detect respondents who were inconsistent in their responses. An example is where agreement to the first item signifies the respondent's acceptance of the issue contained in the item and where agreement to the matched item signifies its rejection.

Ways in which the researcher can identify or estimate response errors for specific responses have been discussed up to this point. However, if the response error is motivated, other replies are also likely to be distorted. Consequently, the researcher must be concerned with identifying motivated response bias. One possible method of identifying "motivated response bias" begins with the assumption that response errors due to confusion or misunderstanding on the part of the respondent are less likely to be repeated in the survey than are motivated response errors. Thus, by examining several different questions for response error, the analyst can identify those respondents who consistently erred in their replies.

While the logic of this strategy is sound, there are serious problems associated with its implementation that reduce its feasibility for many studies. First, this

technique requires that the researcher have available for analysis at least one internal check for each question of concern to be used in detecting motivated response bias. In addition, each of these questions must be sensitive to the same type of motivated response. Furthermore, if consistent response errors across different questions are the measure of motivated response, it is crucial that all internal checks be commensurate with one another in detecting this bias so that the data will be comparable.

A more practical solution is to measure and then statistically control for the motivation which is believed to bias the response. Thus, Bachman and associates included in their survey instrument a scale that attempted to measure the respondent's propensity to elicit socially desirable or approved responses. However, the researchers failed to use the measure systematically as a test for response error.

The basic rationale of the internal check is the assumption that responses to the check item are less susceptible to response error. When this assumption can be defended the analyst can intelligently speculate which reply is likely to be more valid when inconsistent responses are observed. If such an assumption is tenuous, however, the problems associated with interpreting response discrepancies become more complex because the response inconsistency will not clearly indicate which replies are the most valid. If this situation occurs when the response inconsistency is

small, probably the best strategy is to ignore it. If, on the other hand, the discrepancy is large enough to make specific interpretations problematic, the analyst should inform the reader about the possible existence of response error and illustrate the implications of this confounding factor by deleting the inconsistent responses and recalculating the statistics accordingly.

The major weakness of internal checks is that they are not ascertained independently of the testing situation. Consequently, there is a greater probability that the responses to the check items will be contaminated by the same factors operating in the survey that influenced the replies to the other questions. By using external checks, however, the analyst can usually overcome these problems.

The ideal external check in educational research is to cross-validate survey responses with comparable information contained in official records such as school files. Thus, Bachman could have estimated the degree of response error to some of the more extreme items on his deviant behavior scale (especially underreporting) by comparing the student's responses with the school records. Another important source of information is the student's family and friends. Educational surveys often ask the student to report on his home life or his interaction with peers. The analyst can then check the validity of the responses he receives by interviewing the student's family and friends directly. Coleman,

for example, compared the responses of 700 students in two school districts in Tennessee with information that was taken from school records and the students' parents to determine if the matched sets of data were in agreement. Similarly, Bachman and associates cross-validated self-reported grade point average with school records and determined that both measures were in agreement.

When a fairly large percentage of students is sampled from each school in the survey, a useful technique for checking the validity of responses dealing with the student's peer relations is to ask each student to list the names of his closest schoolmates. Then, by analyzing the replies of the friends that were included in the sample, the analyst can make some judgments concerning their consistency. The investigator can also make such sociometric patterns an explicit aspect of his research design.

Frequently, educational researchers interview teachers as well as students. Usually, the survey instrument administered to the teacher shows little similarity to the one administered to the student. However, the responses of the teacher to questions that are comparable to the ones asked of the student can often be conceptualized as an external check to the replies of the student and vice versa. But the investigator should be cautious in using this procedure; inconsistencies may not indicate response error as much as contrasting perspectives, definitions and value orientations of the occupants of these two status positions.

These techniques are often prohibitively expensive and too time consuming to use for every individual in the sample. However, the analyst does not require a complete enumeration of a particular external check before inferences can be made. Drawing a random subsample of responses for comparison with some external check will prove extremely valuable in estimating response error. Unfortunately, however, not all or even most of the responses can be checked in such a straightforward manner.

A second type of external check which has great potential in educational surveys involves using the "split ballot" or alternate form procedure. When the analyst suspects that the specific wording of questions or the sequence in which they are asked may influence the distribution of responses, the sample can be randomly divided into two or more groups with each subsample receiving a different version of the survey instrument. Then, by comparing the responses of the different groups, the researcher can determine the degree to which the findings are independent of the specific procedures that were differentially administered to the groups.

An equally promising method for externally validating the responses of a particular sample is to perform the same basic study on equivalent samples using different personnel and survey instruments. Thus, instead of having one large survey to investigate the impact of the junior college system on lower income students, two smaller but independent

surveys designed to investigate the same phenomena might prove more useful. If the relationships uncovered in one study are truly independent of specific methodologies, then the sister survey should also report similar results.

4. Correcting scores: Occasionally, analysts attempt to correct for response error that originates from the respondent by using the detection criteria to adjust raw scores according to the amount of systematic error that can be attributed to the biasing factor. Messick (1961), for example, modified a formula originally prepared by Helmstadter (1957) in order to eliminate response error on personality tests which had no a priori correct answers. Another approach is to eliminate systematic response error by including in the survey a "content-free" measure of the bias and using a regression equation to eliminate the contaminating variables. At the moment, however, there is no generally accepted procedure for correcting biased reports arising from the intentional or unintentional distortions of the respondent.

The Interviewer. There is considerable research that indicates that the characteristics of interviewers differentially affect the results of the interviewing process (see, e.g., Cannell and Kahn, 1968; Cicourel, 1964; Phillips, 1971). Chief among these characteristics are the skills of the interviewers, their background characteristics, related attitudes, degree of motivation, expectations of the respondents and their involvement in the data to be reported.

Important skills of the interviewers include not only their facility at raising questions and probes, but their ability to establish rapport with respondents, their insights into responses and perceptions of various cues communicated by the respondents. Influential background characteristics include social status, race, religion, sex, age, values, attitudes, beliefs and such personality characteristics as anxiety, need for social approval, hostility versus warmth of feelings and degree of authoritarianism. Underlying motivation is the interviewer's interest in carrying out an effective interview. Degree of involvement in the data to be reported signifies the interviewer's amount of interest in the study in which he is participating and the results of the study.

Singly and together these characteristics have been shown to affect greatly the rapport established between the interviewer and respondent, resultant response bias, the objective and consistent recording of responses and the comprehensiveness, emphasis and accuracy of the responses generally.

For example, Katz (1964), Rankin and Campbell (1955), Hyman (1954), Athey (1960), and Summers and Hammonds (1966) clearly demonstrate that the racial mix between respondent and interviewer can produce systematic response error. As a case in point, Summer and Hammonds report that when both investigators were White, 52 percent of the respondents showed themselves to be prejudiced, while only 37 percent of the respondents were judged prejudiced when one of the investigators was Black.

The evidence relating interviewer sex to response error is equally clear (Binder et al., 1957; Benney, 1956; Friedman, 1967; Sarason and Harmatz, 1965; Stevenson and Allen, 1964; and Stevenson and Odom, 1963). Benney et al. found that the least inhibited communication occurred between people of the same sex and age, while the most inhibited communication occurred between people of different sex but of the same age. The study by Lenski and Legget (1960) is one of many that indicates the social class of the interviewer can produce response error. They report that Black respondents and respondents of low status acquiesced to simple agree-disagree questionnaire items when the interviewer was of middle class background.

There is a distinction between response error that results from the interviewer's biases and consistency of recording (interviewer variance). Errors of carelessness or misunderstanding frequently cancel one another out, but the greater the systematic rather than compensating errors the greater the interviewer variance. Both types of errors are pernicious in survey research and consequently should be assessed for their presence and effect whenever possible. Interviewer bias is greatly more difficult to determine, and is best controlled through the survey design which can include checks against original records and reinterviewing. Interviewer variance can be estimated through replicated or interpenetrating sampling (see Moser and Kalton, 1971).

> . . . with a suitable sample design, response variance can be studied by setting up a theoretical model according to which the response variance arising from different components, notably interviewers, can be estimated. Such models are useful in helping to clarify how response errors may affect the estimates one makes from a survey, both those of the population values and of total variances; and they are helpful in trying to determine the best sample design. Reinforced with cost data, they should make it possible to decide in advance what is the optimum number of interviewers to be used, how much it is worth spending on training and supervision, and so forth (Moser and Kalston, 1971, p. 407).

A number of techniques have been employed to enhance the accuracy and quality of the methodology of interviewing. These include the determination of the extent of interview error; the avoidance of bias producing content and language; the removal of ambiguous or "double-idea" items; checking of items for disparate frames of reference between the interviewer and/or researcher and the respondent; reduction of interviewer improvisation; strict instructions for and control of the interview; selection and assignment of interviewers to minimize undesirable interactions with respondents; training of interviewers in techniques designed to maximize optimum respondent motivation and minimize distorted or disparate cues; the use of "nondirective" or "nonleading" interviewing; and probability sampling of respondents.

Increasing the number of interviewers is also used as a method for reducing interviewer bias. An interviewer who systematically influences the responses he receives will have less impact upon the sample estimates when the number of

interviewers is increased. Finally, matching respondent and interviewer on key personal characteristics is often suggested as a method for reducing interviewer bias. Common sense matching should occur (e.g., matching on the basis of race in a study of racial prejudice); however, there is little data to indicate that matching in less obvious situations will reduce interviewer effects. In fact, there is evidence suggesting that too good a match can produce response error. Weiss (1968), for example, reports that interviewers who established a high degree of rapport with their respondents were more likely to obtain biased responses than interviewers who established moderate rapport with their respondents.

Although half of the studies under review included interview material, scant attention was given to the reliability and validity of the data, particularly with respect to possible distortions of the data resulting from the characteristics of the interviewers. Kagan and Moss did report rater reliabilities in their treatment of their interview data. Trent and Medsker refined the items and language used in their interviews on the basis of extensive pretesting. Their interviewers also went through presurvey training and rater reliabilities were calculated on their interview data. This information was not reported in the literature, thus preventing future researchers from profiting from these experiences and from evaluating them. Other studies

under review may have employed these and other techniques to increase the quality of their data, but again, as long as their efforts remain unreported, they must remain unevaluated and of no use to future research. These matters bear on yet another issue. According to Cannell and Kahn (1968, p. 583):

> These issues are so mundane that they are seldom discussed in the research literature; yet they are so important that they are never left undiscussed when interviewing for social research is being attempted on any substantial scale. Gross underestimation of time and costs involved in data collection and in preparation for data collection is all too common in social research. These matters are no less amenable to empirical study than other aspects of research methodology, but unfortunately few relevant studies have been done and fewer published.

The Questionnaire. The mailed questionnaire is one of the most popular techniques used in educational research. It is widely used in educational surveys because this means of gathering information permits wide coverage over large geographical areas for a minimum expense in terms of time, effort, and money. In addition, people who are difficult to locate and interview can be reached; the respondent retains a sense of privacy; there is greater uniformity in the manner in which questions can be posed; interviewer effects are avoided; and, finally, the mailed questionnaire affords a simple means of continual reporting over time.

The major disadvantage of this technique, the problem of nonresponse, was discussed in a previous section of

this chapter. Most researchers are aware of and take steps to deal with the problem of nonresponse. More specific problems apparent in the Analytical Review studies concern the clarity and consistency of both questionnaire items and instructions.

For example, a number of items used in Lehmann and Dressel's (1963) study were ambiguous. Students were asked to identify those experiences which "reinforced" their behavior and also those which "modified" it. There is some overlap in the denotative value of these two verbs, leading to the possibility that some of the respondents did not understand the difference. This suspicion is borne out by the subjects' responses. For instance, Group I females identified as experiences which very much reinforced their behavior--being away from home and "bull sessions." These same two items also appeared in the five experiences most commonly cited by this group as very much <u>modifying</u> their behavior. Similar repetitions occur for the other groups. Another ambiguous item, and noted as such by the authors, is a statement with which respondents were asked to agree or disagree: "A college education should place equal emphasis on academic and social development" (p. 58). To disagree allows no interpretation of which area the respondent favors.

Several problems of questionnaire design are found in Thistlethwaite's study. First, although the 1961 and 1963 instruments used in Thistlethwaite's study overlapped in

content somewhat, many items in the scales were worded differently, or the scales contained entirely different statements. This was true, for example, in the comparison of responses to scale 10, upperclass press to lowerclass press, both measuring "Faculty Press for Affiliation." In addition, the students received a different set of instructions regarding the reference group for each form: in 1961, they were asked to characterize the entire college environment (faculty and peers), while in 1963 they were directed to characterize faculty and peers in their major fields only. Thus, the author notes that his hypothesis concerning the differences between lowerclass and upperclass environments may be due to the differential wording of the instructions. Although the author feels that there is some evidence supporting this hypothesis, confirmation is tenuous, and his finding that it is the upperclass presses which exert the strongest pressures on students to seek advanced training is questionable.

To avoid such ambiguity and misinterpretation, items and instructions must be stated clearly and consistently, particularly in longitudinal studies. Otherwise, not only are tne findings invalidated but the focus of the entire study may change. For example, in their first survey, Tillery et al. used Warner's Index to measure socio-economic status, whereas occupational prestige was the measure of socio-economic status in the final questionnaire. The investigators

state that since the meaning of the variables changed from year to year and since there was no continuity of the questions, the focus of the study was constantly changing.

Confusion on the part of respondents which may result in measurement error can also occur if one item in the questionnaire contributes to the biased response to another item. Thus, Thistlethwaite notes that in the 1963 questionnaire form, the two criterion questions regarding aspirations and entry are juxtaposed and that this may have influenced responses; that is, because they are found together, students may have answered questions regarding aspirations and entry similarly (either positively or negatively) regarding disposition to go to graduate school. Thistlethwaite also defined the entry criterion as immediate entry into graduate school, that is, within one year following graduation. The survey questions are phrased so that a student who was postponing entrance longer than one year would be missed. It is possible that many 1963 graduates might have had plans to join the Peace Corps, which was the subject of much public interest at that time, and delayed their entrance into graduate school until after their two years with the Corps was over. This is just one possible source of error in determining factors involved in immediate entry.

In a longitudinal study, the effects of repeated test measurements must be considered. For example, three of

the major tests used in the Lehmann and Dressel studies, the Inventory of Beliefs, the Test of Critical Thinking, and the Differential Values Inventory, were given five times during the four year period. Though random sampling was used in the interval years, every student enrolled took the tests at least three times and some took them four or even five times in a four year period. Repeated test effects might be assumed to be operating; however, the investigators do not deal with the ways in which this repeated effect might limit their findings.

As mentioned previously, test scores are the operational definitions of the variables. Thus the meaningfulness of the findings is directly tied to what the tests are measuring. For example, the Test of Critical Thinking used in Lehmann and Dressel's studies is a test of five specific abilities incorporated in the processes of critical thinking; it is not a test of critical thinking, per se. The discussion of the use of this instrument is brief and the authors do not give any rationale for the use of this instrument. Although the diligent reader may refer elsewhere to learn the technical details of the Critical Thinking scale authored by Prince, the point here is that Lehmann and Dressel do not really justify their use of the scale in terms of the objectives of their study.

Lack of discussion of this kind is important since it leaves unanswered such questions as: in what ways is it

valid to think of critical thinking in terms of the five skills tested by the Test of Critical Thinking? What important cognitive skills are not tested? What skills important in critical thinking have been identified that might be difficult to test for with existing instruments? Were tests other than this one considered?

Apparently, however, the investigators were aware of the problem of test error. In speaking of types of changes, they note: "It was assumed that any subject who appeared to become a poorer critical thinker actually did not but that this phenomenon was due to errors in measurement" (1962, p. 67). Since the findings and conclusions depend so heavily on test scores from this instrument, the investigators' failure to make clear the basis of their choice is a serious omission. The same comment applies equally to all the instruments used in these studies, as well as many if not most others under review.

Finally, there may be problems in interpretation when questions are used which either invoke predictable response biases or obscure objective information. For example, Lehmann and Dressel, as well as Trent and Medsker and Katz and associates, asked respondents to assess their own changes on a number of variables.

There is always some question whether students' self-reported changes reflect true changes. For example, in Lehmann and Dressel's (1963) study, large numbers of

subjects did not believe that their values had changed in the four years covered by the study. The authors, however, did not note whether or not these students had changed according to objective measures and the lack of self-perceived changes may have been due to the students' unwillingness (either conscious or unconscious) to accept their own former beliefs. For example, a student who responded that he was the same as a senior as he was as a freshman with respect to tolerance of others, may not have wanted to acknowledge that four years previously he was less tolerant.

Surely, survey questionnaires must be expertly designed and skillfully introduced in order that the kinds of errors described above may be avoided. (For excellent guides to the selection and construction of questionnaires, see Isaac S. and W. B. Michael, 1971; D. Miller, 1970; and A. N. Oppenheim, 1966.) Most important, questionnaires must be pretested on a group of respondents representative of the survey sample. In fact, according to Isaac and Michael (1971), one of the best ways to develop good objective questions is to administer an open-ended form of the question to a small sample of representative subjects. These more lengthy answers will provide the data from which objective-type answers may be derived.

The Effects of Response Errors. The most common belief concerning the effects of response error is that errors in the dependent variable can reduce the correlation but will

not systematically bias the estimate of the relationship while errors in the independent variable will usually reduce the beta coefficient or the slope of the regression line. In reality, the problem is much more complicated.[8]

Consider the following example in which X and Y represent the observed values of the independent and dependent variable and x and y equal the true values of X and Y. Then, X and Y are equal to x and y plus an error term g and h.

$$X = x + g$$

$$Y = y + h$$

$$\text{and } Y - h = y$$

$$\text{and } X - g = x$$

If the variables are assumed to be related in a linear fashion such that

$$y = \beta X + a$$

then we can calculate the relation between the observed values by simple algebraic substitution

$$Y - h = a + \beta(X - g)$$

$$Y = a + \beta X - \beta g + h$$

$$\text{and } Y = a + \beta X + E \text{ where } E = h - \beta g$$

It is apparent that E is not independent of X because it includes the term $-\beta g$ and g is a component of x. Consequently, the least squares procedure will yield a biased (nonindependent) estimate of a and β even when the sample size is

---

[8]The following discussion is based upon that of Lansing and Morgan (1971, pp. 309-314).

infinite and the mean value of the error term is zero.

To estimate β we need five pieces of information:

X_i, $\bar{X}$, Y_i, $\bar{Y}$ and N where

$X_i = X_1 \ldots X_N$

$Y_i = Y_1 \ldots Y_N$

$\bar{X}$ = the arithmetic mean of the values on variable X

$\bar{Y}$ = the arithmetic mean of the values on variable Y

N = number of cases

Then

$$\beta = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2}$$

The following equation to estimate β includes the error terms:

$$\beta = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y}) + \Sigma (x_i - \bar{x})(h_i - \bar{h}) + \Sigma (y_i - \bar{y})(g_i - \bar{g}) + \Sigma (g_i - \bar{g})(h_i - \bar{h})}{\Sigma (x_i - \bar{x})^2 + 2\Sigma (x_i - \bar{x})(g_i - \bar{g}) + \Sigma (g_i - \bar{g})^2}$$

It is apparent that the estimate of β will depend on

(1) Whether the error terms g and h are independent of one another [i.e., $\Sigma (g_i - \bar{g})(h_i - \bar{h})$]

(2) Whether each of the true variables x and y is independent of the error term in the other [i.e., $\Sigma (x_i - \bar{x})(h_i - \bar{h}) + \Sigma (y_i - \bar{y})(g_i - \bar{g})$]

(3) the size of the error variance in variable x [i.e., $\Sigma (g_i - \bar{g})^2$]

(4) Whether the variable x is independent of the error in x [i.e., $2\Sigma (x_i - \bar{x})(g_i - \bar{g})$]

Obviously, then, one cannot simply state whether the beta coefficient will be biased in a certain direction. As the above equation indicates, the value of each of the five terms may involve either g or h or both g and h. However, by reflecting on the nature of response errors in survey research, especially motivated response bias, some of the possible outcomes of response error under varying conditions can be identified.

(1) When the researcher suspects that the response errors in variables X and Y are not independent, as is usually the case in motivated response bias, the effects will be to increase the beta coefficients if the responses are positively correlated and to decrease the slope of the regression line when the responses are negatively correlated. In both cases, however, the strength of the relationship can be artificially increased due to the correlation between the error terms in each variable.

(2) It often occurs that the true value of one variable is correlated with the error term in another variable. Thus, a researcher might discover that his political ideology scale elicits more responses consistent with the dominant student subculture when the respondent is a senior rather than a freshman. If there is a positive correlation between the true value of one variable and the error term in the other, the beta coefficient will rise, and when the correlation is negative the regression line will drop.

(3) It is also quite common in survey analysis that the true value of a variable will be correlated with its own error term. Thus, the investigator might expect more response error in estimating family income from students who come from low income families. If there is a positive correlation between the true value of a variable and its error term, the beta coefficient will be increased.

(4) Perhaps the most important response error occurs in the measurement of the independent variable. Many independent variables in educational research consist of student attitudes and motivations which are especially susceptible to response error. It is apparent from the equation above that the larger the squared error in variable X the smaller the estimate of the beta coefficient.

Researchers have not adequately investigated the problem of response error or the implications that response error has for their particular studies. There has been almost a complete lack of concern on the part of educational researchers to plan systematically for the collection of data bearing upon this threat to the internal validity of their research. Apparently, many researchers accept as a matter of principle that their questions elicit valid and reliable responses. Such blind faith, however, cannot be defended. Because survey researchers rely upon the quality of the written or verbal report they must continually question and investigate the validity of their data. In

short, the validity of the researchers' inferences can be no greater than the validity of the data used to justify such inferences.

CHAPTER VI

SUMMARY AND RECOMMENDATIONS

The Analytical Review Project undertook as one of its objectives, a critical appraisal of the research design and methodology of the studies reviewed, including an assessment of the conceptualization of the phenomena under investigation, the choice and selection of appropriate samples, the choice of statistical design and data analysis strategies.

The process of assessing these various components was complex and involved the development, collation and refinement of broad based criteria upon which such assessments could be made. Relevant literature in the field of methodology was consulted and a preliminary list of criteria was established based upon Campbell and Stanley's (1963) validity check list for experimental research. By necessity, the process was an evolving one. That is, as work on the evaluation of the studies proceeded, some of the criteria originally conceptualized proved inadequate for survey research and the search for new criteria required additional examination of the literature. As a result, an important "spin off" of this review of longitudinal studies has been the development of a "cookbook" of recommended guidelines for educational research based upon a check list of methodological criteria which serves concomitantly as a means by which survey research studies may be evaluated and also as a beginning collection of evaluative

criteria. In essence, then, what follows is a summary distillation of the key points and recommendations raised in the preceding chapters.

The additional time spent on the development of this check list precluded assessing all of the studies according to each of the proposed criteria. In fact, even after many of the chapters in this volume had been completed, re-evaluations were made and additional criteria and guidelines added based on the problems that were observed in the studies under review.

Many of the problems inherent in educational evaluation, from evaluation of teaching effectiveness to large scale institutional evaluation, stem from the lack of well defined criteria upon which such evaluations can be based. Thus, the one intent of the development of the criterion list of methodology and its accompanying enumeration of recommended guidelines for the conceptualization and implementation of research studies on educational impact is to help fill this lack in educational research.

The criterion check list which follows represents only a beginning. It is anticipated and in fact recommended, that as additional critical evaluations of survey research are undertaken, evaluators accept as one of their research objectives the contribution of additional methodological criteria to the Analytical Review criterion list and the improvement of the criteria and guidelines presented in this concluding chapter of Volume III.

The problems discussed herein can be subdivided into two general categories: (1) problems characteristic of survey research and (2) problems more uniquely characteristic of educational surveys.

Under the first category four major methodological issues have been identified:

A. Response bias.

Included under response bias are response errors due to interviewer and instrument effects.

B. Nonresponse bias.

C. Identifying causal relationships in survey data.

D. Selecting an appropriate sampling design.

Two problems that are more specifically related to educational surveys were discussed:

A. The absence of elaborating operations in data analysis.

B. Evaluating the impact of specific school variables, including:

1. The analysis of change.

2. Selecting an appropriate statistical model.

Common Problems in Survey Research

Response Bias

The responses from faculty and students to questions contained in the survey instrument can become distorted or biased for a number of reasons including their confusion, lack of concern, conscious or unconscious motivation, the

manner in which the questions are worded or ordered in the instrument and the characteristics or expectations of the interviewers. Because inferences drawn from survey data are so heavily dependent upon the validity of the written or verbal response, it is extremely important that the analyst empirically determine the extent and possible effects of these error factors.

Three sources of response error were discussed: the respondent, interviewer and questionnaire. The respondent may deliberately distort his answers for a number of reasons, including the desire for social approval and self-enhancement, distrust of the interviewer and the presence of a third person during the interview session. Respondent errors can also be unintentional due to selective perception or a faulty memory.

Four basic methods were suggested for reducing response errors that arise from the characteristics or attributes of the respondent.

Improving the Administration of the Questionnaire. The interview should be nonthreatening to the respondent. In addition, it is important that the respondent realize that he is making an important contribution to the study, which is ultimately dependent upon the validity of his responses. The interview should be task-oriented, and the respondent should be told that all replies will be held in strict confidence.

Modifying the Questionnaire. Response error can also be reduced by modifying the questionnaire. Exceptionally long questionnaires may tax the respondent's ability to

concentrate and give serious and thoughtful responses. To counteract a socially desirable response set, the respondent should be told that there are no right or wrong answers. In addition, the analyst may be able to "legitimize" a socially undesirable response by introducing a question with a statement indicating that many people endorse the less acceptable position. A number of excellent books on questionnaire construction are available, as enumerated in Chapter V.

Detecting Response Error. The analyst can also identify and exclude from analysis those respondents who erred in their replies. The following procedures were recommended for estimating response error.

1. Open-ended questions that require the respondent to elaborate upon his initial response either through a series of interlocking or contingency questions or through carefully designed interviewer probes.

The information gleaned from these questions can provide the analyst with valuable insights into the validity of the original answer and the consistency of overall responses.

2. Separate matched or contingency questions from one another in the questionnaire.

Open-ended questions are not effective error detecting devices when the errors are deliberate or intentional because the responses are usually not independent of one another. One solution to this problem is to separate the matched or contingency questions from one another in the survey instrument or, in the case of longitudinal surveys, to repeat the

items in future surveys when the responses to these questions should remain stable or have a known degree of change (e.g., age, class year, etc.).

3. Compare the responses to questions that have a theoretical or conceptual connection to one another.

Even when the analyst separates the matched or contingency questions he may not be successful in detecting intentionally motivated response bias. A more covert method that can often circumvent this problem involves comparing the responses to different questions that have a theoretical connection to one another. For example, the analyst may have reason to expect a considerable amount of normative or socially desirable responses from the students when asked to evaluate the performance of their teachers. To reduce this possible response bias, the researcher may ask the students to identify "objectively" the five most prominent characteristics of their teachers from a check list of teacher attributes, and at a later date ask the students to identify the five most desirable characteristics of a good teacher from the same list. By comparing the inconsistencies in these two responses the analyst can estimate the degree to which the students are really satisfied with the performance of their instructors.

4. Reduce the effects of motivated response by statistically controlling for the assumed motivation.

A practical solution to the problem of detecting motivated response bias is to measure and then statistically control for the motivation which is believed to account for some or all of the response error. Bachman, for example, administered the Crowne Marlowe scale of social approval. This index could have been used to statistically control for this response set in the analysis of selected relationships.

5. Include a question that is designed to reveal the response bias.

Another technique for detecting motivated response bias is to include a question that is specifically designed to identify the motivated response by making such a response invalid by definition. This technique is commonly employed to test for a socially desirable response set. For example, in a dichotomous answer-forced choice question (e.g., agree-disagree; yes-no, etc.) one answer may be highly desirable but never true (I always accept criticism graciously) while the alternative answer is always true but less desirable (I sometimes resent criticism). The analyst can then identify the respondents who desired social approval by looking at those who agreed with the highly desirable answer.

6. Reverse the wording of the question.

The most common procedure for identifying an acquiescence response set is to include matched questions that are worded in opposite directions. An example is where agreement to the first item signifies the respondent's acceptance of the issue contained in the item and where agreement to the matched item signifies rejection.

7. Cross-validate survey responses with comparable information contained in official records (e.g., through school files, interviews with the student's family, friends, schoolmates, or teachers).

Although official records and the student's "significant others" can often supply the analyst with a rich body of data that can be used to evaluate the validity of the student's responses, such investigations are frequently too expensive and time consuming. However, the analyst can draw a random subsample of responses for comparison with one of the above sources of information and estimate the magnitude of the response bias.

8. Cross-validate a question or set of responses on the basis of some factual characteristic that should influence the respondent's answer to a particular question or scale.

The analyst may be able to determine if the responses to a question appear reasonable by ascertaining whether another measurement, which is presumed to be both reliable and valid, is strongly associated with these responses in the expected direction.

9. Employ the split-ballot technique.

When the analyst suspects that the specific wording of questions or the sequence in which they are asked may influence the distribution of responses, the sample can be randomly divided into two or more groups with each subsample receiving a different version of the survey instrument. Then, by comparing the responses of the different groups, the researcher can determine the degree to which the findings are

independent of the specific procedures that were differentially administered to the groups.

10. Use independent surveys and/or samples.

Another method for validating the responses of a particular sample is to perform the same basic study on equivalent samples using the same or similar survey instruments. If the relationships uncovered in one study are truly independent of specific methodologies, then the sister survey should report similar results.

Correcting Scores. A final technique is to correct the scores for response bias. However, a generally accepted correction strategy does not exist at the present time.

Response bias can also result from sources which are external to the subject. The two primary sources of external response bias are the interviewer and the measurement instrument.

The observer or interviewer can artifactually influence the results of a study in a number of ways. For example: observers or interviewers often either become more adept as they gain experience with the demands of their job or become less adept and even careless due to boredom or fatigue. Observers or interviewers can also unconsciously impart a bias in the data by systematically recording equivocal statements or behaviors according to their personal biases or expectations of what the analysis should reveal. Moreover, the opinions or actual behaviors of the subject can actually be influenced by the characteristics, behaviors or expectations

of the observer or interviewer. Thus, the subject may respond to a question on the basis of what he believes is the expected answer rather than on the basis of his own judgment. There is, in addition, a considerable amount of data to suggest that certain personal characteristics of the interviewer (e.g., age, sex or race) may artifactually increase the probability that certain types of responses will occur.

Recommended procedures for preventing and/or controlling for response bias due to interviewer effects are as follows:

1. The detection of the amount and effect of interviewer variance and bias.
2. The avoidance of bias producing content and language.
3. The removal of ambiguous or "double idea" items.
4. Checking items for disparate frames of reference between the interviewer (or researcher) and respondent.
5. Reduction of interviewer improvisation.
6. Strict instructions for and control of the interviewer.
7. Selection of and assignment of interviewers to minimize undesirable interactions with respondents.
8. Training of interviewers in techniques designed to maximize optimal respondent motivation and minimize distorted or disparate cues.
9. The use of "nondirective" or "nonleading" interviewing.
10. Probability sampling of respondents.

The measurement instruments themselves can also be a major source of external response bias. In particular is the problem of test sensitization. Several researchers (e.g., Anastasi, 1958; Cane and Heim, 1950; French and Dear, 1959; Yates, 1953; James, 1953; Dempster, 1954 and Weisman, 1953) have demonstrated that students taking achievement, intelligence, or objective speeded tests for the second time usually do better than students taking the same or an alternative form of the test for the first time. Thus, taking a pretest or filling out a questionnaire at time one can be a learning experience which has the effect of artifactually producing change on the posttest or questionnaire administered at time two.

Secondly, there is the problem of instrumentation factors which result from changes in the measuring instrument between the pretest and posttest or, in the case of a single measurement, during the data collection phase of the investigation. For example, the response referent in a particular question can change in meaning over time. Thus, if college students were asked at the beginning of their freshmen year and again at the end of their senior year to describe their political orientation in terms of "radical--liberal" versus "moderate--conservative," any observed change or shift in the political orientation of these students could be due to (1) a real change in their political orientation or (2) changes that occurred during the four-year period in the connotations of the words radical, liberal, moderate

or conservative. Observing any change in the characteristics of the respondents over time may, in fact, reflect changes in the meaning of items contained in the measuring instrument as perceived by the respondents rather than true or real change.

Third, the generalizability of a research finding can be severely limited whenever the measuring device has reactive effects forming a stimulus for real rather than artifactual change. As Campbell and Stanley point out:

> It has long been a truism in the social sciences that the process of measuring may change that which is being measured. . . . The reactive effect can be expected whenever the testing process is in itself a stimulus to change rather than a passive record of behavior (1963, p. 9).

Administering survey questionnaires is not a passive method of recording behavior. Consequently the questionnaire itself can become a stimulus for real change. For example, questionnaires may form or influence the actual or real opinions of the respondent if they are heavily loaded with either negative or positive items that pertain to a particular issue, or contain quotes or the opinions of recognized experts. In addition, the responses to previous questions can frequently influence the responses to subsequent questions. Thus, students who are first made aware of their idealized self-concept may be more likely to answer questions according to this frame of reference than are students who are not first reminded of their idealized

self-concept. Even though the investigator may be measuring the real sentiments or behaviors of the respondent, if these responses are in some way influenced or dependent upon the construction of the questionnaire, then the generalizability of these responses will be restricted.

To test and/or prevent the biasing effect of test sensitization, the following procedures are recommended:

> Administer alternative forms of the same test to the entire sample.
>
> Randomly divide the sample into two or more groups and stagger the administration of the questionnaire or test. Thus, group one may receive the first test one year prior to group two, which receives the first test one year prior to group three, and so forth. The analyst can then compare the changes occurring between time periods in which different, but equivalent comparison groups, have received a limited number of tests over an extended period of time, as illustrated in Figure 12.

	Time 1	Time 2	Time 3	Time 4
Group 1	X			X
Group 2		X		
Group 3			X	
Group 4				X

Figure 12. Staggering the Administration of the Questionnaire to Random Subsamples

By dividing the sample in this fashion, the analyst has created a trend study within a panel design. For example, comparing the differences between the first and fourth time periods within group one represents a panel design, while the comparison of groups one and four represents a trend design. Selected items on the questionnaire can also be randomly assigned in this manner. When imperfect random assignment occurs, Covariance Analysis can be used to adjust the initial differences between subgroups.

The following procedures are recommended for reducing the occurrence of instrumentation factors.

> Provide the respondent with a specific definition of the response referent.
>
> Ask the respondents to define the response referent before they answer the question.
>
> Ask the respondents to describe any changes which may have occurred in the meaning of a particular response referent.

To reduce and/or identify the reactive effects of the questionnaire, the following procedures are recommended.

> Administer alternative forms of the same questionnaire to equivalent subsamples. If the obtained results are truly independent of the specific methodologies used, then equivalent subsamples should reveal similar results.
>
> Carefully pretest the questionnaire.
>
> Consult a text on questionnaire design. A number of excellent books are available, as enumerated in the previous chapter.

The methods for reducing and detecting response bias that were discussed above should not be considered exhaustive of all possible techniques. The methods used by the analyst will ultimately depend upon the variables involved, the characteristics of the respondents, the temporal and economic constraints placed upon the researcher and his creativity in conceptualizing and developing techniques for evaluating the validity of the responses he receives. Nevertheless, response errors are serious threats to the internal validity of survey research. Therefore, it is recommended that:

> Investigators should be required to delineate in their proposals what procedures they anticipate using in order to reduce or control for the problems of response bias.

Nonresponse Bias

Nonresponse bias refers to the differential loss or nonparticipation of respondents which biases the data. The validity of a study is particularly threatened when the nonresponse is systematic, that is, when particular subgroups of a sample have significantly lower rates of participation than others, and the nonresponse is associated with the independent or dependent variable.

There are two basic types of nonresponse in survey research: First, the rate or percentage of nonresponse for the entire survey instrument or "questionnaire nonresponse," and second, the nonresponse rate for particular questions or "item nonresponse." Questionnaire nonresponse may occur

at the time of the original survey as well as during subsequent follow ups in a longitudinal design (sample mortality).

Recommended procedures that can be employed to reduce the rate of initial questionnaire nonresponse are as follows:

Schedule the administration of the questionnaire at a time and place when the largest proportion of target respondents will be present. For example, although response rates are higher when questionnaires are administered in school than when they are mailed to respondents, student absenteeism from school is usually higher on Mondays and Fridays, when special extra-curricular activities are taking place and when the climate makes outdoor activities especially attractive.

Increase the number of attempts to contact the nonrespondents ("call-backs"). When this strategy is economically unfeasible for the entire nonrespondent population, draw a random subsample of nonrespondents for further call-backs.

Mail a postcard to the respondent (or his parents) prior to the interview or administration of the questionnaire explaining the nature and importance of the survey as well as the need for a high response rate. "Reminder postcards" can also be used to increase the rate of response for mailed questionnaires.

Obtain endorsements from the respondent's significant others such as school officials, peers, student leaders, PTA or government officials.

Employ methods to increase the student's identification with the project (e.g., periodically mailing newsletters, issuing project membership cards).

Recommended methods for adjusting the rate of initial questionnaire nonresponse are as follows:

It may be possible to extrapolate a curve to include the probable responses of the nonrespondent on the basis of call-back data.

Weight subsamples inversely to their response rate. This procedure should be used with caution, however, since in weighting responses the analyst substitutes missing responses with the responses of individuals who were successfully contacted. Consequently, weighting procedures can introduce bias as well as reduce it.

Recommended methods for reducing sample mortality are:

Maintain contact with the respondents during the interim period between measurements. This strategy will serve several useful functions. It will impress upon the respondent the importance of his participation in the survey and the concern for nonresponse. In addition, the analyst will be better informed about any plans of the respondent that may influence subsequent nonresponse such as vacations, school transfers and residential mobility. As a general rule, the longer the delay in discovering that a respondent is missing the greater is the likelihood that

he will remain lost. Periodic contacts will increase the probability that early detection of missing respondents will occur.

Contact may be further maintained by collecting extensive information about the respondents, including the names, addresses and occupations of close friends, siblings, neighbors and spouses, since these people may be instrumental in tracking down missing respondents.

Recommended methods for adjusting for sample mortality are:

Develop weighting procedures that incorporate information gathered from the missing respondent during an earlier measurement period and the responses of other respondents who have similar characteristics and responses.

Item nonresponse refers to the nonresponse rate for particular questions or items on the questionnaire. In longitudinal designs, this problem becomes intensified through item mortality, or the subsequent loss of responses to given items that were obtained in the initial survey.

Recommended procedures that can be employed to reduce item nonresponse are:

Carefully screen the questionnaires as they are turned in.

Arrange the format of the questionnaires so that the marked responses to the items will appear on the margins thereby making it easier to identify missing responses.

Contact the respondent by telephone. This strategy can also be used with a random subsample of nonrespondents.

Recommended procedures that can be used to adjust the rate of item nonresponse are as follows:

Assign each omission a response value equal to the arithmetic mean of those respondents with similar characteristics who answered the particular question.

Assign each omission the mean response plus some random coefficient so that these omissions will be randomly distributed throughout the distribution of responses.

Both of the above procedures serve to simplify tabulating procedures for later work and also remove bias from the sample estimates. However, using these strategies will delay processing the data. Consequently, when the item nonresponse rate is low, the best procedure may simply be to ignore it altogether.

Recommended methods for reducing item mortality are:

Use the same techniques suggested for item nonresponse.

Improve the design of the questionnaire based upon an evaluation of the item nonresponse rate from the previous survey.

Recommended procedures that can be employed to adjust the rate of item mortality:

Use the same techniques suggested for item mortality.

It is important to remember that weighting techniques and other adjustment strategies are not necessarily the solution to the problem of nonresponse. They can just as easily introduce bias as reduce it. Consequently, researchers should make every effort initially to reduce the rate of nonresponse. When the use of a specific technique for reducing or correcting for the rate of nonresponse is judged necessary but is prohibitively expensive to implement for the entire nonrespondent population, a random sample of nonrespondents can be drawn for further treatment.

Nonresponse is a serious methodological problem in survey research, particularly in educational surveys where the nonresponse rate typically runs as high as 40 percent. Therefore, it is recommended that:

> Investigators should be required to delineate in their proposals the methods they intend to use to reduce the rate of nonresponse and, if necessary, to correct the biases that result from a significant nonresponse rate.

Causal Analysis

Before the survey analyst can infer that one variable has "caused" another, it must be established that: (A) the variables concomitantly vary with one another; (B) the dependent variable does not precede the independent variable in temporal sequence; and (C) the observed relationship is not spurious or due to other factors which are temporally antecedent to both the independent and dependent variables.

A. Statistical Association. The first requirement involves selecting appropriate statistical test to measure the strength or significance of the covariation between two or more variables.

B. Establishing the Temporal Sequence of Variables. A causal relationship cannot be inferred unless the temporal sequence of the dependent and independent variables has been established.

The problem of establishing the proper temporal sequence of variables is only partially ameliorated by longitudinal analysis. Even in longitudinal designs it is often difficult to determine which of two variables in a relationship precedes the other. Several techniques that can be used to identify the most likely direction of causality are enumerated below.

1. Develop and test hypotheses derived from a conceptual model that specifies different outcomes involving a third variable depending upon which temporal sequence is operating.
2. Ask the respondent to clarify or elaborate upon the temporal sequence of variables through retrospective questions.
3. Consult external sources of information (e.g., teachers, parents or school records) which may contain clues to the probable temporal sequence of variables in a particular relationship.
4. Perform a cross-lagged panel correlation. The basic assumption of this technique is that the larger of two "cross-lagged" correlation coefficients identifies the most probable temporal sequence.

C. Testing for Spurious Relationships. An additional requirement for inferring causal relationships is to demonstrate that the relationship in question is not due to the effects of a third, antecedent variable.

The analyst must statistically control for those antecedent variables which could most likely produce a spurious relationship. If the relationship is sustained or replicated when the effects of these antecedent variables are statistically reduced the analyst may conclude that these antecedent factors do not likely account for the original relationship. However, if the relationship is sharply reduced when the effects of these variables have been controlled, then the analyst has reason to suspect that the original relationship is spurious.

The most important idea to remember concerning the problems of temporal sequence and spuriousness is anticipation. If the analyst is unable to anticipate these problems before the questionnaire reaches final revision or before field work begins, then he is unlikely to collect the type of information that is necessary to determine the direction of causality or test for spuriousness. Analysts frequently pretest and evaluate their instruments on the basis of the clarity and wording of the questions or on the basis of the distribution of responses they receive. It is equally important that the researcher determine in advance whether his data will be adequate to identify temporal sequences or spurious relationships.

Therefore, it is recommended that:

> Investigators develop a conceptual model which identifies the major hypotheses of the investigation. In addition, the analyst should discuss in detail credible rival hypotheses, and the methods to be used in testing them. Finally, the specific procedures to be employed in determining the temporal sequence of variables should be discussed for each major hypothesis.

Selecting an Appropriate Sampling Design

Educational samples should be random and probability samples. There are four basic types of random-probability samples; simple random samples, systematic samples, stratified samples and clustered samples. Most surveys, however, combine two or more of these designs. The major criterion for selecting a sampling design is the size of the standard error obtained at a given level of cost; the smaller the standard error, the more precise the sampling design. The size of the standard error can be decreased by increasing the size of the sample and/or decreasing the sample variance. Both of these factors should be taken into consideration when selecting a sampling design.

As a general rule, systematic random samples yield a sample variance approximating the sample variance obtained through simple random samples. Stratified probability samples are of two types; proportional and nonproportional. Proportionately stratified samples employ equal sampling fractions, while nonproportional samples use unequal sampling

fractions. The biggest gains from stratification occur when the stratifying criteria are associated with the variables to be studied. Disproportional stratification should not be used without first consulting a sampling expert who is familiar with the substantive area under investigation. Cluster samples are most appropriate when the target population can be divided into heterogenous subgroups. Again, cluster samples should be employed only after consulting with a sampling expert.

> Since the selection of an appropriate sampling design is a crucial aspect of survey methodology investigators should be required to justify their choice of sampling designs in terms of its projected effect upon the magnitude of the standard error at a fixed cost.

Specific Problems in Educational Surveys

Explicating Two Variable Relationships

Important information can be gained by introducing additional variables into the analysis which specify and interpret the relationships in question. To perform both specification and interpretation analysis, the analyst statistically controls for a third variable. In successful specification analysis the strength of the original relationship is increased and/or decreased in the partial relationships. There are no temporal constraints on specification

test factors. To perform interpretation analysis, the analyst statistically reduces the effects of a third variable which intervenes between the independent variable and dependent variable. In successful interpretation analysis, the original relationship disappears.

As a general rule, educational researchers have been negligent in explicating two variable relationships through interpretation and specification analysis. This is unfortunate because frequently the most important contributions to educational theory and policy are made when the investigator introduces third variable test factors into the analysis.

Educational researchers are currently preoccupied with developing prediction equations that link a number of independent variables to a specific dependent or criterion variable. Little or no effort has been made to clarify the dynamics of these relationships through systematic interpretation and specification analysis. Due to the central importance of these elaborating operations in the development of educational theory and enlightened policy recommendations, a serious effort should be made to supplement the current orientation of simply predicting educational outcomes to one that emphasizes the need to explicate predictive or causal relationships.

Therefore, it is recommended that:

> Investigators should be required to develop a conceptual model that identifies specific variables to be used as interpretation and

specification test factors. In addition, each major hypothesis of the investigation should be outlined and discussed in terms of which variables will be used to interpret and specify the expected results.

Impact Analysis

There are two basic conceptualizations of impact in educational research; impact as change and impact as outcome. When the analyst conceptualizes impact in terms of change, gain or difference scores are usually computed. The impact of a particular intervening stimulus is evaluated on the basis of the statistical association between selected independent variables and the criterion measure of change.

There are two major problems in interpreting gain scores. First, initially extreme scores have a tendency to regress toward the mean independently of the impact of an independent variable (regression effects). Secondly, it is less likely that extreme scores will become more extreme because of "ceiling" and "floor" effects. Although a great deal of attention has been devoted to both problems, there is no generally accepted solution to either. Consequently, constructing gain scores as the criterion measure in a study of impact is not recommended. A more satisfactory method of studying impact in a test-retest situation is to evaluate the outcomes of a particular experience by regressing the posttest score on the pretest score, along with selected independent variables.

Separating the effects of student background characteristics and other nonschool factors from the effects of the school is another major problem in impact analysis. Three basic modes of analysis are readily available to the educational researcher: (1) the causal-comparative model which is particularly useful for studies involving nominally qualified types of impact; (2) the input-output model which allows for assessing the impact of various levels or amounts of certain school characteristics; (3) the process model which allows for the analysis of a system in which school characteristics and student characteristics act together to produce outcome levels.

In the first two models the student background characteristics must be controlled in order to allow the impact of the variables associated with the educational system to be assessed. The process model is a framework for path analytic or simultaneous equation models and has the advantage of estimating the direct and indirect effects of school as well as nonschool factors. Since the statistical model employed by the investigator is a critically important aspect of data analysis, it is recommended that:

> Investigators should be required to justify in their proposals their choice of statistical models for estimating school impact.

BIBLIOGRAPHY

Anastasi, A. Differential psychology. (3rd ed.) New York: Macmillan, 1958.

Astin, A. W. Differential college effects. Journal of Educational Psychology, 1963, 54 (1), 63-71.

Astin, A. & Panos, R. J. The educational and vocational development of college students. Washington: American Council on Education, 1969.

Athey, K. R. et al. Two experiments showing the effects of the interviewer's racial background on responses to questionnaires concerning racial issues. Journal of Applied Psychology, 1960, 44, 244-246.

Averch, H. A., Carroll, S. J., Donaldson, T. S., Kiesling, H. J., & Pincus, J. How effective is schooling? A critical review and synthesis of resource findings. California: Rand Corp., 1972.

Bachman, J. G., Kahn, R. L., Mednick, M. T., Davidson, T. N. & Johnston, L. D. Youth in transition: Blueprint for a longitudinal study of adolescent boys. Vol. I. Ann Arbor: Institute for Social Research, University of Michigan, 1967.

Bachman, J. G. Youth in transition: The impact of family background and intelligence on tenth-grade boys. Vol. II. Ann Arbor: Institute for Social Research, University of Michigan, 1970.

Ball, J. The reliability and validity of interview data obtained from 59 narcotic drug addicts. American Journal of Sociology, 1967, 72, 650-654.

Bass, B. M. Authoritarianism or acquiescence? Journal of Abnormal and Social Psychology, 1955, 51, 616-623.

Bauman, L., Rogers, T. F., Lipson, S., Cantor, A. & Weiss, C. H. Bibliography on respondent-interviewer interaction in the research interview. Bureau of Applied Social Research, Columbia University, December, 1970. (Mimeo)

Benny, M., Riesman, D. & Star, S. Age and sex in the interview. American Journal of Sociology, 1956, 62, 143-153.

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

Berg, I. A. & Rapapert, G. M. Response bias in an unstructured questionnaire. Journal of Psychology, 1954, 38, 93-107.

Blalock, H. M. Social statistics. New York: McGraw-Hill, 1960.

Blalock, H. M. Theory construction from verbal to mathematical formulation. New Jersey: Prentice-Hall, 1970.

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

Blau, P. M. Determining the dependent variable in certain correlations. Public Opinion Quarterly, 1955, 19, 100-105.

Bohrnstadt, G. W. Observations on the measurement of change. In E. F. Borgatta (Ed.), Sociological methodology. San Francisco: Jossey Bass, Inc., 1969.

Bowles, S. & Levin, H. M. The determinants of scholastic achievement--an appraisal of some recent evidence. Journal of Human Resources, 1968, 3, 3-29. (A)

Bowles, S. & Levin, H. M. More on multicollinearity and the effectiveness of schools. Journal of Human Resources, 1968, 3, 393-400. (B)

Bryant, E. et al. Responses on racial attitudes as affected by interviewers of different ethnic groups. Journal of Social Psychology, 1966, 43, 53-61.

Bunker, J. P. et al. National Halothane study. Washington, D.C.: U.S. Government Printing Office, 1969.

Cahalan, D. Correlates of respondent accuracy in the Denver validity survey. Public Opinion Quarterly, 1968, 32, 607-621.

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

Campbell, D. T. The indirect assessment of social attitudes. Psychological Bulletin, 1950, 47, 15-38.

Campbell, D. T. From description to experimentation. In C. W. Harris (Ed.), Problems in measuring change. Madison: University of Wisconsin Press, 1962.

Campbell, D. T. & Clayton, K. N. Avoiding regression effects in panel studies of communication impact. Studies in Public Communication, 1961, No. 3, 99-118.

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

Cane, V. R. & Heim, A. W. The effects of repeated testing: III. Further experiments and general conclusions. Quarterly Journal of Experimental Psychology, 1950, 2, 182-195.

Cannell, C. Reporting of hospitalization in the health interview survey. Health Statistics Series D, No. 4. Public Health Service, U.S. Dept. of Health, Education and Welfare. Washington, D.C., May, 1961.

Cannell, C. F. & Kahn, R. L. Interviewing. In G. Lindzey & E. Aronson (Eds.), The handbook of social psychology. Vol. 2. (2nd ed.) Reading, Mass.: Addison-Wesley, 1968.

Cicourel, A. V. Method and measurement in sociology. New York: Free Press, 1964.

Clark, J. P. & Tifft, L. L. Polygraph and interview validation of self-reported deviant behavior. American Sociological Review, 1966, 31, 516-523.

Clausen, A. R. Response validity: Vote report. Public Opinion Quarterly, 1968, 32, 588-606.

Cochran, W. G. Sampling techniques. New York: Wiley, 1953.

Coleman, James S. Equality of educational opportunity. Washington: U.S. Government Printing Office, U.S. Department of Health, Education and Welfare, 1966.

Cook, S. W. & Selltiz, C. A multiple-indicator approach to attitude measurement. Psychological Bulletin, 1964, 62, 36-55.

Couch, A. & Keniston, K. Yeasayers and naysayers: Agreeing response set as a personality variable. Journal of Abnormal and Social Psychology, 1960, 60, 144-157.

Creager, J. A. On methods for analysis of differential input and treatment effects on educational outcomes. Paper read at the 1969 Annual Meeting of the American Educational Research Association. (A)

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

Creager, J. A. & Boruch, R. F. Orthogonal analysis of linear composite variance. Paper read at the 77th Annual Meeting of the American Psychological Association, 1969.

Cronbach, L. J. Response sets and test validity. Educational and Psychological Measurement, 1946, 61, 54-77.

Cronbach, L. J. & Furby, L. How should we measure change--or should we? Psychological Bulletin, 1970, 71, 68-80.

Crowne, D. P. & Marlowe, D. A new scale of social desirability independent of psychopathology. Journal of Consulting Psychology, 1960, 24, 349-354.

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

Deming, W. E. Some theory of sampling. New York: Dover Publications, 1950.

Dempster, J. J. B. Symposia on the effects of coaching and practice in intelligence tests. III. The South Hampton investigation and procedure. British Journal of Educational Psychology, 1954, 24, 1-5.

Duncan, O. D. Partials, partitions and paths. In E. F. Borgatta & G. W. Bohrnstedt (Eds.), Sociological methodology. San Francisco: Josey Bass, 1970.

Durbin, J. & Stuart, A. Callbacks and clustering in sample surveys. Journal of the Royal Statistical Society, 1954, 117, 387-428.

Edwards, A. L. The relationship between the judged desirability of a trait and the probability that the trait will be endorsed. The Journal of Applied Psychology, 1953, 53, 90-93.

Edwards, A. L. The social desirability variable in personality assessment and research. New York: Dryden, 1957.

Edwards, A. L. Social desirability or acquiescence in the MMPI? A case study with the SD scale. Journal of Abnormal and Social Psychology, 1961, 63, 351-359.

Edwards, A. L. A factor analysis of experimental social desirability and response set scales. Journal of Applied Psychology, 1963, 47, 308-316.

Edwards, A. L., Diers, C. I. & Walker, J. N. Response sets and factor loadings on sixty-one personality scales. Journal of Applied Psychology, 1962, 46, 220-225.

Edwards, A. L. & Walker, J. N. A note on the Couch and Keniston measure of agreement response set. Journal of Abnormal and Social Psychology, 1961, 62, 163-174. (A)

Edwards, A. L. & Walker, J. N. Social desirability and agreement response set. Journal of Abnormal and Social Psychology, 1961, 62, 180-183. (B)

Edwards, A. L. & Walsh, J. A. A factor analysis of ? scores. Journal of Abnormal and Social Psychology, 1964, 69, 559-563.

Elashoff, J. D. Analysis of covariance: A delicate instrument. American Educational Research Journal, 1969, 6, 383-402.

Evans, F. R. & Patrick, C. Antecedents and patterns of academic growth of school dropouts. In T. L. Hilton (Ed.), A study of intellectual growth and vocational development. New Jersey: Educational Testing Service, 1971.

Ferber, R. Item nonresponse in a consumer survey. Public Opinion Quarterly, 1966, 30, 399-415.

Ferber, R. & Wales, H. Detection and correction of interviewer bias. Public Opinion Quarterly, 1952, 16, 107-127.

Fisher, R. A. Statistical methods for research workers. (10th ed.) Edinburgh: Oliver and Boyd, 1946.

Flanagan, J. C., Dailey, J. T., Shaycoft, M. F., Orr, O. B., & Goldberg, J. Studies of the American high school. Final report to the U.S. Office of Education, Cooperative Research Project No. 226. Pittsburgh: Project TALENT Office, University of Pittsburgh, 1962.

Flanagan, J. C., Davis, F. B, Dailey, J. T., Shaycoft, M. F., Orr, O. C., Goldberg, I. & Neyman, C. A., Jr. The American high school student. Final report to the U.S. Office of Education, Cooperative Research Project No. 635. Pittsburgh: Project TALENT Office, University of Pittsburgh, 1964.

Flanagan, J. C., Shaycoft, M. F., Richards, J. M., Jr., & Claudy, J. G. Five years after high school. Final report to the U.S. Office of Education. Palo Alto: Project TALENT Office, American Institute for Research and University of Pittsburgh, 1971.

French, J. W. & Dear, R. E. Effect of coaching on an aptitude test. Educational and Psychological Measurements, 1959, 19, 319-330.

Friedman, N. The social nature of psychological research. New York: Columbia University Press, 1967.

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

Gentleman, W. M., Gilbert J. P. & Tukey, J. W. The smear and sweep analysis. In J. P. Bunker et al., National Halothane study. Washington, D.C.: U.S. Government Printing Office, 1969.

Gilbert, J. P. & Mosteller, F. Urgent need for experimentation. In F. Mosteller & D. Moynihan (Eds.), On equality of educational opportunity. New York: Random House, 1972.

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

Glock, C. Y. (Ed.) Survey research in the social sciences. New York: Russell Sage Foundation, 1967.

Goldberger, A. S. Econometrics and psychometrics: A survey of commonalities. Psychometrika, 1971, 36, 83-107.

Goodman, L. & Kruskall, W. H. Measures of association for cross-classification. Journal of American Statistical Association, 1954, 49, 732-764.

Gould, Leroy C. Who defines delinquency: A comparison of self-reported and officially reported indices of delinquency for three racial groups. Social Problems, 1969, 16, 325-336.

Gray, C. W. & Crisp, H. Credibility of pure response set. Paper read at Southeastern Psychological Association, Guttenburg, April, 1961.

Graybill, F. A. Introduction to linear statistical models. Vol. I. New York: McGraw-Hill, 1961.

Gullahorn, J. T. & Gullahorn, J. E. Increasing returns from nonrespondents. Public Opinion Quarterly, 1959, 23, 119-121.

Gullahorn, J. E. & Gullahorn, J. T. An investigation of the effects of three factors on response to mail questionnaires. Public Opinion Quarterly, 1963, 27, 294-296.

Gurin, P. & Katz, D. Motivation and aspiration in the Negro college. Office of Education, U.S. Department of Health, Education and Welfare Project No. 5-0787. Ann Arbor, Mich.: Survey Research Center, Institute for Social Research, University of Michigan, 1966.

Hansen, M. H., Hurwitz, W. N. & Madow, W. G. Sample survey methods and theory. Vol. I. New York: Wiley, 1953.

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

Helmstadter, G. C. Procedures for obtaining separate set and content components of a test score. Psychometrika, 1957, 22, 381-393.

Hilton, T. L. A study of intellectual growth and vocational development. New Jersey: Educational Testing Service, 1971.

Hites, R. W. Change in religious attributes during four years of college. Journal of Social Psychology, 1965, 66, 51-63.

Houseman, E. E. Statistical treatment of the nonresponse problem. Agricultural Economics Research, 1953, 5, 12-18.

Hovland, C. I., Lumsdaine, A. A. & Sheffield, F. D. Experiments on mass communication. Princeton: Princeton University Press, 1949.

Husen, I. (Ed.). International study of achievement in mathematics: Comparison of twelve countries. Vol. I & II. New York: Wiley; Stockholm: Almquist and Wiksell, 1967.

Hyman, H. Interviewing in social research. University of Chicago Press, 1954.

Hyman, H. Survey design and analysis. New York: The Free Press, 1955.

Isaac, S. & Michael, W. B. Handbook in research and evaluation. San Diego: Knapp, 1971.

Jackson, D. S. & Messick, S. Response styles and assessment of psychopathology. In S. Messick & J. Rose (Eds.), Measurement in personality and cognition. New York: Wiley, 1962.

James, W. S. Symposia on the effects of coaching and practice in intelligence tests. II. Coaching for all recommended. British Journal of Educational Psychology, 1953, 23, 155-162.

Johnston, J. Econometric methods. New York: McGraw-Hill, 1963.

Jones, M. C., Bayley, N., MacFarlane, J. W. & Honzik, M. P. (Eds.). The course of human development. Waltham, Mass.: Xerox College Publishing, 1971.

Kagan, J. & Moss, H. A. Birth to maturity. New York: Wiley, 1962.

Katz, D. Do interviewers bias poll results? Public Opinion Quarterly, 1942, 6, 248-268.

Katz, I. Body language: A study in unintended communication. Unpublished doctoral dissertation, Harvard University, 1964.

Katz, J. & Korn, H. A. and others. No time for youth. San Francisco: Jossey Bass, 1968.

Keesling, J. W. Maximum likelihood approaches to causal flow analysis. Unpublished Ph.D. dissertation, University of Chicago, Chicago, Ill., 1972.

Keesling, J. W. & Wiley, D. E. Some problems in the application of cross-lagged panel correlation. Paper presented at American Educational Research Association Meeting, 1969.

Kendall, M. G. & Stuart, A. The advanced theory of statistics. Vols. 1-3. New York: Hafner, 1961-1966.

Kephart, W. M. & Bressler, M. Increasing the responses to mail questionnaires: A research study. Public Opinion Quarterly, 1958, 22, 123-132.

Kish, L. Confidence intervals for clustered samples. American Sociological Review, 1957, 22, 154-165.

Kish, L. Survey sampling. New York: Wiley, 1965.

Kmenta, J. Elements of econometrics. New York: Macmillan, 1971.

Lansing, J. B. & Morgan, J. N. Economic survey methods. Ann Arbor: Institute for Social Research, University of Michigan, 1971.

Lazarsfeld, P. F. & Rosenberg, M. The language of social research. New York: The Free Press, 1955.

Lehmann, I. J. & Dressel, P. Critical thinking, attitudes and values in higher education. Michigan State University, 1962.

Lehmann, I. J. & Dressel, P. Changes in critical thinking ability, attitudes and values associated with college attendance. Michigan State University, 1963.

Lenski, G. E. & Leggett, J. C. Caste, class and deference in the research interview. American Journal of Sociology, 1960, 65, 463-468.

Lipset, S. M., Lazarsfeld, P. F., Barton, A. H. & Linz, J. The psychology of voting: An analysis of political behavior. In G. Lindzey (Ed.), Handbook of social psychology. Cambridge, Mass.: Addison-Wesley, 1954. Pp. 1124-1175.

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

Lord, F. M. Further problems in the measurement of growth. Educational and Psychological Measurement, 1958, 18, 437-451.

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

Lord, F. M. Statistical adjustments when comparing pre-existing groups. Psychological Bulletin, 1969, 72, 336-337.

Maccoby, E. Pitfalls in the analysis of panel data: A research note on some technical aspects of voting. American Journal of Sociology, 1956, 61, 359-362.

Maccoby, E. & Hyman, R. Measurement problems in panel studies. In E. Burdick & A. J. Brodbeck (Eds.), American voting behavior. Glencoe, Ill.: Free Press, 1959.

McGee, R. K. The relationship between response style and personality variable. Journal of Abnormal and Social Psychology, 1962, 64, 229-234.

McNemar, Q. On growth measurement. Educational and Psychological Measurement, 1958, 18, 47-55.

Mahler, I. Yeasayers and naysayers: A validating study. Journal of Abnormal and Social Psychology, 1962, 64, 317-318.

Messick, S. Separate set and content scores for personality and attitude scales. Educational and Psychological Measurement, 1961, 21, 915-923.

Messick, S. Response style and content measures from personality inventories. Educational and Psychological Measurement, 1962, 22, 41-56.

Messick, S. & Jackson, D. N. Desirability scale values and dispersions for MMPI. Psychological Reports, 1961, 8, 409-414.

Miklich, D. R. Item characteristics and agreement-disagreement response set. Unpublished doctoral dissertation, University of Colorado, 1965.

Miller, D. C. Handbook of research design and social measurement. New York: David McKay, 1970.

Moser, C. A. & Kalton, G. Survey methods in social investigation. (2nd ed.) New York: Basic Books, 1971.

Murray, J. R. Statistical models for qualitative data with classification errors. Unpublished Ph.D. dissertation, University of Chicago, Chicago, Ill., 1971.

Murray, J. R., Wiley, D. E. & Wolfe, R. G. New statistical techniques for evaluating longitudinal models. Human Development, 1971, 14, 142-148.

Neely, Twila B. A study of error in the interview. (Privately printed), 1937.

Newcomb, T. M. Personality and social change: Attitude formation in a student community. New York: Holt, 1943.

Newcomb, T. M., Koenig, K. E., Flacks, R. & Warwick, D. P. Persistence and change: Bennington College and its students after twenty-five years. New York: Wiley, 1967.

Oppenheim, A. N. Questionnaire design and attitude measurement. New York: Basic Books, 1966.

Parry, H. & Crossley, H. M. Validity of responses to survey questions. Public Opinion Quarterly, 1950, 14, 61-80.

Peabody, D. Authoritarianism scales and response bias. Psychological Bulletin, 1966, 65, 11-23.

Pelz, D. C. & Andrews, F. M. Detecting causal priorities in panel study data. American Sociological Review, 1964, 29, 836-848.

Phillips, D. L. Knowledge from what? Chicago: Rand McNally, 1971.

Platt, J. R. Strong inference. Science, 1964, 146, 347-353.

Politz, A. & Simmons, W. R. An attempt to get the "Not at homes" into the sample without callbacks. Journal of American Statistical Association, 1949, 44, 9-31.

Price, D. O. On the use of stamped return envelopes with mail questionnaires. American Sociological Review, 1950, 15, 672-673.

Rankin, R. & Campbell, D. T. Galvanic skin response to negro and white experimenters. Journal of Abnormal and Social Psychology, 1955, 51, 30-33.

Rao, C. R. Linear statistical inference and its applications. New York: Wiley, 1965.

Reynolds, H. T. Making causal inferences with ordinal data. Working paper No. 5. Chapel Hill: University of North Carolina Institute for Research in Social Science, 1971.

Rice, Stuart A. Contagious bias in the interview: A methodological note. American Journal of Sociology, 1929, 35, 420-423.

Robins, L. N. The reluctant respondent. Public Opinion Quarterly 1963, 27, 276-286.

Robinson, R. A. & Agism, P. Making mail surveys more reliable. The Journal of Marketing, 1951, 15, 415-424.

Rorer, L. G. The great response-style myth. Psychological Bulletin, 1965, 63, 129-156.

Rosenberg, M. J. When dissonance fails: On eliminating evaluation apprehension from attitude measurement. Journal of Personality and Social Psychology, 1965, 1, 28-42.

Rosenberg, M. J. The logic of survey analysis. New York: Basic Books, 1968.

Rosenberg, M. J. The conditions and consequences of evaluation apprehension. In R. Rosenthal & R. Rosnow (Eds.), Artifact in behavioral research. New York: Academic Press, 1969.

Rosenthal, R. Experimenter effects in behavioral research. New York: Appleton-Century-Crofts, 1966.

Rosenthal, R. & Rosnow, R. L. (Eds.). Artifact in behavioral research. New York: Academic Press, 1969.

Sarason, I. G. & Harmatz, M. Test anxiety and experimenter condition. Journal of Personality and Social Psychology, 1965, 1, 499-505.

Schutz, R. E. & Foster, R. J. A factor analytic study of acquiescent and extreme response set. Educational and Psychological Measurement, 1963, 23, 435-447.

Selltiz, C., Jahoda, M., Deutsch, M. & Cook, S. Research methods in social relations. New York: Holt, Rinehart and Winston, 1959.

Shaycoft, M. F. The high school years: Growth in cognitive skills. Pittsburgh: Project TALENT Office, University of Pittsburgh and American Institutes for Research, 1967.

Siegel, S. S. Nonparametric statistics. New York: McGraw-Hill, 1956.

Simmons, W. R. A plan to account for "not at homes" by combining weighting and callbacks. Journal of Marketing, 1954, 19. 42-53.

Skager, R., Holland, J. L. & Braskamp, L. A. Changes in self-ratings and life goals among students at colleges with different characteristics. ACT Research Report No. 14. Iowa City, Iowa: American College Testing Program, 1966.

Sonquist, J. A. & Morgan, J. N. The detection of interaction effects: A report of a computer program for the selection of optimal combinations of explanatory variables. Monograph No. 35. Ann Arbor: Institute for Social Research, 1964.

Stanley, J. C. (Ed.). Improving experimental design and statistical analysis. Chicago: Rand McNally, 1967.

Stephan, F. F. & McCarthy, P. S. Sampling opinion. New York: Wiley, 1958.

Stevenson, H. W. & Allen, S. Adult performance as a function of the sex of experimenter and sex of subject. Journal of Abnormal and Social Psychology, 1964, 68, 214-216.

Stevenson, H. W. & Odum, R. D. Visual reinforcement with children. Unpublished manuscript, University of Minnesota, 1963.

Stricker, L. J. Acquiescence and social desirability response styles, item characteristics, and conformity. Psychological Reports, 1963, 12, 319-341.

Summers, G. F. & Hammonds, A. D. Effect of racial characteristics of investigator on self-enumerated responses to a negro prejudice scale. Social Forces, 1966, 44, 515-518.

Super, D. E. & associates. Floundering and trial after high school. Career patterns study, monograph IV. Teachers' College, Columbia University, New York, 1967.

Taietz, P. Conflicting group norms and the "third" person in the interview. American Journal of Sociology, 1962, 68, 31-39.

Taylor, J. B. What do attitude scales measure: The problem of social desirability. Journal of Abnormal and Social Psychology, 1961, 62, 386-390.

Theil, H. Principles of econometrics. New York: Wiley, 1971.

Thistlethwaite, D. L. Effects of college upon student aspirations. Nashville: Vanderbilt University, 1965.

Thomson, G. H. A formula to correct for the effect of errors of measurement on the correlation of initial values with gains. Journal of Experimental Psychology, 1924, 7, 321-324.

Thomson, G. H. An alternative formula for the true correlation of initial value with gains. Journal of Experimental Psychology, 1925, 8, 323-324.

Thorndike, E. L. The influence of the chance imperfections of measures upon the relation of initial score to gain or loss. Journal of Experimental Psychology, 1924, 7, 225-232.

Tillery, D. SCOPE: School to college; opportunities for post-secondary education. Berkeley: Center for the Study of Higher Education, University of California, 1966.

Tillery, D. & Collins, C. College-going in four states: A study of differential outcomes of high school graduates. SCOPE Project, Center for Research and Development in Higher Education and the College Entrance Examination Board. New York, 1972. (A)

Tillery, D. & Kildegaard, T. Educational goals, attitudes and behaviors: A differential study of high school seniors. SCOPE Project, Center for Research and Development in Higher Education and the College Entrance Examination Board. New York, 1972. (B)

Trent, J. W. & Medsker, L. L. Beyond high school. San Francisco: Jossey-Bass, 1968.

Trow, M. Education and survey research. In C. Y. Glock (Ed.). Survey research in the social sciences. New York: Russell Sage Foundation, 1967.

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

Tukey, J. W. Causation, regression and path analysis. In O. Kempthorne et al. (Eds.), Statistics and mathematics in biology. Iowa: State College Press, 1954.

Webster, H. Extension of a simple psychometric model to measure change. Berkeley: Center for the Study of Higher Education, 1963. (Mimeo)

Webster, H. Factors that measure true change. Unpublished monograph, 1968.

Weiss, C. Validity of welfare mothers' interview responses. Public Opinion Quarterly, 1968, 32, 622-633.

Weiss, C. Comments on interviewer biasing effects: Toward a reconsideration of findings. Public Opinion Quarterly, 1969, 33, 127-129.

Weiss, D. J. & Davis, R. V. An objective validation of factual interview data. Journal of Applied Psychology, 1960, 44, 117-123.

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

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

Werts, C. E. & Watley, D. J. Analyzing college effects: Correlation vs. regression. American Educational Research Journal, 1968, 5, 585-598.

Wiley, D. E. & Wiley, J. A. The estimation of measurement error in panel data. American Sociological Review, 1970, 35, 112-117.

Willcock, H. D. The effects of interviewers' own opinions about minorities and foreigners on the opinion about negroes which they obtain from informants. The Social Survey, Methodology Series No. M45. London: Central Office of Information, 1951. Pp. 26-35.

Wiseman, S. Symposia on the effects of coaching and practice in intelligence tests. IV. The Manchester experiment. British Journal of Educational Psychology, 1953, 23, 5-8.

Wiseman, S. & Wrigley, J. The comparative effects of coaching and practice on the results of verbal intelligence tests. British Journal of Psychology, 1953, 44, 83-94.

Wright, S. Statistical methods in biology. Journal of the American Statistical Association, 1931, 26, 155-163.

Yates, A. Symposia on the effects of coaching and practice in intelligence tests: I. An analysis of some recent investigation. British Journal of Educational Psychology, 1953, 23, 147-155.

Yates, F. Sampling methods for censuses and surveys. New York: Hafner, 1949.

Zarkovich, S. S. Sampling methods and censuses. Vol. II. Quality of Statistical Data. Rome: FAO, (draft) 1963.

Zieve, L. Note on the correlation of initial scores with gains. Journal of Educational Psychology, 1940, 31, 391-394.

ADDENDUM

Bell, C. G. & Buchanan, W. Reliable and unreliable respondents: Party registration and prestige pressure. Western Political Quarterly, 1966, 29, 37-43.

Binder, A. D. et al. Verbal conditioning as a function of experimenter characteristics. Journal of Abnormal and Social Psychology, 1957, 55, 309-314.

Gergen, K. J. & Back, K. W. Communication in the interview and the disengaged respondent. Public Opinion Quarterly, 1966, 30, 385-398.